Modelling Categorical Data: Loglinear models and logistic regression

Brendan Halpin,
G&S, Limerick University

August 26–30, 2002
Preliminary info

Contact information:
Brendan Halpin
Email: brendan.halpin@ul.ie
Sociology Group
Dept of Government & Society
University of Limerick
Limerick
Ireland.
Overview

- An introduction the analysis of categorical data
  - Loglinear models
  - Logistic regression

- **Applied** focus:
  - More emphasis on learning to run and interpret analyses;
  - Less emphasis on maths and statistical theory.

- Teaching times
  - 08h30 to 18h00
Models

- We will look at a number of related ways of modelling categorical data

- Loglinear models: model the structure of multi-dimensional tables of categorical variables – “association” models $\Rightarrow$ no dependent variable
  - 2-dimensional tables
  - Higher dimensions
  - Special topics such as square tables or “matched pairs”
  - Models for ordinal variables
• Grouped logistic regression: modelling tabular data with a binary dependent variable

• Individual-level logistic regression: micro-data with a binary dependent variable, continuous covariates possible

• Multinomial logistic regression: dependent variable with more than two categories

• Ordinal logistic regression: dependent variable with more than two categories, which are ordered
Software

- **SPSS**
  - For data manipulation and most analysis;
  - Syntax favoured over menu interface;
  - Other programs partly supported if demand exists: Stata, LispStat, GLIM, SAS.

- **Excel Spreadsheet**
  - Very useful for manipulating tables and arbitrary calculations
Modelling Categorical Data
Introduction: the analysis of categorical data

- Categorical data is perhaps the most common form of data in the social sciences
- But the best-established techniques favour continuous data
  - t-tests, correlations
  - Regression analysis, ANOVA
  - Factor Analysis, etc.
- However, a battery of techniques are available for categorical data, including
  - cross-tabulation
  - loglinear modelling, and
  - binary, multinomial and ordinal logistic regression
What’s Categorical Data?

- Categorical data consists of variables with a finite number of values (really, a small number of discrete values).

- It can be:
  - nominal
  - ordinal
  - interval
  - or ratio data,

- But it can’t be continuous.
• Categorical data arise in a number of ways
  – Simple counts
  – Binary variables – yes/no, pass/fail, live/die
  – Unordered multinomial: christian, jew, muslim, atheist
  – Ordinal:
    * Pure: degree, complete second level, incomplete second level, primary only, no education.
    * Imperfect scale measurement: *e.g.*, Likert’s 5-point scale
    * Grouped variables: *e.g.*, income in bands

• Some of these are interval–ratio variables:
  – Count of number of children
  – Income in bands (imprecise ratio measurement)

• and we can use appropriate summaries, such as means and correlations.
• However, the really powerful technique of **Ordinary Least Squares Regression** won’t do:
  – OLS requires a dependent variable that is **conditionally normally distributed**
  – OLS may well predict impossible values – negative counts, probabilities outside the 0–1 range

• So something else is required
• Consider first count data: non-negative integers
• Where the mean of a count variable is sufficiently large, its distribution approximates the normal and **OLS will do okay**
• Counts are distributed as Poisson:

\[ P(Y = y) = \frac{e^{-\mu} \mu^y}{y!} \]

• This is a *discrete* function
• Asymmetric but approximated by the normal distribution for large values of \( \mu \)
• Standard deviation is \( \sqrt{\mu} \)
Poisson Regression

- An alternative to OLS:

\[ \log(\text{count}) = \alpha + \beta X + \varepsilon \]

or

\[ \text{count} = e^{\alpha+\beta X + \varepsilon} \]

- Predicted values less than zero are impossible

- Error term has appropriate distribution
Binary dependent variables

- For binary dependent variables we are interested in the probability of one outcome versus another: 0–1 range or 0–$n$ range for $n$ chances

- Binary dependent variables have a binomial distribution that depends on the number of trials and the probability of a “yes” in any single trial

$$P(Y = y) = \frac{N!}{y!(N-y)!} \pi^y (1 - \pi)^{N-y}$$
Binary regression

- With this distribution for the error term, and

- a suitable transformation for the dependent variable

  1. the logit: $\log\left(\frac{p}{1-p}\right)$ or “log of the odds"
  2. or the inverse of the cumulative normal distribution: $\Phi^{-1}$

- we have a form of regression for binary dependent variables
These two mappings translate 0 ↔ 1 into −∞ ↔ ∞

They are symmetric, and quite similar in effect

For extreme probabilities, a bigger change in f(P) is required for the same change in P as for probabilities around 0.5

Logit and probit regressions have the same form:

\[ g(p) = \alpha + \beta X + \varepsilon_{b,n} \]

The logit transform gives us “logistic regression”

This is favoured over probit in recent years, for its mathematical tractability
Models for categorical data

- Poisson regression is directly useful for natural counts and rates
- However, we often create counts from nominal data: tables
- Loglinear modelling uses poisson regression to model the structure of tables in terms of cell counts
The logit transformation gives us logistic regression with binary dependent variables, either

– grouped (N successes out of M trials), or
– individual (each individual fails or succeeds, 0 or 1 success out of one trial)

By extension it gives us

– multinomai logistic regression
– ordinal logistic regression
Tabulation and association

- Categorical data lends itself to tabulation:
  - Tabulation retains all the information in the data, and
  - makes the **structure** of the data clearer.
<table>
<thead>
<tr>
<th>FJBJSTAT</th>
<th>Row</th>
<th>1</th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-employed</td>
<td></td>
<td>546</td>
<td>175</td>
<td>721</td>
</tr>
<tr>
<td>Employed</td>
<td></td>
<td>2359</td>
<td>2382</td>
<td>4741</td>
</tr>
<tr>
<td>Unemployed</td>
<td></td>
<td>261</td>
<td>132</td>
<td>393</td>
</tr>
<tr>
<td>Retired</td>
<td></td>
<td>708</td>
<td>950</td>
<td>1658</td>
</tr>
<tr>
<td>Maternity leave</td>
<td></td>
<td>34</td>
<td>34</td>
<td>.4</td>
</tr>
<tr>
<td>Family care</td>
<td></td>
<td>24</td>
<td>795</td>
<td>819</td>
</tr>
<tr>
<td>FT studt, school</td>
<td></td>
<td>313</td>
<td>360</td>
<td>673</td>
</tr>
<tr>
<td>LT sick, disabld</td>
<td></td>
<td>193</td>
<td>153</td>
<td>346</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>22</td>
<td>15</td>
<td>37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column</th>
<th>4426</th>
<th>4996</th>
<th>9422</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>47.0</td>
<td>53.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

(Source: BHPS wave 6)
• Tabulation allows us to see four things not apparent from the raw data:
  1. The overall male–female distribution;
  2. The overall employment-status distribution;
  3. The different employment-status distributions within the sexes; and
  4. The different sex proportions in each employment status.

• The latter two constitute the **association** between the two variables.

• We can see this more clearly with percentages
  – To focus on how employment status varies across sex we use column percentages . . .
  – . . . to get proportions within each sex.
### FSEX

<table>
<thead>
<tr>
<th>FJBSTAT</th>
<th>1</th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-employed</td>
<td>546</td>
<td>175</td>
<td>721</td>
</tr>
<tr>
<td>Employed</td>
<td>2359</td>
<td>2382</td>
<td>4741</td>
</tr>
<tr>
<td>Unemployed</td>
<td>261</td>
<td>132</td>
<td>393</td>
</tr>
<tr>
<td>Retired</td>
<td>708</td>
<td>950</td>
<td>1658</td>
</tr>
<tr>
<td>Maternity leave</td>
<td>34</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>Family care</td>
<td>24</td>
<td>795</td>
<td>819</td>
</tr>
<tr>
<td>FT studt, school</td>
<td>313</td>
<td>360</td>
<td>673</td>
</tr>
<tr>
<td>LT sick, disabld</td>
<td>193</td>
<td>153</td>
<td>346</td>
</tr>
<tr>
<td>Other</td>
<td>22</td>
<td>15</td>
<td>37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>546</td>
<td>546</td>
<td></td>
</tr>
<tr>
<td></td>
<td>175</td>
<td>175</td>
<td></td>
</tr>
<tr>
<td></td>
<td>721</td>
<td>721</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2359</td>
<td>2359</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2382</td>
<td>2382</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4741</td>
<td>4741</td>
<td></td>
</tr>
<tr>
<td></td>
<td>261</td>
<td>261</td>
<td></td>
</tr>
<tr>
<td></td>
<td>132</td>
<td>132</td>
<td></td>
</tr>
<tr>
<td></td>
<td>393</td>
<td>393</td>
<td></td>
</tr>
<tr>
<td></td>
<td>708</td>
<td>708</td>
<td></td>
</tr>
<tr>
<td></td>
<td>950</td>
<td>950</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1658</td>
<td>1658</td>
<td></td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td></td>
<td>819</td>
<td>819</td>
<td></td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>360</td>
<td>360</td>
<td></td>
</tr>
<tr>
<td></td>
<td>673</td>
<td>673</td>
<td></td>
</tr>
<tr>
<td></td>
<td>193</td>
<td>193</td>
<td></td>
</tr>
<tr>
<td></td>
<td>153</td>
<td>153</td>
<td></td>
</tr>
<tr>
<td></td>
<td>346</td>
<td>346</td>
<td></td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column</th>
<th>4426</th>
<th>4996</th>
<th>9422</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>47.0</td>
<td>53.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

- Percentages facilitate comparison by ‘controlling’ for marginal (i.e., row and column distribution) differences.

### Higher dimensions

- Tables can have any number of dimensions, though they get harder to read and format.

- This table has four dimensions: status at age 25, status at age 35, cohort and sex:
### General status, age 25 by age 35, by cohort and sex

<table>
<thead>
<tr>
<th>State at 25</th>
<th>Men</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort 1</td>
<td>1</td>
<td>15</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>14</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6</td>
<td>128</td>
<td>1</td>
<td>0</td>
<td>150</td>
<td>256</td>
<td>3</td>
<td>83</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>42</td>
<td>0</td>
<td>2</td>
<td>45</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>14</td>
<td>53</td>
<td>14</td>
<td>0</td>
<td>2</td>
<td>30</td>
<td>1</td>
<td>30</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>36</td>
<td>185</td>
<td>58</td>
<td>0</td>
<td>19</td>
<td>298</td>
<td>18</td>
<td>115</td>
<td>33</td>
<td>3</td>
</tr>
<tr>
<td>Cohort 2</td>
<td>1</td>
<td>41</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>43</td>
<td>24</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>14</td>
<td>285</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>303</td>
<td>3</td>
<td>132</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>58</td>
<td>0</td>
<td>1</td>
<td>60</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>20</td>
<td>43</td>
<td>8</td>
<td>0</td>
<td>6</td>
<td>77</td>
<td>9</td>
<td>103</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>76</td>
<td>331</td>
<td>67</td>
<td>210</td>
<td>486</td>
<td>464</td>
<td>36</td>
<td>204</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Cohort 3</td>
<td>1</td>
<td>91</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>98</td>
<td>34</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>15</td>
<td>346</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>366</td>
<td>7</td>
<td>109</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>34</td>
<td>0</td>
<td>0</td>
<td>37</td>
<td>4</td>
<td>3</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>22</td>
<td>22</td>
<td>103</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>116</td>
<td>360</td>
<td>38</td>
<td>1</td>
<td>9</td>
<td>524</td>
<td>67</td>
<td>219</td>
<td>28</td>
<td>3</td>
</tr>
<tr>
<td>Cohort 4</td>
<td>1</td>
<td>176</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>189</td>
<td>78</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>33</td>
<td>434</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>487</td>
<td>14</td>
<td>204</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>33</td>
<td>6</td>
<td>0</td>
<td>45</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>16</td>
<td>27</td>
<td>187</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>219</td>
<td>451</td>
<td>38</td>
<td>2215</td>
<td>745</td>
<td>125</td>
<td>403</td>
<td>2211</td>
<td>311</td>
<td>311</td>
</tr>
<tr>
<td>Cohort 5</td>
<td>1</td>
<td>184</td>
<td>15</td>
<td>0</td>
<td>9</td>
<td>4</td>
<td>212</td>
<td>92</td>
<td>15</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>61</td>
<td>380</td>
<td>5</td>
<td>41</td>
<td>14</td>
<td>501</td>
<td>32</td>
<td>215</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>12</td>
<td>7</td>
<td>2</td>
<td>37</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td>9</td>
<td>0</td>
<td>12</td>
<td>5</td>
<td>28</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>11</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>25</td>
<td>36</td>
<td>155</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>264</td>
<td>419</td>
<td>17</td>
<td>71</td>
<td>32</td>
<td>803</td>
<td>169</td>
<td>396</td>
<td>19</td>
<td>8</td>
</tr>
<tr>
<td>Grand total</td>
<td></td>
<td>711</td>
<td>1,746</td>
<td>218</td>
<td>96</td>
<td>85</td>
<td>2,856</td>
<td>415</td>
<td>1,337</td>
<td>144</td>
<td>31</td>
</tr>
</tbody>
</table>
Loglinear Analysis

Unit 1

- It summarises 6,322 cases in a table with 500 cells.

- It arranges the status-at-25/status-at-35 subtables line cells in a cohort by sex table.

- Careful inspection reveals a lot of structure, but the four-way relationship is too complex to grasp by simple examination.
Independence

- Structure: Is there association?
  - Does the distribution of one variable differ across the categories of the other variable?
  - Examine row or column percentages:
    * For instance, retirement is the status of 19% of women but only 16% of men,
    * for self employment the corresponding figures are 3.5% and 12.3%.
    * That is, the distribution of employment status clearly differs by sex (and vice versa).

- Association is a non-directional:
  - row or column percentages equally useful.
• If there is no association the variables are said to be independent. In a table showing independence, the percentage distributions within any row (or column) will be the same.
<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>10</td>
<td>40</td>
<td>50</td>
<td>II</td>
<td>14</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>24</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Row percents</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>20%</td>
<td>80%</td>
<td>50</td>
<td>II</td>
<td>20%</td>
<td>80%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>24</td>
<td>96</td>
<td>120</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Column percents</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>41.67%</td>
<td>41.67%</td>
<td>50</td>
<td>II</td>
<td>88.33%</td>
<td>88.33%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>24</td>
<td>96</td>
<td>120</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Given a set of marginals (i.e., row and column totals) we can calculate the expected values under the assumption of independence according to the following formula:

$$E_{ij} = \frac{n_i n_j}{n_{++}}$$
Measures of association

- There are many summary measures of association in tables.

- In the simplest case of a $2 \times 2$ table, a popular measure is $\phi$ (phi):
  
  - Given this table:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>2</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

  - Phi is defined as:

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(b+d)(a+c)}}$$
– Where there are higher than expected numbers in cells a and/or d, $\phi$ will tend to 1, and where the opposite is true, to $-1$. 
• For tables of variables with more than two categories an analogous measure is Cramér’s $V$.
  – For $2 \times 2$ tables it is equivalent.
  – Same range $-1 < V < +1$.
  – Based on Pearson’s $X^2$, a more general measure (Cramér’s $V$ is a version of $X^2$ scaled to be independent of table size).

• Pearson’s $X^2$ is based on the deviation between the observed values ($O$) and the expected ($E$) values under the assumption of independence:

$$X^2 = \sum \frac{(O - E)^2}{E}$$

That is, for each cell it calculates a measure of the observed–expected difference, and adds them up.
Pearson’s $X^2$ has a $\chi^2$ (chi-squared) distribution which allows us to make inferences about association:

- Sampling from a population where two variables are truly independent will result in tables which do not exactly match the independence table.
- The probability distribution of the calculated $X^2$ value under these circumstances follows a $\chi^2$ distribution, with degrees of freedom equal to $(r - 1)(c - 1)$.
- By comparing the $X^2$ for a real table with the cumulative $\chi^2$ distribution we can test the null hypothesis that there is no association.
- e.g., if the $X^2$ is at least as big as the value you could theoretically get from truly independent variables no more than, say, 1% of the time, then you can be 99% confident that there is really association.

- Many measures relevant to loglinear models approximate a $\chi^2$ distribution.
The Odds ratio

- The **odds ratio** is a very useful device for the analysis of categorical data.
  - It measures association
  - and underlies the maths behind loglinear models and logistic regression.

- What are **odds**?
  - The odds of outcome 1 versus outcome 2 are the probability (or frequency) of outcome 1 divided by the probability (or frequency) of outcome 2.
  - Contrast this with probability or proportions: one category’s probability is its frequency divided by the total frequency:
\[ O_{ij} = \frac{f_i}{f_j} \]

\[ \pi_i = \frac{f_i}{f_+} \]

- Where there are only two outcomes (e.g., winning and losing):

\[ O_{ij} = \frac{f_i}{f_+ - f_i} \equiv \frac{p_i}{1 - p_i} \]

- A horse given odds of 5:2 has 2 chances in 7 (= 5 + 2) of winning \((p = \frac{2}{7} = 0.286)\).
• Odds measure the frequency or probability of one outcome relative to another.
- **Odds ratios** involve comparing the odds of a pair of outcomes on one variable, for a pair of categories of a second variable: \( OR_{ij,kl} = \frac{O_{i,kl}}{O_{j,kl}} \)

<table>
<thead>
<tr>
<th></th>
<th>Owners</th>
<th>Private renters</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cons</td>
<td>1977</td>
<td>211</td>
<td>2188</td>
</tr>
<tr>
<td>Lab</td>
<td>2353</td>
<td>378</td>
<td>2731</td>
</tr>
<tr>
<td>Total</td>
<td>4330</td>
<td>589</td>
<td>4919</td>
</tr>
</tbody>
</table>

- The overall odds of owning versus renting are \( \frac{4330}{589} = 7.35 \). That is, there are far more owners.

- However, separately for Conservative and Labour voters:

  - Conservative: \( \frac{1977}{211} = 9.37 \)
  - Labour: \( \frac{2353}{378} = 6.22 \)
  - Ratio: \( \frac{9.37}{6.22} = 1.51 \)
• The Odds-Ratio for conservative versus labour voters being owners versus private renters is 1.51. That is, even though labour voters are predominantly house-owners, conservative voters are much more likely to be house-owners.

• A $2 \times 2$ table under independence will have an odds ratio of 1, whatever the marginals.

• If we know the odds ratio and the marginals for a $2 \times 2$ table we can calculate all the cell sizes.

• For bigger 2-D tables, we can reconstruct all the cell counts from the marginals and a set of odds ratios that involve all the cells ($(r - 1) \times (c - 1)$ is sufficient).

• The odds ratios are therefore equivalent to the structure of association.
Loglinear models essentially define a pattern of odds ratios, apply the marginals to them, and compare the resulting table with the observed table, in pretty much the same way we apply the Pearson $X^2$ test for association. The big difference is the pattern we define can be much more complicated than independence.
Data representations

• Categorical data is neat:
  – small number of values
  – a set of $N$ variables can be represented as an $N$-dimensional table without loss of information.

• There are three main ways of representing categorical data and we can move between them:
  1. Conventional data set: one row per case;
  2. Table: one dimension per variable; and
Conventional rectangular data set: one row per case, one column per variable:

Figure 1: A data matrix, in the SPSS data editor...

Micro-data of this form is the most general because it allows each case
to have different values on any variable: for instance, it is the only way to store continuous-variable data, where it is possible that every case has a different value.

- Micro-data consisting exclusively of categorical variables can be represented as a table; the table can be re-expressed in machine-readable form, where each cell is given a row, containing the values of each of the classifying variables and the cell count:
Table 1: Two table formats

<table>
<thead>
<tr>
<th>A ‘natural’ format</th>
<th>A ‘machine-readable’ format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable 1</td>
<td>1 1 11</td>
</tr>
<tr>
<td>1</td>
<td>11 12</td>
</tr>
<tr>
<td>2</td>
<td>21 22</td>
</tr>
<tr>
<td>3</td>
<td>31 32</td>
</tr>
<tr>
<td>Variable 2</td>
<td>1 2 12</td>
</tr>
</tbody>
</table>

- The machine-readable format is used as input to stats programs (necessary format for programs like GLIM and Stata, acceptable for programs like SPSS and SAS).
• This format can be easily produced by stats programs (for exporting data to other programs). For instance, in SPSS:

```plaintext
procedure output out='temp.dat'.
crosstabs /var=a(1,5) b(1,3)
    /tables=a by b
    /write=all.
```

• And finally, data in this form can be expanded back to micro-data by using weights:

```plaintext
data list free /dose outcome count.
begin data.
  1 3 15
  1 2 1
  1 1 281
```

Brendan Halpin, Sociology Group, Dept of Government & Society, University of Limerick
Swiss Summer School, Lugano August 2002
2 3 17
2 2 0
2 1 225
3 3 22
3 2 7
3 1 283
4 3 38
4 2 59
4 1 202
5 3 144
5 2 132
5 1 9
end data.

weight by count.
• Weights make the stats procedures act as if, for instance, there were 15 different cases with dose=1 and outcome=3.

• In short, three equivalent, interchangeable formats.
Loglinear models for 2-way tables

In this unit we consider simple loglinear models for two-way tables, and how to choose between models on the basis of goodness of fit.
Analysing the structure of tables

- The simplest test we can apply to a table is independence: a null hypothesis of no association:
  - Pearson’s $X^2$ test.
- For more detailed patterns of association, or higher dimensional tables we need a more general tool:
  - Loglinear models.
• A loglinear model is a **generalised linear model** which estimates the cell counts of a table using different combinations of the ‘margins’.

  – In a two-way table the margins are the row and column totals.
  – In higher dimensional tables, the margins are the sub-tables of lower dimension: *e.g.*, in an A*B*C table, the A*B, A*C and B*C tables (and indeed the A, B and C one-way frequency distributions) are the margins.
Tangent: GLMs

- Generalised Linear Models are a set of statistical models which have a dependent variable which is a function of a linear combination of variables and parameters, with an error component. Normal regression is the simplest example; loglinear models are fitted using ‘poisson regression’.

\[ f(y_i) = \alpha + \beta x_i + e_i \]

- The program GLIM used to be used extensively for loglinear modelling because it was the first package to do Generalised Linear Models.
- Most statistics packages now have GLM modules so even if they don’t have explicit loglinear commands they can fit loglinear models.
A loglinear model takes this form:

$$\log f_{ijk} = \lambda_0 + \lambda_i I + \lambda_j J + \lambda_k K + e_p$$

or

$$\log F_{ijk} = \lambda_0 + \lambda_i I + \lambda_j J + \lambda_k K$$

where $F_{ijk} = f_{ijk} + e_p$. This is exactly equivalent to

$$F_{ijk} = \tau_0 \tau_i \tau_j \tau_k$$

where $\lambda = \log \tau$. 
Independence in loglinear terms

- Independence is equivalent to the following loglinear model:

$$F_{ij} = \tau_0 \tau^A_i \tau^B_j$$

- That is, the fitted or expected value in the $ij$th cell is calculated by multiplying a ‘grand mean’ effect, an effect for row $i$, and an effect for column $j$. Compare with the formula we already used to calculate expected values under independence:

$$E_{ij} = \frac{n_{i+j} n_{++}}{n_{++}}$$
Loglinear allows us other models

- We can define several other loglinear models of a 2-way table:
  - Grand mean:
    \[ F_{ij} = \tau_0 \]
  - Equiprobability on one dimension:
    \[ F_{ij} = \tau_0 \tau_i \]
    and
    \[ F_{ij} = \tau_0 \tau_j \]
  - Independence:
    \[ F_{ij} = \tau_0 \tau_i \tau_j \]
– Association:

\[ F_{ij} = \tau_0 \tau_i \tau_j \tau_{ij} \]
What these models mean

- Grand mean:

\[ F_{ij} = \tau_0 \]

- All cells in the table have the same expected value
- *e.g.*, in a 1-D table (frequency distribution) we might expect the same amount in each category
- An example would be suicides by day of the week: we might first try a base model that says day of the week has no effect on numbers of suicides
- A 2-D example could be suicides by day-of-week and season: a reasonable base model might say neither affects the number of suicides, and we should expect approximately the same number on a Sunday in winter as a Tuesday in autumn.
• Equiprobability on one dimension:

\[ F_{ij} = \tau_0 \tau_i \]

– For instance, we might find that the suicide rate varies by day-of-week, but not by time of year.

• Independence:

\[ F_{ij} = \tau_0 \tau_i \tau_j \]

– We might find that there is a seasonal effect as well as a day-of-week effect, but that these don’t interact: e.g., the seasonal effect is approximately the same for all days, or the weekly effect is the same all year round.
Association:

\[ F_{ij} = \tau_0 \tau_i \tau_j \tau_{ij} \]

- Where independence doesn’t hold we have association.
- For at least some combinations of categories of the two variables there is an extra effect: another way of saying this is that there is an ‘interaction’ between the variables.
- In the day-of-week/season example, there could be a weekly effect that is different in different seasons, or equivalently a seasonal effect that is different for different days of the week (e.g., winter Fridays may be exceptionally high compared with other Fridays or other winter days, summer Tuesdays exceptionally low, etc.).
Association

- Association in a two-way table is represented by the $\tau_{ij}$ term which implies that for every combination of the row and column categories there is a different, additional, effect over and above the row and column specific effects.

- This reproduces the observed data exactly and is called the saturated model.

- It has no degrees of freedom because the number of parameters to be estimated is equal to the number of cells.

- Because it exactly reproduces the observations it has no explanatory power.
• It is therefore usually uninteresting but sometimes it can be useful to look at its parameter estimates.

• Later in the course we will look at ways of constraining the $\tau_{ij}$ term to fit restricted forms of association that retain some degrees of freedom, and explanatory power.
The $\chi^2$ distribution

- The $\chi^2$ distribution is a theoretical probability distribution like the normal distribution, taking one parameter for degrees-of-freedom.

- It’s non-symmetrical and starts at zero.

- The higher the degrees of freedom the flatter and more rightward the distribution.
The $\chi^2$ distribution

Degrees of freedom: 4, 7 and 11
Probabilities are areas under the curve

- Areas under the curve correspond to the probability of a value falling into the relevant interval.

- For instance, 50% of the time, a $\chi^2$ variable with DF of 5 will be in the range 0–4.35. 99% of the time the value will fall under 15.09, meaning that there is only a 1% chance of getting a random $\chi^2$ variable with 5 degrees of freedom above this level.
Areas under the curve and probabilities
Calculating probabilities

- Keep a spreadsheet on hand to calculate $\chi^2$ probabilities. For instance, in Excel, the CHIDIST function gives you the area under the curve to the right of $X^2$: enter "=CHIDIST(dev, df)" in a cell to get the probability of getting a $X^2$ at least as large as $dev$ for $df$ degrees of freedom.
Fitting models in SPSS

- SPSS (version 6.1 and later) uses the command `GENLOG` to fit loglinear models.

- There are two important parts to the command:
  - A definition of the table being modelled:
    
    ```
    genlog housing vote fsex
    ```
  - And a definition of the model being fitted to it:
    
    ```
    /design housing vote fsex vote by fsex.
    ```

- This example defines the table to be analysed as the `housing by vote` table, and fits a model to it which contains the three ‘main effects’ and the `vote*fsex` interaction:
genlog housing vote fsex  
/design housing vote fsex vote by fsex.

- It produces an intimidating amount of output, including graphs. Much of it is useful but for the moment we will ignore most of it, except the model fit statistics (look for Likelihood Ratio towards the end of the output.)
Goodness of fit

- The standard test for 2-way independence uses the Pearson $X^2$ with expected values defined as $E_{ij} = \frac{n_i \cdot n_j}{n_{++}}$.

- This also applies to models other than independence: we can use the expected value under any model.

- In loglinear modelling estimation is by maximum likelihood methods which generate their own analogue of $X^2$, the ‘deviance’ or $G^2$:

$$G^2 = 2\Sigma O_{ij} \log \frac{O_{ij}}{E_{ij}} = -2[L_m - L_s]$$
• This is usually very close in value to Pearson’s $X^2$, but has some mathematical properties which make it preferable.

• The interpretation is the same: if the $G^2$ is higher than the critical level of the theoretical $\chi^2$ distribution, the model does not fit.
Comparing models

- An alternative way of examining how well models fit is to compare nested models.
- One model is 'nested' within another when the other has all the parameters it has, plus some extra ones.
- When two models are nested, we test the effect on fit of adding the extra parameters as a block, to test whether the bigger model is a better fit than the one nested within it.
- We do this by taking the difference in $G^2$, and the difference in degrees of freedom, and comparing with the theoretical $\chi^2$ distribution. When the reduction in $G^2$ is big relative to the reduction in degrees of freedom, the more complex model is an improvement over the simpler one.
Calculating model fit

- To calculate the fit of a model, consider the area under the curve to the right of $G^2$ for the appropriate degrees of freedom ($\text{CHIDIST}(G^2, df)$): the smaller the $G^2$ the better.
  - ‘What is the chance of getting at least this much deviance if the model is correct?’
  - Small $G^2$ ⇒ moderate/high probability.

- To calculate the significance of an improvement, consider the area to the right of the $\Delta G^2$ for $\Delta df$ ($\text{CHIDIST}(\Delta G^2, \Delta df)$): the bigger the drop in $G^2$ the better.
  - ‘What is the chance of getting this much reduction in deviance if the second model is not actually better than the first?’
  - Big $\Delta G^2$ means small probability (high significance).
The fit of a single model can also be considered as how much the saturated model could improve on it (saturated model has $G^2 = 0$).

- If the model has a small deviance ($G^2$) reducing it to zero will only constitute a small improvement.
Loglinear Analysis

Parameter estimates

- When we fit models we get a great deal of output (from SPSS in particular) and most of it is useful sometimes. We will consider three aspects in particular:
  - overall fit
  - fitted values and residuals, and
  - parameter estimates.

- We will consider parameter estimates first.
For convenience, consider a 1-dimensional table. This allows two simple models:

$$\log F_i = \lambda_0$$

and

$$\log F_i = \lambda_0 + \lambda_i$$

- The former is the grand-mean or equiprobability model
- the latter is the saturated model, because we have only one dimension.

These are mathematical statements of the model: what do they look like in principle?

The grand mean model estimates just a constant, the intercept in a poisson regression. This constant is to be interpreted as the log of the expected count in all \( I \) cells.
• The saturated model is written as if it has \( I + 1 \) terms, where \( I \) is the number of categories the variable has. Since we have only \( I \) pieces of information we cannot determine all \( I + 1 \) values, so we impose a simple constraint.

• Other methods are possible, but the typical approach is to set one parameter, often that for the first category, to zero. All other parameters consequently take a relative form: the effect of being in category \( i \) relative to category 1.

• SPSS uses the last category as the reference category.
Thus for all (say 5) cells in the 1-D table we have the following arrangement of estimated effects:

\[
\begin{align*}
\log F_1 &= \lambda_0 + \lambda_1 = \lambda_0 \\
\log F_2 &= \lambda_0 + \lambda_2 \\
\log F_3 &= \lambda_0 + \lambda_3 \\
\log F_4 &= \lambda_0 + \lambda_4 \\
\log F_5 &= \lambda_0 + \lambda_5
\end{align*}
\]
In the case of the housing variable in smalltab.sav the saturated model generates the following parameter estimates:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>SE</th>
<th>Z-value</th>
<th>Asymptotic 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.6073</td>
<td>.0367</td>
<td>179.80</td>
<td>6.54 to 6.68</td>
</tr>
<tr>
<td>2</td>
<td>1.9677</td>
<td>.0392</td>
<td>50.15</td>
<td>1.89 to 2.04</td>
</tr>
<tr>
<td>3</td>
<td>.5192</td>
<td>.0464</td>
<td>11.19</td>
<td>.43 to .61</td>
</tr>
<tr>
<td>4</td>
<td>.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This generates the following results:

Housing = 1: log freq = 6.6073 + 1.9677 = 8.575 => freq = 5297
Housing = 2: log freq = 6.6073 + .5192 = 7.127 => freq = 1244
Housing = 3: log freq = 6.6073 + .0000 = 6.607 => freq = 740

which correspond to the observed values (with a little bit of rounding error).
Parameter estimates for higher dimensions

- For higher dimensions the story is similar. Consider independence in two dimensions:

\[
\log F_{ij} = \lambda_0 + \lambda_i^A + \lambda_j^B
\]

- Remembering that \( \lambda_1 = 0 \) for rows and columns:

\[
\begin{array}{ccc}
1 & 2 & 3 \\
1 & \lambda_0 & \lambda_0 + \lambda_2^B & \lambda_0 + \lambda_3^B \\
2 & \lambda_0 + \lambda_3^A & \lambda_0 + \lambda_2^A + \lambda_2^B & \lambda_0 + \lambda_2^A + \lambda_3^B \\
3 & \lambda_0 + \lambda_2^A & \lambda_0 + \lambda_3^A + \lambda_2^B & \lambda_0 + \lambda_3^A + \lambda_3^B \\
\end{array}
\]
Interactions

- Interactions work in a similar manner, but the way they are constrained is a little more complicated.

- For a 2-way interaction, \( \lambda_{ij}^{AB} \) there is a non-zero effect for every \( ij \) combination, except where \( i = 1 \) or \( j = 1 \) (or whatever the reference category is):

\[
\lambda_{i1}^{AB} = \lambda_{1j}^{AB} = 0
\]

- This means the saturated model looks like this:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \lambda_0 )</td>
<td>( \lambda_0 + \lambda_2^B )</td>
<td>( \lambda_0 + \lambda_3^B )</td>
</tr>
<tr>
<td>2</td>
<td>( \lambda_0 + \lambda_2^A )</td>
<td>( \lambda_0 + \lambda_2^A + \lambda_2^B + \lambda_{22}^{AB} )</td>
<td>( \lambda_0 + \lambda_2^A + \lambda_2^B + \lambda_{23}^{AB} )</td>
</tr>
<tr>
<td>3</td>
<td>( \lambda_0 + \lambda_3^A )</td>
<td>( \lambda_0 + \lambda_3^A + \lambda_2^B + \lambda_{32}^{AB} )</td>
<td>( \lambda_0 + \lambda_3^A + \lambda_3^B + \lambda_{33}^{AB} )</td>
</tr>
</tbody>
</table>
Tables with more than 2 dimensions

The structure of 3-D tables

- In 2-way tables, independence means the cell counts are given by the marginals.

- In 3-way tables, we have three pairs of variables (A*B, A*C and B*C), but a 3-way tabulation (A*B*C): more complex:
  - We could have overall 3-way independence: knowing the 1-D A, B and C margins (frequencies) is enough to predict the cell counts in the 3-way table.
- We could have any one, two or all three of the 2-way interactions – pairwise associations.
- We could need the A*B*C interaction, if the association between e.g., A and B differs across the categories of C (if this is true then in general the association between A and C differs across B, etc.).

- The equation for a saturated model of a three-way table is

\[ F_{ijk} = \tau_0 \tau_i^A \tau_j^B \tau_k^C \tau_{ij}^{AB} \tau_{ik}^{AC} \tau_{jk}^{BC} \tau_{ijk}^{ABC} \]

- A grand mean;
- Three ‘main’ or first-order effects;
- Three two-way interactions;
- One three-way interaction.
Interpreting 3-D models

- In a 3-way table, of say religion by party by cohort, the first order terms mean that there are different numbers in the categories of the marginal distributions of each variable. For instance, cohort size is not a constant, different parties have different levels of support, etc.

- The two-way terms, e.g., religion by cohort mean that there is an association between religion and cohort: the distribution of religion is different in the different cohorts.

- Having more than one 2-way term means that there is more than one 2-way association, even when you take the other(s) into account.

- You may find an apparent association between say, religion and
party (e.g., if you created the religion by party table and tested for independence), but this is really due to associations between religion and cohort and between party and cohort. That is, you may have marginal association but not conditional association.

- This can arise by many means: for instance, assume there is no mechanism associating religion and party, but that the distribution of these states changes over time (across cohort). Let's say there is an increase in religious non-participation and an entirely unrelated rise in the vote for the Social Democratic Party. Early cohorts will have lower non-participation and lower SDP, later will have higher. In the religion by party marginal table this will look like association, but in each cohort-specific religion by party subtable there will be no association.
Conditional vs Marginal Association: an example

- This is a $2 \times 2 \times 2$ table of fictional data, say religion by party at two different times. In each period there is no association between religion and party (perfect independence) but the distribution of religion and party change between the periods:

<table>
<thead>
<tr>
<th>Religion</th>
<th>Period 1</th>
<th></th>
<th>Period 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Party A</td>
<td>Party B</td>
<td>Total</td>
<td>Party A</td>
</tr>
<tr>
<td>I</td>
<td>20</td>
<td>30</td>
<td>50</td>
<td>72</td>
</tr>
<tr>
<td>II</td>
<td>20</td>
<td>30</td>
<td>50</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>60</td>
<td>100</td>
<td>80</td>
</tr>
</tbody>
</table>

Period 1 OR: $\frac{20}{30} / \frac{20}{30} = 1$

Period 2 OR: $\frac{72}{18} / \frac{8}{2} = 1$
Loglinear Analysis Unit 3

Religion by party marginal

<table>
<thead>
<tr>
<th>Religion</th>
<th>Party</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>I</td>
<td>92</td>
<td>48</td>
</tr>
<tr>
<td>II</td>
<td>28</td>
<td>32</td>
</tr>
<tr>
<td>Total</td>
<td>120</td>
<td>80</td>
</tr>
</tbody>
</table>

(\textit{observed}): Religion Party Total

<table>
<thead>
<tr>
<th>Religion</th>
<th>Party</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>I</td>
<td>84</td>
<td>56</td>
</tr>
<tr>
<td>II</td>
<td>36</td>
<td>24</td>
</tr>
<tr>
<td>Total</td>
<td>120</td>
<td>80</td>
</tr>
</tbody>
</table>

(\textit{expected}): Religion Party Total

Pearson’s $X^2$ is 6.35 for 1 df ($p = 0.012$).

Observed OR is 2.19.

- The marginal table shows association.

- But when you control for the association between party and period and religion and period the association disappears.
If we fit loglinear models to this data we get the following results (R is religion, P is party, T is time period):

<table>
<thead>
<tr>
<th>Model</th>
<th>$G^2$</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>R+P+T</td>
<td>75.2215</td>
<td>4</td>
</tr>
<tr>
<td>R+P+T+R*T</td>
<td>34.5218</td>
<td>3</td>
</tr>
<tr>
<td>R+P+T+P*T</td>
<td>40.6997</td>
<td>3</td>
</tr>
<tr>
<td>R+P+T+R*P</td>
<td>68.9432</td>
<td>3</td>
</tr>
<tr>
<td>R+P+T+R<em>T+P</em>T</td>
<td>.0000</td>
<td>2</td>
</tr>
</tbody>
</table>

The R+P+T model doesn’t fit well but including either the religion by time or party by time 2-way interactions improve the fit by c50%.

Adding the religion by party term reduces the deviance too, though not quite as much.
- The model with both the religion by time and party by time interactions fits perfectly: there is no religion by party association once you control for the associations with time.

- This is made-up data: in real sample data the fit wouldn’t be perfect. Instead, adding the religion by time interaction to this model would not improve the fit significantly.
Death penalty example

- In the practicals a well known data set is/was modelled:

<table>
<thead>
<tr>
<th>Victim’s race</th>
<th>Defendant’s race</th>
<th>Death Penalty</th>
<th>%age yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>White</td>
<td>53</td>
<td>414</td>
</tr>
<tr>
<td>Black</td>
<td>White</td>
<td>11</td>
<td>37</td>
</tr>
<tr>
<td>Black</td>
<td>White</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>Black</td>
<td>Black</td>
<td>4</td>
<td>139</td>
</tr>
<tr>
<td>Total</td>
<td>White</td>
<td>53</td>
<td>430</td>
</tr>
<tr>
<td>Total</td>
<td>Black</td>
<td>15</td>
<td>176</td>
</tr>
</tbody>
</table>

(Source: Agresti p. 54ff.)

- A $2 \times 2 \times 2$ table of defendant’s race, victim’s race and sentence for US murder cases.
Careful inspection of the “% yes” column tells us much of the story:

- On the whole, black defendants are less likely than white to be sentenced to death.
- However, the percentages conditional on the victim’s race are quite different: for black or white victims taken separately, black defendants have a higher rate of receiving the death penalty.

This is another case of conditional and marginal association diverging, and the explanation lies in other two 2-way associations:

- Black killers have more black victims and white killers have more white victims;
- Killers of white victims are more likely to receive the death penalty.
This emerges directly in the modelling:

1. Adding \texttt{drace by verdict} does not improve on the 3-way independence model ($G^2$ falls from 402.84 to 401.30, for 1 df: $p = 0.22$) (very different from 2-D table!)
2. Adding the \texttt{vrace by verdict} term does improve ($G^2$ falls to 385.96)
3. So does adding \texttt{vrace by drace} (dramatically), and the model with \texttt{vrace by drace} and \texttt{vrace by verdict} approaches a good fit ($p=0.07$).
4. However, if we now add the \texttt{drace by verdict} term it improves the fit even more, raising the fit to $p=0.5$. 
The parameter estimates are informative:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>3.96685</td>
<td>0.137378</td>
</tr>
<tr>
<td>Penalty (No)</td>
<td>2.05946</td>
<td>0.145845</td>
</tr>
<tr>
<td>VicRace (B)</td>
<td>-5.66961</td>
<td>0.645704</td>
</tr>
<tr>
<td>DefRace (B)</td>
<td>-1.55253</td>
<td>0.326164</td>
</tr>
<tr>
<td>Penalty (No).DefRace (B)</td>
<td>-0.867797</td>
<td>0.367062</td>
</tr>
<tr>
<td>VicRace (B).DefRace (B)</td>
<td>4.59497</td>
<td>0.313523</td>
</tr>
<tr>
<td>Penalty (No).VicRace (B)</td>
<td>2.40444</td>
<td>0.600437</td>
</tr>
</tbody>
</table>

- Two very large positive terms linking white victims to the death penalty and linking defendant’s and victim’s race
- One smaller, though significant, negative term indicating black defendants are more likely to receive the death penalty, after controlling for the confounding effect of the other two relationships.
Careful inspection of the table would have told the same story but the modelling has two advantages:

- a framework for formal hypothesis testing
- a tool valid for tables too large to analyse by hand.
Odds Ratios and what they mean

- We understand the definition of the odds ratio, and how it is one way of parameterising association: ORs plus margins give you the whole table.

- What of interpreting odds ratios in context? What does an OR mean?

- In a simple table: white victim panel of the death penalty data:

<table>
<thead>
<tr>
<th>Defendent’s Race</th>
<th>Death Penalty</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td></td>
<td>53</td>
<td>414</td>
</tr>
<tr>
<td>Black</td>
<td></td>
<td>11</td>
<td>37</td>
</tr>
</tbody>
</table>

(Source: Agresti p. 54ff.)
• The odds ratio for whites versus blacks getting versus not getting the death penalty is \( \frac{53/414}{11/37} = 0.43 \). That is, the odds of a white person convicted of killing a white person getting the death penalty are only 43% of those of a black person.

• How do we map this arithmetic concept onto social-science or everyday concepts? Consider the general case of an OR describing the relationship between two types of people and two outcomes, one of which is preferable, \( e.g. \), people from two rich versus poor families, completing or not completing second level-education.
Here we have two tables with the same odds ratios but different margins:

<table>
<thead>
<tr>
<th>Education</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period 1</td>
<td>OR=4.2</td>
<td></td>
</tr>
<tr>
<td>Rich</td>
<td>12</td>
<td>188</td>
</tr>
<tr>
<td>Poor</td>
<td>3</td>
<td>198</td>
</tr>
<tr>
<td>Period 2</td>
<td>OR=4.2</td>
<td></td>
</tr>
<tr>
<td>Rich</td>
<td>103</td>
<td>97</td>
</tr>
<tr>
<td>Poor</td>
<td>40</td>
<td>160</td>
</tr>
</tbody>
</table>

What has happened between the two periods? The provision of education has increased enormously, and both rich and poor people have taken advantage of it. However, the rich group have much more of the extra places, so that their odds of completing have risen by as much as have the odds for poor people.
- So, if we’re concerned with social justice, is this progress or not? What does it mean in social-justice terms that the OR stays constant?

- At one level there is unmistakable progress: more education all round. But the relative effect of family background has remained constant.

- The odds ratio measures what might be considered equality of opportunity: as it deviates from one there is inequality, and if its size remains the same the inequality of opportunity remains the same, even though the ‘quantity’ of opportunity changes.

- This is an artificial example, but it is not unrealistic. In terms of loglinear models, it is the case that arises when a model like $B*O + B*P + O*P$ fits well (Background, Outcome, Period) suggesting that there is no 3-way $B*O*P$ interaction.

- That is, that the $B*O$ interaction is not significantly different at the different levels of $P$. 

That is, the odds-ratios that the $B^*O$ interaction term represents are constant despite the changes in the distribution of Outcome ($O^*P$) and changes in the distribution of Background ($B^*P$).

**Processes generating odds-ratios**

- It may seem surprising that we often find models that fit like this: big changes in Background and Outcome over Period are not accompanied by a change in the association between Background and Outcome.

- However, some simple processes of allocation can give rise to near constant odds ratios.

- Assume outcomes are decided according to a threshold measurement:
anyone over a certain score gets the good outcome. Increases in the supply of the outcome cause a lowering of the threshold.

- Assume we have two subpopulations with different distributions of the score variable: let’s say same standard deviation but different mean.

- The odds ratio is defined by the intersection of the distributions and the threshold.

- What happens to the OR as we increase the supply?

- This depends on the distribution: if it is a logistic distribution (very like the normal distribution with slightly fatter tails, but easier to deal with mathematically), the odds ratio is constant as long as the difference in means is constant.
• If the distribution is normal, the odds ratio changes quite slowly: a large change in the supply/threshold may have little effect on the odds ratio.

• The odds ratio will be most close to equal the closer the threshold is to the means of the distributions: it has a shallow U-shaped curve.
• In summary, two points:
  – It’s not necessarily surprising when odds ratios are relatively stable despite big marginal change.
  – Odds ratios *may* directly relate to simple processes generating the data: they are not necessarily statistical abstractions.
Fitting models to large tables

- As the number of dimensions increase the number of possible models increases rapidly.

- It is no longer practical to examine them all, so we need a way of searching efficiently.

- We could consider automatic stepwise selection as SPSS will do by default if you go Statistics⇒Loglinear⇒Model Search (this invokes the `HILOGLINEAR` command. This is not a good idea: to quote from the Stata mailing list (for fuller detail see http://teaching.sociology.ul.ie/~brendan/CDA/stepwise.text).
1. It yields R-squared values that are badly biased high.
2. The F and chi-squared tests quoted next to each variable on the printout do not have the claimed distribution.
3. The method yields confidence intervals for effects and predicted values that are falsely narrow (See Altman and Anderson Stat in Med).
4. It yields P-values that do not have the proper meaning and the proper correction for them is a very difficult problem.
5. It gives biased regression coefficients that need shrinkage (the coefficients for remaining variables are too large; see Tibshirani, 1996).
6. It has severe problems in the presence of collinearity.
7. It is based on methods (e.g. F tests for nested models) that were intended to be used to test pre-specified hypotheses.
8. Increasing the sample size doesn’t help very much (see Derksen and Keselman).
9. It allows us to not think about the problem.
10. It uses a lot of paper.
Note that "all possible subsets" regression does not solve any of these problems.

- On the other hand, non-automatic stepwise model building is sometimes a good idea *e.g.*, :

  - **Forward:**
    1. Begin with first order terms one at a time, and include all which seem important.
    2. Add 2-way interactions, first one at a time and then cumulatively, including all which significantly reduce the deviance. Pay a lot of attention to what seems very significant, and what is marginal or non-significant.
    3. If you have enough dimensions, continue from your favoured two-way model adding three-way terms that are possible (*i.e.*, their hierarchical descendents are already in the model), and so on.
4. Pay attention to the possibility that a higher-order interaction might be significant despite its lower-order interaction being insignificant.
   – Backward
     1. Pick a model with all, say, 3-way interactions.
     2. See what terms you can eliminate without raising the deviance.
     – But always pay attention to what the addition or removal suggests about the model.

- Another approach is to start with a ‘minimum-interest’ model, which contains terms to take account of all that you are not interested in: for instance, if you really want to know whether sentencing policy differs according to defendant's race:

   – fit the model \( \text{drace} + \text{vrace} + \text{verdict} + \text{vrace*verdict} + \text{vrace*drace} \).
   – Add the term of interest, \( \text{drace*verdict} \), and test for improvement.
– Then start refining (adding or deleting terms) to get rid of unnecessary terms or add others that improve the fit.
– If you arrive at a different model this way, retest your term of interest.
Residuals

- A model fits well if the expected and observed counts approximately coincide.

- We normally use an overall measure of this match, such as $G^2$ or $X^2$.

- However, this is based on cell-wise comparisons of fitted and observed values, and we can look at the **residuals**.

- Simple residuals are defined as:

  $$R_{ij} = O_{ij} - E_{ij}$$

  That is, if the model underfits that cell there is a positive residual.
The size of residuals will be related to the cell size, so they are usually scaled. A ‘Pearson Residual’ is defined as

\[ e_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}} \]

Notice that \( X^2 \) is thus equal to \( \Sigma e^2 \).

We can also partition the \( G^2 \) across the cells to give what are called deviance residuals.

If the model fits, these residuals have a large-sample standard normal distribution so residuals with an absolute value much greater than 2 are worthy of attention.
• But remember that in a big table there are many residuals so we expect to see a few in the tails of the probability distribution.

• In practice we will use adjusted residuals. These are equivalent to Pearson residuals scaled by $\sqrt{1/(1-h_i)}$ where $h_i$ is the leverage of that case (cell):

$$e_{ij}^{adj} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}(1-h_{ij})}}$$

• Examination of residuals can tell us a lot about how the model fails to fit. SPSS normally displays observed and expected counts, and raw, adjusted and deviance residuals in the GENLOG output. It can also save them as a variable for examination elsewhere.
fsex housing vote
/MODEL=POISSON
/DESIGN fsex housing vote
/SAVE ADJRESID .

This will create a variable \texttt{ADJ\_n} where \texttt{n} is a number containing the adjusted residuals.

- Using a macro available in \url{http://teaching.sociology.ul.ie/~brendan/CDA/macros/} these can be inspected in tables:

\begin{verbatim}
  temp.
  select if fsex=1.
  tabu housing vote adj_1.
  temp.
\end{verbatim}
select if fsex=2.
tabu housing vote adj_1.

Something similar can be achieved using the menu interface, via Statistics⇒Custom Tables⇒Basic Tables; the macro handles the case for two classifying variables.
Men

<table>
<thead>
<tr>
<th>VOTE</th>
<th>Conservat</th>
<th>Labour</th>
<th>Lib-dem</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HOUSING</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Owned/mortgage</td>
<td>Adjusted</td>
<td>Residual</td>
<td>8.65</td>
<td>-2.87</td>
</tr>
<tr>
<td>LA etc rented</td>
<td>Adjusted</td>
<td>Residual</td>
<td>-10.05</td>
<td>4.81</td>
</tr>
<tr>
<td>Private rented</td>
<td>Adjusted</td>
<td>Residual</td>
<td>-18</td>
<td>.79</td>
</tr>
</tbody>
</table>

Women
| HOUSING          | | | | |
| Owned/mortgage   | | | | |
| Adjusted         | | | | |
| Residual         | 3.62 | -8.63 | 4.57 | -2.09 |
| LA etc rented    | | | | |
| Adjusted         | | | | |
| Residual         | -7.51 | 12.35 | -.85 | 2.28 |
| Private rented   | | | | |
| Adjusted         | | | | |
| Residual         | -2.92 | .36 | .80 | 1.13 |
Loglinear models for square tables

Square Tables

- Square tables have
  - 2 dimensions
  - the same sets of categories on each dimension

- Square tables arise very often for many reasons
  - measuring the same thing at two different times, *e.g.*, vote at successive elections
  - comparing two judges’ or raters’ opinions

Brendan Halpin, Sociology Group, Dept of Government & Society, University of Limerick
Swiss Summer School, Lugano August 2002
– comparing the same concept measured on linked objects, *e.g.*, mobility tables linking parents’ class (origin) to respondents’ class (destination)
– any before-and-after comparison.
The squareness often gives these tables special structure which we can take advantage of with some specialised models.

Often the diagonal is overpopulated:

- the raters agree
- the respondent is in the same state before and after
- the child has the same social class as the parent

Sometimes the distribution of the two variables is (approximately) the same: **marginal homogeneity** may arise e.g., if two raters are trying to get the ‘correct’ proportions in each category.
• Sometimes there is **symmetry**: moves from $A_i$ to $B_j$ are as likely as moves from $A_j$ to $B_i$ and the table is symmetrical around the main diagonal.

• However, if there is not marginal homogeneity
  - e.g., the second rater is more prone to grade as ‘good’ and ‘very good’
  - the class distribution has changed substantially while the respondent was growing up
  - the government got thrown out

there cannot be true symmetry: the model that fits ‘as much symmetry as can be expected’ is the **quasi-symmetry** model.
Fitting models to square tables

- Square tables have only two dimensions, so only the following basic models can be fitted:
  - Grand mean: \( F_{ij} = \tau_0 \)
  - Equiprobability on one dimension: \( F_{ij} = \tau_0 \tau_i \) and \( F_{ij} = \tau_0 \tau_j \)
  - Independence: \( F_{ij} = \tau_0 \tau_i \tau_j \)
  - Association: \( F_{ij} = \tau_0 \tau_i \tau_j \tau_{ij} \)

- The association model is the saturated model, and the independence model is usually a poor fit for square tables: we need to model a reduced form of the association.
• Simplest case is the **mover–stayer** model: there are two sorts of cases, those who move, for whom there is independence (all destinations equally likely) and those who stay.

• This is also known as quasi-independence because it means independence everywhere except the diagonal.

**NB:** some movers may end up in the same place by chance, so the diagonal will include both movers and stayers. Off the diagonal there are only movers, and there is independence off the diagonal.
To fit: drop the diagonal and fit independence on the rest of the table:

1. compute a variable equal to zero if the case is on the diagonal and equal to one if not:
   
   compute offdiag = VAR1<>VAR2.
   
   where VAR1 and VAR2 are the classifying variables.

2. **Weight** the table according to this variable:

   genlog VAR1 VAR2
   
   /cstructure=offdiag
   
   . . .
   
   or Statistics⇒Loglinear⇒General and put offdiag into the Cell Structure box.

3. Fit it:

   genlog VAR1 VAR2
   
   /cstructure=offdiag
   /design = VAR1 VAR2.
• Why is it called ‘mover–stayer’?
  – It assumes there are two types of people, those who have a non-zero probability of moving, and those who will not move. By weighting out the diagonal, we exclude the stayers, and can then test the movers for independence.

• Weighting out the diagonal is equivalent to fitting it exactly by including cell-specific terms. Weighting out reduces the degrees of freedom by dropping data points; fitting cell-specific terms uses up the same degrees of freedom. The deviance is the same in both cases because the cells, if fitted exactly, have residuals of zero.

• This fact is useful for comparing pairs of models (i.e., weighting out is equivalent to certain parameters being included so we can establish whether certain pairs of models are nested).
• We will get a clearer idea what fitting cell-specific terms means in the following sections.
Symmetry

- Symmetry implies \( \pi_{ij} = \pi_{ji} \), that the probability of changing from \( i \) to \( j \) is the same as that from \( j \) to \( i \). This can be expressed as \( F_{ij} = \tau_0 \tau_i \tau_j \tau_{ij} \) where \( \tau_i = \tau_j \) and \( \tau_{ij} = \tau_{ji} \). One way of fitting this model is to create a new variable with a different value for each cell in one half of the table, with the other half of the table being a mirror image: that is, for each pair of cells \( c_{ij}, c_{ji} \) there is a distinct value of the new variable.

- Since for the diagonal there is only one cell for each \( \pi_{ii} \) it is usually weighted out.
• For a $6 \times 6$ table the symmetry variable looks like:

\[
\begin{array}{ccccc}
- & 1 & 2 & 3 & 4 & 5 \\
1 & - & 6 & 7 & 8 & 9 \\
2 & 6 & - & 10 & 11 & 12 \\
3 & 7 & 10 & - & 13 & 14 \\
4 & 8 & 11 & 13 & - & 15 \\
5 & 9 & 12 & 14 & 15 & - \\
\end{array}
\]

• The diagonal values are represented by ‘-’ because they are weighted out in the analysis.

• We can use the macro `tabu` from [http://teaching.sociology.ul.ie/~brendan/CDA/categ-macs.sps](http://teaching.sociology.ul.ie/~brendan/CDA/categ-macs.sps) to confirm the appearance of the symmetry variable:
tabu VAR1 VAR2 symm.

This will lay it out as in the table above.
• One way of generating an design variable like this in SPSS is the following (for var1 by var2, each with 5 categories):

```spss
compute #c=1.
loop #i = 1 to 5.
    . loop #j = #i+1 to 5.
    . if var1=#i and var2=#j sym = #c.
    . if var1=#j and var2=#i sym = #c.
    . compute #c=#c+1.
    . end loop.
end loop.
if var1=var2 sym=1.
```

• This will work for individual level data and for grouped data; with grouped data it is also possible to enter it by hand.
Code to do this as a macro is also available in categ-macs.sps: the syntax is

```
makesym VAR1 VAR2 n NEWVAR
```

this takes the two classifying variables VAR1 and VAR2, each with n categories, and creates a variable called NEWVAR with the symmetry pattern in it.

What does this model say about the data?

- About the diagonal cells, nothing: since they are weighted out, they can have any value.
- Off the diagonal \( \log F_{ij} = \log F_{ji} = \lambda_0 + \sigma_{ij} \).
- Row \( i \) and Column \( i \) will have the same totals: marginal homogeneity.
• This is a somewhat strange model: if we consider the symmetry term as a restricted two-way interaction, it is almost like fitting the interaction term without its first-order children, though the interaction term has consequences for the row and column totals (they match).

• It has (at least) one similarity with the mover–stayer model: it drops the diagonal entirely.
• Fitting this model takes another step, and another macro.
  – Having used `makesym` to create the symmetry variable, we need to make dummy variables from it.
  – This is because SPSS only recognises the classifying variables as categorical; extra variables go in to the model as covariates, which are expected to be quantitative.
  – The macro `factor` makes a set of dummy variables
    - `factor symm s 5` creates variables `s1` to `s5` where `s3=0` unless `symm=3 when s3=1, etc.`
  – Covariates are entered in the first part of the `GENLOG` command:
    - `genlog avote evote with s2 s3 s4 s5`.
  – Note the `with` separating classifying variables from the covariates.
  – Note that `s1` is not included: this is the reference category.
  – The `/design` statement contains only the `n – 1` dummy variables, no ‘real’ variable.
The symmetry model is fitted on a $4 \times 4$ table thus:

- compute offdiag = avote<>evote.
- makesym avote evote 4 symm .
- factor symm s 6.
- genlog avote evote with s2 s3 s4 s5 s6
  /cstructure = offdiag
  /print=est/plot=none
  /design= s2 s3 s4 s5 s6
  /save=adjresid.

- The last subcommand /save=adjresid is not essential but it is always good to look at the pattern of residuals:
  tabu avote evote adj_1.
Quasi-Symmetry

- Where marginal homogeneity does not hold, true symmetry is not possible: when a category grows, it is not generally possible for flows out of the category to equal flows into the category, *e.g.*, if there is a swing to Labour this means there are more Conservative to Labour switchers than Labour to Conservative switchers.

- There are models which directly test the hypothesis of marginal homogeneity, but they are not loglinear models and require a different approach to fit (possible with an iterated GLM).

- Where marginal homogeneity does not hold the symmetry model will not fit, but we can test for **quasi-symmetry**, that is, that there is as much symmetry as possible given the changing marginal distributions.
To fit this, simply add the row and column variables to the symmetry model:

```
genlog avote evote with s2 s3 s4 s5 s6
   /cstructure = offdiag
   /print=est/plot=none
   /design= avote evote s2 s3 s4 s5 s6
```

We can interpret this as the independence model with a overlay that says the departure from independence is symmetrical. For instance, after taking account of the fall in the Conservative vote and the rise in the Labour vote (and the net Conservative→Labour flow arising from this: $F_{cons,lab} > F_{lab,cons}$) there are equal (in the logs) flows from Con to Lab and from Lab to Con.
Weighting and cell-specific fitting

- Weighting out the diagonal is exactly equivalent to fitting it exactly.

- To do this, add new values to the `symm` variable, one for each diagonal cell, and fit without the `/cstructure` command:

  ```
  compute symm2=symm.
  if avote=evote symm2=avote+6.
  factor symm2 d 10.
  
  genlog avote evote with d2 d3 d4 d5 d6 d7 d8 d9 d10
  /print=est/plot=none
  /design=avote evote d2 d3 d4 d5 d6 d7 d8 d9 d10.
  ```

- The $G^2$ and degrees of freedom will be the same, and the residuals will be
identical (diagonal residuals will be either 0.0 or missing).

- The parameter estimates will be different because there is a different set of parameters (but they combine to the same predicted values).
Other models

- Using this technique of partitioning the two-way association by fitting an extra variable to the model (without making this variable a classifying variable!) we can fit a lot of other models, some *ad hoc* and some expressing particular theoretical expectations.

- One is the ‘distance’ model: if we feel there is an order in the categories, such that moves to adjacent categories are more likely than moves to further categories we can create a distance variable to take account of this:

  compute dist = 1+abs(aopfama-eopfama).

  tabu aopfama eopfama dist.
To fit this we first make a set of dummy variables and include them as covariates, with the row and column variables:

```
factor dist ds 5 .
genlog aopfama eopfama with ds2 ds3 ds4 ds5
   /print=none/plot=none
   /design = aopfama eopfama ds2 ds3 ds4 ds5.
```

This implies that moving is more difficult the farther you have to go.
The parameter estimates for the dummy variables should confirm this by monotonically decreasing (the farther you are from the diagonal the lower the expected frequency).

- An example of an *ad hoc* usage would be to take account of a single cell that has a high residual. We can define a variable with the value 1 for that cell and 0 otherwise, and add it to the model. This fits the cell exactly and has the same effect of weighting out the cell.
A more efficient case is where you can group a number of cells and give them one parameter: then you’re saying those cells are under-/over-populated for a single reason. For instance, when tabulating opinion at two times, you may feel that there are some people with very strong opinions who will always opt for strongly agree or strongly disagree and other people with no connection between their opinion at both times. If this is the case, the top-left and bottom-right corners of the table will be overpopulated and a model that doesn’t take account of this will show positive residuals in these cells. We can fit a single term to take account of both:

```
compute corner=0.
if a=1 and e=1 corner = 1.
if a=5 and e=5 corner = 1.
```

It is important to realise that when we start looking at residuals and fitting models to take account of them the normal conditions of hypothesis testing
no longer hold. We are instead trying to model or ‘smooth’ the data, and we must be aware of the different nature of our findings.
Fitting and interpreting models

- The file modelvote.sps contains the code necessary to fit all these models to the \texttt{AVOTE} by \texttt{EVOTE} table from opinion.por data file (BHPS data on opinions and vote intention for 1991 and 1995).

Example output from this file is in modelvote.out. Corresponding files analysing a pair of opinion variables are in modelop.sps and modelop.out. This input file differs slightly from the vote example to take account of the greater number of categories.
## Vote intentions

<table>
<thead>
<tr>
<th>Model</th>
<th>Chi-Square</th>
<th>Degrees of Freedom</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>independence model</td>
<td>5382.4057</td>
<td>9</td>
<td>.0000</td>
</tr>
<tr>
<td>quasi-independence</td>
<td>14.1967</td>
<td>5</td>
<td>.0144</td>
</tr>
<tr>
<td>symmetry</td>
<td>5358.7148</td>
<td>10</td>
<td>.0000</td>
</tr>
<tr>
<td>quasi-symmetry</td>
<td>521.2890</td>
<td>6</td>
<td>.0000</td>
</tr>
<tr>
<td>distance</td>
<td>135.7280</td>
<td>6</td>
<td>.0000</td>
</tr>
</tbody>
</table>
Models for ordered categories and trends (i)

- A lot of variables we work with have some ordinality.
  - Levels of education, or strength of opinion.
  - Some are clearly quantitative, as counts (number of children) or as grouped continuous variables (age groups, income groups).
  - Some have specific values of a variable attached to each category: school size where categories are different schools, dose where categories are experimental groups, etc.

- We can model these as nominal and get sensible results but it is more efficient to take account of the ordinality: we can make better fitting, simpler models and make stronger claims.
There are several different approaches to imposing ordinality:

- imposing a linear effect on a variable
- imposing a set on non-linear scores
- calculating a best-fit set of scores
- imposing true ordinality: steps

Imposing a linear effect simply says that one (or more) variables has an effect which increases proportionately from category to category.

$$\log F_{ij} = \lambda_0 + \lambda_j + \beta i$$

This model suggests that the effect of the row variable can be approximated by a linear term: There are $x$ in row 1, $x.e^\beta$ in row 2, $x.e^{2\beta}$ in row three and so on.
- If we try this on AOPFAMC by EOPFAMC like this:

compute alin = aopfamc.
genlog aopfamc eopfamc with alin
   /print=est/plot=none/design= eopfamc alin .

we get $G^2$ of 6604.7571 for 19 df.

- This model is a constrained version of independence which says that the $\lambda_i$ set of parameters can be approximated by a linear effect. The fit of independence is 1167.1117 for 16, much better: the row distribution is not well approximated by a linear effect in this table.

- If we look at the parameter estimates for the $\lambda_i$s we see why a linear effect (of 0.0646) a poor fit:
<table>
<thead>
<tr>
<th>i</th>
<th>p e</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.4948</td>
</tr>
<tr>
<td>2</td>
<td>1.4743</td>
</tr>
<tr>
<td>3</td>
<td>2.5569</td>
</tr>
<tr>
<td>4</td>
<td>1.9288</td>
</tr>
<tr>
<td>5</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
Linear terms become more interesting at higher orders. Here they impose linearity on the association. If we are classifying social class by opinion on welfare, it may be plausible that for a particular category the association effect has a linear form: very unlikely to agree strongly, very likely to disagree strongly, with a smooth gradient in between. Each category may have a different gradient.
Consider the **EVOTE by EOPFAMF** table (vote intention in 1995, by attitude on ‘Husband should earn, wife stay at home’).

<table>
<thead>
<tr>
<th>Count</th>
<th>Strongly Agree</th>
<th>Neither</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>agree</td>
<td>di</td>
<td>disagree</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EVOTE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>123</td>
<td>383</td>
<td>538</td>
<td>703</td>
<td>230</td>
</tr>
<tr>
<td>Consve</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>214</td>
<td>492</td>
<td>732</td>
<td>1172</td>
<td>613</td>
</tr>
<tr>
<td>Labo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>39</td>
<td>121</td>
<td>237</td>
<td>302</td>
<td>177</td>
</tr>
<tr>
<td>Lib b/SDP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>28</td>
<td>47</td>
<td>82</td>
<td>60</td>
</tr>
<tr>
<td>Othe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column</td>
<td>386</td>
<td>1024</td>
<td>1554</td>
<td>2259</td>
<td>1080</td>
</tr>
<tr>
<td>Total</td>
<td>6.1</td>
<td>16.2</td>
<td>24.7</td>
<td>35.8</td>
<td>17.1</td>
</tr>
</tbody>
</table>

Brendan Halpin, Sociology Group, Dept of Government & Society, University of Limerick
Swiss Summer School, Lugano August 2002
• Independence gives $G^2$ of 100.4906 for 12. If we fit a linear effect for opinion in association with vote, what happens?

• First, make a copy of the opinion variable. Then enter the copy as a covariate, and include it in the `/design`. By entering a copy as a covariate, SPSS treats the values 1, 2, 3, 4 and 5 as quantitative scores.

```
compute elin = eopfamf.
genlog evote eopfamf with elin
    /print=est/plot=none
    /design= eopfamf evote elin by evote.
```
$G^2$ falls to 42.8971 for 9 df, a big improvement (but still not a well fitting model). The interaction uses up only 3 df because we fit parameters only for $4 - 1$ categories of vote.

$$\log F_{ij} = \lambda_0 + \lambda_i + \lambda_j + \beta_{ji}$$

- The parameter estimates are worth looking at:

<table>
<thead>
<tr>
<th>EVOTE</th>
<th>ELIN</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.3307</td>
<td>0.0657</td>
</tr>
<tr>
<td>2</td>
<td>-0.1848</td>
<td>0.0646</td>
</tr>
<tr>
<td>3</td>
<td>-0.1337</td>
<td>0.0697</td>
</tr>
<tr>
<td>4</td>
<td>0.0000</td>
<td>.</td>
</tr>
</tbody>
</table>
• For conservatives, we see a decline as we move from agreement to disagreement that is larger than that for labour, with LibDems being even smaller, with the other category being zero.

• That is, after taking account of the margins, the association takes the form that Conservatives are least likely to agree, and ‘other’ voters most, with the other parties in between.
The Linear-by-Linear model

- An extension of this approach to two dimensions is the well-known ‘linear-by-linear association model’. (Agresti p182ff.)

- This takes the form

\[ \log F_{ij} = \lambda_0 + \lambda_i + \lambda_j + \beta_{ij} \]

- In this model the association is described by a single number, \( \beta \), which is multiplied by the product of the row number and column number:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>9</td>
<td>12</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>12</td>
<td>16</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>
• \( \beta \)'s sign is related to the direction of the association: if higher values of \( I \) are associated with higher values of \( J \), it will be positive, and vice versa.

• The model effectively means all 'adjacent' odds ratios have the same value, \( e^\beta \).

• Independence is given by \( \beta = 0, e^\beta = 1 \), i.e., the odds ratios are all 1.0.

• To fit this model we simply create a new variable that is the product of the column and row numbers:

\[
\text{compute assoc } = \text{aopfamc*eopfamc}.
\]

and add it to the model as a covariate:

\[
\text{genlog aopfamc eopfamc with assoc}
\]
\[
/\text{print=est/plot=none}
\]
\[
/\text{design= aopfamc eopfamc assoc}.
\]
Score models

- We can generalise these linear models by applying scores other than the row and column numbers.

- We may feel there is a smaller distance between ‘strongly agree’ and ‘agree’ than between ‘agree’ and ‘neutral’, and score them thus:
  - 1 strongly agree
  - 2 agree
  - 4 neutral
  - 6 disagree
  - 7 strongly disagree
- Alternatively, we may have quantitative information about the categories.
  - Medical experiments may have several groups of subjects given different doses of the treatment.
  - Categories may refer to entities about which we have other information.
  - They may be constructed from continuous variables such as income and we can use mid-point values etc.
- An example: pregnant mice are exposed to differing levels of a toxin and three outcomes are recorded: normal foetus, abnormal foetus, dead foetus:

<table>
<thead>
<tr>
<th>Dose</th>
<th>Normal</th>
<th>Abnormal</th>
<th>Dead</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00</td>
<td>281</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>62.50</td>
<td>225</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>125.0</td>
<td>283</td>
<td>7</td>
<td>22</td>
</tr>
<tr>
<td>250.0</td>
<td>202</td>
<td>59</td>
<td>38</td>
</tr>
<tr>
<td>500.0</td>
<td>9</td>
<td>132</td>
<td>144</td>
</tr>
</tbody>
</table>
Independence gives $G^2$ of 911.6904 for 8 df, a very poor fit with these residuals:

<table>
<thead>
<tr>
<th>Dose</th>
<th>Normal</th>
<th>Abnormal</th>
<th>Dead</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00</td>
<td>10.50</td>
<td>-7.58</td>
<td>-5.95</td>
</tr>
<tr>
<td>62.50</td>
<td>8.65</td>
<td>.</td>
<td>-4.34</td>
</tr>
<tr>
<td>125.0</td>
<td>9.13</td>
<td>-6.72</td>
<td>-5.06</td>
</tr>
<tr>
<td>250.0</td>
<td>-.90</td>
<td>3.30</td>
<td>-1.96</td>
</tr>
<tr>
<td>500.0</td>
<td>-27.30</td>
<td>17.71</td>
<td>17.34</td>
</tr>
</tbody>
</table>
There is obviously association between the toxin and bad outcomes, so let's fit a linear term for dose (1, 2, 3, 4, 5) in interaction with outcome. $G^2$ falls to 155.2920 for 5 df, a significant improvement.

The residuals are smaller but still many are too big:

<table>
<thead>
<tr>
<th>Dose</th>
<th>Normal</th>
<th>Abnormal</th>
<th>Dead</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00</td>
<td>-8.92</td>
<td>1.93</td>
<td>8.66</td>
</tr>
<tr>
<td>62.50</td>
<td>-3.76</td>
<td>.</td>
<td>3.76</td>
</tr>
<tr>
<td>125.0</td>
<td>3.97</td>
<td>-1.78</td>
<td>-3.32</td>
</tr>
<tr>
<td>250.0</td>
<td>6.94</td>
<td>1.08</td>
<td>-8.17</td>
</tr>
<tr>
<td>500.0</td>
<td>-10.69</td>
<td>-.69</td>
<td>11.37</td>
</tr>
</tbody>
</table>
- But we have quantitative information on dose: what if we fit this as row scores: $G^2$ of 48.9363 for 5 df.

<table>
<thead>
<tr>
<th>Dose</th>
<th>Normal</th>
<th>Abnormal</th>
<th>Dead</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>-3.15</td>
<td>-1.07</td>
<td>4.43</td>
</tr>
<tr>
<td>62.50</td>
<td>-2.82</td>
<td></td>
<td>2.82</td>
</tr>
<tr>
<td>125.0</td>
<td>1.66</td>
<td>-2.18</td>
<td>-.36</td>
</tr>
<tr>
<td>250.0</td>
<td>3.08</td>
<td>2.90</td>
<td>-5.70</td>
</tr>
<tr>
<td>500.0</td>
<td>-4.23</td>
<td>-2.80</td>
<td>6.69</td>
</tr>
</tbody>
</table>

- It may be that the effect of the toxin is non-linear in the dose: since the residuals for the 500mg/kg dose are so high, it may be that the effect increases faster than dose: we can use the square of dose to model this: $G^2$ of 37.1931 for 5 df, a somewhat better fit.
• Outcome can be considered ordinal as well (true ordinal rather than quantitative as dose). We can fit a variant of the linear-by-linear model, using the square of dose and the column index for outcome:

compute assoc = outcome*dose*dose.

genlog dosecat outcome with assoc
   /print=est/plot=none
   /design = dosecat outcome assoc.
This doesn’t fit well: $G^2$ is 288.4299 for 7 df. Compared with the less parsimonious model with the square of dose in interaction with outcome, it is not preferred. Looking back at the residuals from independence, this is not surprising:

<table>
<thead>
<tr>
<th>Dose</th>
<th>Normal</th>
<th>Abnormal</th>
<th>Dead</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00</td>
<td>10.50</td>
<td>-7.58</td>
<td>-5.95</td>
</tr>
<tr>
<td>62.50</td>
<td>8.65</td>
<td></td>
<td>-4.34</td>
</tr>
<tr>
<td>125.0</td>
<td>9.13</td>
<td>-6.72</td>
<td>-5.06</td>
</tr>
<tr>
<td>250.0</td>
<td>-.90</td>
<td>3.30</td>
<td>-1.96</td>
</tr>
<tr>
<td>500.0</td>
<td>-27.30</td>
<td>17.71</td>
<td>17.34</td>
</tr>
</tbody>
</table>
Log-multiplicative models

- Log-multiplicative models are analogous to models which fit quantitative scores to summarise interaction. They differ in that the values of the score variable(s) are to be estimated by the model. This makes them
  - no longer loglinear; and
  - no longer directly estimable in the GLM framework: iteration is necessary.

- Where the score models assume that the effect comes from a dimension that we have measured (e.g., the actual value of the dose) logmultiplicative models assume that the effect can be represented on an unobserved dimension, with the actual locations on this dimension to be estimated. They are thus more general.
The equation of a log-multiplicative model estimating one scale is

$$\log F_{ij} = \lambda_0 + \lambda_i^A + \lambda_j^B + \alpha_i \zeta_j$$

or

$$F_{ij} = \tau_0 \tau_i^A \tau_j^B e^{\alpha_i \zeta_j}$$

Because of the multiplication of $\alpha_i$ and $\zeta$ it is no longer a linear model in the log format. $\alpha$ is a slope variable, with a value for each row (in this case), and $\zeta$ is the scale, a continuous variable with a separate (estimated) value for each column.

This corresponds closely to the model where we apply a known scale:

$$\log F_{ij} = \lambda_0 + \lambda_i + \lambda_j + \beta_{i,j}$$
Where both row and column effects are estimated, the model is Goodman’s RC log-multiplicative model:

\[ \log F_{ij} = \lambda_0 + \lambda_i^A + \lambda_j^B + \phi \mu_i \nu_j \]

- \( \phi \) is a scaling variable, \( \mu_i \) and \( \nu_j \) are the row and column effects.

Estimation of these models is possible in GLIM using macros available in Lindsey’s *Modelling Frequency and Count Data*, or by using the program /EM (available free via [http://cwis.kub.nl/~fsw_1/mto/](http://cwis.kub.nl/~fsw_1/mto/)).
More dimensions: trend models

- The association summarised by a quantitative covariate may be in tables with dimensions greater than two.

- A linear-by-linear association may be fitted on two dimensions and allowed to vary across categories of a third (nominal) variable. Similarly an association with a single quantitative component could vary by a third variable.

\[
\log F_{ijk} = \lambda_0 + \lambda_i + \lambda_j + \lambda_k + \lambda_{ik} + \lambda_{jk} + \lambda_{ijk}\beta \times i \times j
\]

The interaction in \( \lambda_{ijk}\beta \) constitutes a test whether the overall \( IJ \) association described by \( \beta_{ij} \) is the same across all levels of the \( K \) variable. If all the \( \lambda_{ik} \)'s are equal to one, the association pattern is not changing.
• Alternatively we can find models where we allow a full two-way association, but constrain the three-way association to be linear in the $K$ dimension:

$$\log F_{ijk} = \lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ij}^{AB} + \lambda_{ij}^{AB} \times k$$

• The Payne, Payne and Heath paper in Dale and Davies, *Analyzing Social and Political Change* considers this model in detail. Where the $K$ dimension is something like time, these can be considered trend models, because they allow the association to change in a (log-)linear trending fashion over time.

• The interpretation is simple: there is an $IJ$ association, which changes over time. For instance, there may be over-population of the diagonal, but it is less in earlier times than later, and the change can be approximated by a log-linear effect.
In terms of odds-ratios, they are on the whole significantly different from one, but each one changes in a log-linear fashion across time (towards or away from one).

Other models are available which constrain all the odds ratios to move in the same direction (e.g., Xie’s “unidiff” model): these test whether the association is weakening or strengthening, whereas the trend model simply allows the odds ratios to change
Fitting a trend model is straightforward, since it is not different from the 2-D linear models we have considered:

```
compute time = c.
genlog a b c with time
   /print=est/plot=none
   /design= a b c
       a by b   a by c   b by c
       a by b by time
```

The interpretation is that \( a \) varies across time, \( b \) varies across time, there is an association between \( a \) and \( b \), and this association evolves log-linearly over time.

Extensions of this model by imputing values to \( \text{time} \), or by estimating it log-multiplicatively, are straightforward.
The logit/loglinear equivalence

- For most users of statistics, the loglinear model has one big unaccustomed characteristic: it is an association model; it has no dependent variable.

- There are two other differences from most models which are possibly as important but less disturbing:
  - It works on grouped data (tables);
  - It routinely works with very high order interactions.
• Because it’s grouped it is particularly easy to manipulate, but it also becomes clear that including an extra variable changes the data (equally true but less obvious in other models). We could also fit high-order interactions in OLS, for instance, but it’s less convenient than adding more variables.

• But despite the absence of a meaningful dependent variable in the loglinear model, we estimate it with a conventional generalised linear model and we can exploit this to reformulate the loglinear model as a logistic regression model for grouped data.
Loglinear Analysis  

- Logistic regression takes a binary (0/1, yes–no) dependent variable, with a logit link and a binomial error term:

\[
\log \frac{y_i}{1 - y_i} = \alpha + \beta x_i + e_b
\]

or

\[
\log \frac{\pi_i}{1 - \pi_i} = \alpha + \beta x_i
\]

- Consider a \( n \times 2 \) table, where the 2-category variable is to be thought of as dependent.

\[
\log F_{ij} = \lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}
\]
• The logit or log-odds are \( \log F_{i1}/F_{i2} = \log F_{i1} - \log F_{i2} \) for each row:

\[
\begin{align*}
\log F_{i1}/F_{i2} &= \frac{\lambda_0 + \lambda_i^A + \lambda_i^B + \lambda_i^{AB}}{(\lambda_0 + \lambda_i^A + \lambda_i^B + \lambda_i^{AB})} \\
&= \frac{\lambda_i^B - \lambda_i^A}{\alpha} + \beta_i
\end{align*}
\]

That is, there is a direct relationship between the loglinear parameters and logit parameters.
• Thus, where all the independent variables are categorical, a logistic regression can be estimated using a suitably formulated loglinear model.

• The general rule is the model should contain the \( N - 1 \) and lower order interactions between the independent variables, and for every term in the logistic model formula, an interaction between in and the dependent variable. If \( Y \) is the dependent variable, and \( A, B \) and \( C \) are independent, with an interaction \( B \times C \) the loglinear model will have the following design:

\[
Y + A + B + C + A*B + A*C + B*C + A*B*C \quad \text{ <-- Nuisance terms}
\]
\[
+ Y*A + Y*B + Y*C + Y*B*C \quad \text{ <-- Model terms}
\]

All the initial terms collapse into the intercept or constant term, and all the interactions with \( Y \) are interpreted as the effect of that variable on \( Y \).
Logistic regression

- As an example, we take data from the BHPS, wave 5:
  - Dependent: education; A-level-plus versus less than A-level
  - Sex
  - Age (grouped into 10-year bands, highest 70+)
  - Vote (as before).

- We can fit this as a conventional logistic regression:

```
logistic regression hied with esex ageg vote
 /cat=vote ageg
 /con(vote)=ind/con(ageg)=ind.
```

**HIED** is the dependent variable, **VOTE** and **AGEG** are declared categorical, and simple dummy variable (indicator) contrasts are requested.
<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig</th>
<th>R</th>
<th>Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESEX</td>
<td>-.3597</td>
<td>.0520</td>
<td>47.8442</td>
<td>1</td>
<td>.0000</td>
<td>-.0660</td>
<td>.6979</td>
</tr>
<tr>
<td>AGEG</td>
<td>1203.612</td>
<td>6</td>
<td>.0000</td>
<td>.3367</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGEG(1)</td>
<td>2.4024</td>
<td>.1463</td>
<td>269.6315</td>
<td>1</td>
<td>.0000</td>
<td>.1596</td>
<td>11.0492</td>
</tr>
<tr>
<td>AGEG(2)</td>
<td>2.9158</td>
<td>.1037</td>
<td>790.2274</td>
<td>1</td>
<td>.0000</td>
<td>.2739</td>
<td>18.4627</td>
</tr>
<tr>
<td>AGEG(3)</td>
<td>2.2740</td>
<td>.0913</td>
<td>619.7681</td>
<td>1</td>
<td>.0000</td>
<td>.2425</td>
<td>9.7181</td>
</tr>
<tr>
<td>AGEG(4)</td>
<td>1.7560</td>
<td>.0902</td>
<td>378.9330</td>
<td>1</td>
<td>.0000</td>
<td>.1894</td>
<td>5.7890</td>
</tr>
<tr>
<td>AGEG(5)</td>
<td>1.1135</td>
<td>.0908</td>
<td>150.4871</td>
<td>1</td>
<td>.0000</td>
<td>.1189</td>
<td>3.0450</td>
</tr>
<tr>
<td>AGEG(6)</td>
<td>.4487</td>
<td>.0985</td>
<td>20.7411</td>
<td>1</td>
<td>.0000</td>
<td>.0422</td>
<td>1.5663</td>
</tr>
<tr>
<td>VOTE</td>
<td>212.9447</td>
<td>3</td>
<td>.0000</td>
<td>.1403</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VOTE(1)</td>
<td>1.0285</td>
<td>.0889</td>
<td>133.9467</td>
<td>1</td>
<td>.0000</td>
<td>.1120</td>
<td>2.7968</td>
</tr>
<tr>
<td>VOTE(2)</td>
<td>.2291</td>
<td>.0784</td>
<td>8.5261</td>
<td>1</td>
<td>.0035</td>
<td>.0249</td>
<td>1.2574</td>
</tr>
<tr>
<td>VOTE(3)</td>
<td>.8494</td>
<td>.1040</td>
<td>66.6860</td>
<td>1</td>
<td>.0000</td>
<td>.0785</td>
<td>2.3382</td>
</tr>
<tr>
<td>Constant</td>
<td>-.9208</td>
<td>.1289</td>
<td>51.0232</td>
<td>1</td>
<td>.0000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Logistic through loglinear

- Since all independent variables are categorical we can treat this data set as a table, and fit the model using loglinear.

- This is easiest as a ‘Loglinear Logit Model’ to use SPSS’s terms: simply through the Statistics⇒Loglinear⇒Logit menus or with the following syntax:

```plaintext
GENLOG
   hied BY ageg esex vote
   /MODEL=multinomial
   /PRINT est
   /PLOT NONE
   /DESIGN hied hied*ageg hied*esex hied*vote.
```
This has new elements:

- the BY phrase separating the dependent from the independent variables;
- a multinomial /MODEL; and
- a design with the dependent variable in interaction with all the independent variables.

- This fits a very large number of parameters, which have a direct relationship to the individual-data logistic regression parameters.
Loglinear Analysis

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>SE</th>
<th>Z-value</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>1.6406</td>
<td>.1015</td>
<td>16.17</td>
<td>1.44</td>
<td>1.84</td>
</tr>
<tr>
<td>58</td>
<td>.0000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>60</td>
<td>-2.9169</td>
<td>.1037</td>
<td>-28.12</td>
<td>-3.12</td>
<td>-2.71</td>
</tr>
<tr>
<td>61</td>
<td>-2.2742</td>
<td>.0913</td>
<td>-24.90</td>
<td>-2.45</td>
<td>-2.10</td>
</tr>
<tr>
<td>62</td>
<td>-1.7561</td>
<td>.0902</td>
<td>-19.47</td>
<td>-1.93</td>
<td>-1.58</td>
</tr>
<tr>
<td>63</td>
<td>-1.1136</td>
<td>.0908</td>
<td>-12.27</td>
<td>-1.29</td>
<td>-.94</td>
</tr>
<tr>
<td>64</td>
<td>-.4488</td>
<td>.0985</td>
<td>-4.55</td>
<td>-.64</td>
<td>-.26</td>
</tr>
<tr>
<td>65</td>
<td>.0000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>72</td>
<td>.0000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>73</td>
<td>-.3598</td>
<td>.0520</td>
<td>-6.92</td>
<td>-.46</td>
<td>-.26</td>
</tr>
<tr>
<td>74</td>
<td>.0000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>75</td>
<td>.0000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>76</td>
<td>.0000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>77</td>
<td>-1.0288</td>
<td>.0889</td>
<td>-11.58</td>
<td>-1.20</td>
<td>-.85</td>
</tr>
<tr>
<td>78</td>
<td>-.2292</td>
<td>.0785</td>
<td>-2.92</td>
<td>-.38</td>
<td>-.08</td>
</tr>
<tr>
<td>79</td>
<td>-.8497</td>
<td>.1040</td>
<td>-8.17</td>
<td>-1.05</td>
<td>-.65</td>
</tr>
</tbody>
</table>

Brendan Halpin, Sociology Group, Dept of Government & Society, University of Limerick
Swiss Summer School, Lugano August 2002
• This ‘Loglinear Logit Model’ is simply a special parameterisation of a normal loglinear model which we can fit thus:

GENLOG
  hied ageg esex vote
/MODEL=multinomial /PRINT est /PLOT NONE
/DESIGN hied hied*ageg hied*esex hied*vote
  vote*esex vote*ageg esex*ageg
  vote*esex*ageg vote esex ageg.

• This formulation has a more complex model statement and has messier parameter estimates (with this syntax, pairs of estimates have to be subtracted).
For instance, parameter 59 in the previous model (hied=0 by age-group=1) is parameter 11 (hied=1 by age-group=1) below minus parameter 4 (hied = 0 by age-group=1).

<table>
<thead>
<tr>
<th>p.e.</th>
<th>s.e.</th>
<th>12.04</th>
<th>1.70</th>
<th>2.36</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0324</td>
<td>.1688</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.6406</td>
<td>.1015</td>
<td>16.17</td>
<td>1.44</td>
<td>1.84</td>
</tr>
<tr>
<td>.0000</td>
<td>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-.1545</td>
<td>.1953</td>
<td>-.79</td>
<td>-.54</td>
<td>.23</td>
</tr>
<tr>
<td>-.1456</td>
<td>.1800</td>
<td>-.81</td>
<td>-.50</td>
<td>.21</td>
</tr>
<tr>
<td>.2916</td>
<td>.1742</td>
<td>1.67</td>
<td>-.05</td>
<td>.63</td>
</tr>
<tr>
<td>.1595</td>
<td>.1820</td>
<td>.88</td>
<td>-.20</td>
<td>.52</td>
</tr>
<tr>
<td>.0062</td>
<td>.1949</td>
<td>-.03</td>
<td>-.38</td>
<td>.39</td>
</tr>
<tr>
<td>-.5041</td>
<td>.2323</td>
<td>-2.17</td>
<td>-.96</td>
<td>-.05</td>
</tr>
<tr>
<td>.0000</td>
<td>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.2480</td>
<td>.1953</td>
<td>11.51</td>
<td>1.87</td>
<td>2.63</td>
</tr>
<tr>
<td>2.7714</td>
<td>.1838</td>
<td>15.08</td>
<td>2.41</td>
<td>3.13</td>
</tr>
<tr>
<td>2.5658</td>
<td>.1823</td>
<td>14.08</td>
<td>2.21</td>
<td>2.92</td>
</tr>
</tbody>
</table>

---

Brendan Halpin, Sociology Group, Dept of Government & Society, University of Limerick
Swiss Summer School, Lugano August 2002
• The exact pattern of parameters that are estimated (rather than marked 'aliased') in this syntax depends on their entry order, so sometimes we will find the same effect divided between more than one parameter.

• However, the $G^2$ and fitted values are strictly identical.

• An important point to note is that overall fit of the individual-level logistic regression cannot be judged by the $G^2$; because the number of combinations of different values of variables ("settings") is very large and a function of the sample, $G^2$ no longer has a $\chi^2$ distribution. In contrast, $G^2$ for the grouped logistic, or logistic via loglinear, has the usual properties.

• $\Delta G^2$, however, does have a $\chi^2$ distribution, so pairs of nested models can still be compared.
Multinomial logit models through loglinear

- Fitting a multinomial logistic model on grouped data is very easy: simply specify a multicategory dependent variable in the logit-loglinear version of the GENLOG command.

- But how do we interpret it? Basically for each category of the dependent variable, there are a set of parameters that show the effects of the independent variables for that category compared to the reference category of the dependent variable.

That is, it is as if for $n - 1$ categories of the dependent variable, separate simultaneous binary logit models are fitted, pairing each category with the null category.
This is therefore sometimes known as the **baseline-category logit** model:

$$\log \frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + \beta_j x_i$$
Output looks like this, with ‘constants’ for the interaction of all independent variables, followed by terms for interactions between the dependent variable and the terms in the model:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Aliased Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Constant for [FSEX = 1]*[AGEG = 1.00]</td>
</tr>
<tr>
<td>2</td>
<td>Constant for [FSEX = 1]*[AGEG = 2.00]</td>
</tr>
<tr>
<td></td>
<td>. . .</td>
</tr>
<tr>
<td>13</td>
<td>Constant for [FSEX = 2]*[AGEG = 6.00]</td>
</tr>
<tr>
<td>14</td>
<td>Constant for [FSEX = 2]*[AGEG = 7.00]</td>
</tr>
<tr>
<td>15</td>
<td>[VOTE = 1.00]</td>
</tr>
<tr>
<td>16</td>
<td>[VOTE = 2.00]</td>
</tr>
<tr>
<td>17</td>
<td>[VOTE = 3.00]</td>
</tr>
<tr>
<td>18</td>
<td>x [VOTE = 4.00]</td>
</tr>
<tr>
<td>19</td>
<td>[VOTE = 1.00]*[FSEX = 1]</td>
</tr>
<tr>
<td>20</td>
<td>x [VOTE = 1.00]*[FSEX = 2]</td>
</tr>
<tr>
<td>21</td>
<td>[VOTE = 2.00]*[FSEX = 1]</td>
</tr>
<tr>
<td>22</td>
<td>x [VOTE = 2.00]*[FSEX = 2]</td>
</tr>
<tr>
<td>23</td>
<td>[VOTE = 3.00]*[FSEX = 1]</td>
</tr>
</tbody>
</table>
24 \times [\text{VOTE} = 3.00] * [\text{FSEX} = 2] \\
25 \times [\text{VOTE} = 4.00] * [\text{FSEX} = 1] \\
26 \times [\text{VOTE} = 4.00] * [\text{FSEX} = 2] \\
27 \times [\text{VOTE} = 1.00] * [\text{AGEG} = 1.00] \\
28 \times [\text{VOTE} = 1.00] * [\text{AGEG} = 2.00] \\
29 \times [\text{VOTE} = 1.00] * [\text{AGEG} = 3.00] \\
30 \times [\text{VOTE} = 1.00] * [\text{AGEG} = 4.00] \\
31 \times [\text{VOTE} = 1.00] * [\text{AGEG} = 5.00] \\
32 \times [\text{VOTE} = 1.00] * [\text{AGEG} = 6.00] \\
\ldots
From this we can build equations for each dependent category in contrast with the baseline category. Thus for calculating the expected logit for conservative voters who are male, in their thirties we use the terms for $[\text{VOTE} = 1.00]$, $[\text{VOTE} = 1.00]*[\text{FSEX} = 1]$ and $[\text{VOTE} = 1.00]*[\text{AGEG} = 3.00]$

For individual level data, multinomial logistic models exist with the same interpretation, but a different method for fitting.
Models for ordered categories (ii)

- True ordinality means simply a ranking of categories, with no necessity that their ranking should be related to locations on a continuous dimension as with linear-by-linear, log-multiplicative or other models treated above.

- We can fit a number of models which use logit to take account of such ordinality:
  - the adjacent categories model
  - the continuation ratio model
  - the proportional odds model
• All these models can be considered as similar to the **baseline-category logit** model: they fit multiple simultaneous binary logits to restructured versions of the original table.

• One restriction is that the ordinal variable needs to be considered as the dependent variable (this is not the case with the previously discussed models).
Figure 2: Dataset re-arrangements to use logit models to model multicategory data: (a) baseline category (b) adjacent category (c) continuation ratio (d) proportional odds.
• Figure 2 relates these models visually: the top row represents a table containing a variable with five categories (in the table only one row is shown: imagine there are many). Below the table are shown the arrangements of subtables implicit in the various models:
  
  – The baseline-category logit compares each of categories 2 to 5 with category 1, in four simultaneous logit models.
  – The adjacent-category logit fits logit models to 4 adjacent pairs.
  – The continuation-ratio model fits 4 logits on each successive step, with the ‘zero’ category cumulating, and the ‘one’ category being the next category.
  – The proportional odds approach fits 4 models on the entire data, simply changing the cut-point.

• The adjacent-category model and the continuation-ratio model can be implemented simply in SPSS, but the proportional odds model has its
own module. This is because the four sub-models are independent in the former two (each one contains at least some new data) but not in the proportional odds model, where the same observations are simply re-partitioned.

- To fit the adjacent-category and continuation-ratio models in SPSS we need to restructure the table.
For the adjacent category approach we need to create $n - 1$ subtables, each consisting of an adjacent pair. Thus we end up with a table with one binary variable, and an extra dimension with $n - 1$ categories. That is, an $n \times m$ table becomes a $2 \times m \times n - 1$ table.

\begin{verbatim}
c1 c2 c3
r1 11 12 13
r2 21 22 23
r3 31 32 33
r4 41 32 43
\end{verbatim}

becomes

\begin{verbatim}
c1 c2 c3
l1 r1 11 12 13
\end{verbatim}
r1 21 22 23
12 r1 21 22 23
r1 31 32 33
13 r1 31 32 33
r1 41 32 43
For the mouse-foetus data the transformation is like this:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>281</th>
<th>becomes</th>
<th>0</th>
<th>0</th>
<th>281</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>62.5</td>
<td>225</td>
<td>1</td>
<td>0</td>
<td>62.5</td>
<td>225</td>
</tr>
<tr>
<td>1</td>
<td>125</td>
<td>283</td>
<td>1</td>
<td>0</td>
<td>125</td>
<td>283</td>
</tr>
<tr>
<td>1</td>
<td>250</td>
<td>202</td>
<td>1</td>
<td>0</td>
<td>250</td>
<td>202</td>
</tr>
<tr>
<td>1</td>
<td>500</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>500</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>62.5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>62.5</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>125</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>125</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>250</td>
<td>59</td>
<td>1</td>
<td>1</td>
<td>250</td>
<td>59</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>132</td>
<td>1</td>
<td>1</td>
<td>500</td>
<td>132</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>15</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>62.5</td>
<td>17</td>
<td>2</td>
<td>0</td>
<td>62.5</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>125</td>
<td>22</td>
<td>2</td>
<td>0</td>
<td>125</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>250</td>
<td>38</td>
<td>2</td>
<td>0</td>
<td>250</td>
<td>59</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
<td>144</td>
<td>2</td>
<td>0</td>
<td>500</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>62.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>500</td>
</tr>
</tbody>
</table>
treating outcome (normal, abnormal, dead) as ordinal. The first column is the new variable with \( n - 1 \) categories, indexing the new pairs. This can usually be done ‘by hand’ in a text editor or a spreadsheet, since tables are usually small enough to be convenient.

- The model can then be fitted by

```
GENLOG
  outcome  BY  subtab  with  dose2
  /MODEL=MULTINOMIAL
  /PRINT  estim  /PLOT  none
  /DESIGN  outcome  outcome*dose2
            outcome*subtab.
```

where `subtab` is the index of the subtable.
This is analogous to the baseline-category logit, but the ‘real’ variable (e.g., vote) is replaced by the indicator of the subtable. The interpretation of a parameter is the effect of its independent variable on the log-odds of being in a higher versus a lower outcome in any adjacent pair. If we feel that this effect differs across adjacent pairs, a design term like:

/DESIGN = outcome outcome*dose2
          outcome*subtab
          outcome*dose2*subtab.

will allow separate effects.
- The continuation ratio data manipulation is very similar, except we cumulatively collapse categories.

<table>
<thead>
<tr>
<th>DOSE</th>
<th>OUTCOME</th>
<th>COUNT</th>
<th>SUBTAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>281</td>
<td>1</td>
</tr>
<tr>
<td>62.5</td>
<td>1</td>
<td>225</td>
<td>1</td>
</tr>
<tr>
<td>125</td>
<td>1</td>
<td>283</td>
<td>1</td>
</tr>
<tr>
<td>250</td>
<td>1</td>
<td>202</td>
<td>1</td>
</tr>
<tr>
<td>500</td>
<td>1</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>62.5</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>125</td>
<td>2</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>250</td>
<td>2</td>
<td>59</td>
<td>1</td>
</tr>
<tr>
<td>500</td>
<td>2</td>
<td>132</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>282</td>
<td>2</td>
</tr>
<tr>
<td>62.5</td>
<td>1</td>
<td>225</td>
<td>2</td>
</tr>
<tr>
<td>125</td>
<td>1</td>
<td>290</td>
<td>2</td>
</tr>
<tr>
<td>250</td>
<td>1</td>
<td>261</td>
<td>2</td>
</tr>
<tr>
<td>500</td>
<td>1</td>
<td>141</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>62.5</td>
<td>2</td>
<td>17</td>
<td>2</td>
</tr>
<tr>
<td>125</td>
<td>2</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>250</td>
<td>2</td>
<td>38</td>
<td>2</td>
</tr>
<tr>
<td>500</td>
<td>2</td>
<td>144</td>
<td>2</td>
</tr>
</tbody>
</table>
The same model statement as for the adjacent-category logit applies here. The interpretation is a little different: each estimate relates to the effect on the log odds of being in category $i$ versus categories 1 to $i-1$ combined. The file mice-auto.sps gives an example of doing this automatically within SPSS.

The proportional odds model cannot be fitted directly by this means, as the submodels implied are not independent. Various means are available for fitting it iteratively (using GLIM macros, etc.) but it can be fitted as an individual-level ordered logit model in SPSS, Stata and SAS. The “Ordinal Regression” option in SPSS (PLUM), and the ologit command in Stata fit this model, or proc logistic in SAS (see Agresti, p276). The interpretation of the parameters generated is the effect on the log-odds of being in the higher band versus the lower band.
Using $EEM$

- $EEM$ is a program for loglinear and related analysis of categorical data.
- Written by Jeroen Vermunt, Tilburg University.
- Freely available off the WWW, at [http://www.kub.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html](http://www.kub.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html)
- It fits a very wide variety of models, from simple loglinear to log-multiplicative to latent class models. For anyone wishing to push beyond the bounds of SPSS it is worth examining carefully.
- It runs under Windows with a simple interface: a short syntax file is executed and the results are shown in another window.
The format of the syntax file is straightforward, at least for simple models:

```
man 2      * 2 'manifest' variables
dim 3 3    * N cats, fastest changing last
lab S V    * labels, default to A, B, . . .
mod {S,V}  * Fit the main effects
dat [43 6 9 * Data as a table.
     16 11 18
     3 10 16]
```

This fits independence to the hospital visit data used in the practicals.
This is the death penalty example, set up for the model with all the two way interactions:

\[
\text{man 3} \\
\text{dim 2 2 2} \\
\text{lab V, D, P} \\
\text{MOD \{PV, VD, PD\}} \\
\text{dat \[ 53 414} \\
\text{ 11 37} \\
\text{ 0 16} \\
\text{ 4 139]} \\
\]
This is the VOTE*HIED*AGEG table used for the logit models in the previous session. The fastest changing variable is HIED followed by AGEG with VOTE changing last (the command that created these data was crostabs hied by ageg by vote). It fits the model with all two-way interactions.

```
man 3
dim 4 7 2
dum 4 7 2
lab V A E
mod {VA,EA,VE}
dat |
  9  75
 35 322
 64 349
 89 332
149 257
 .  .  .
113 197
 77 108
 58  59
 40  18
 54  12
```
The `dum 4 7 2` statement over-rides `EM`'s usual effect coding of parameters and forces it to use dummy variables, with the last value as reference category (i.e., like SPSS).

The `dim` statement defines the classifying variables according to the order of the counts: given the following order:

<table>
<thead>
<tr>
<th>c1</th>
<th>c2</th>
<th>c3</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>r1</td>
<td>111 112 113</td>
</tr>
<tr>
<td></td>
<td>r2</td>
<td>121 122 123</td>
</tr>
<tr>
<td></td>
<td>r3</td>
<td>131 132 143</td>
</tr>
<tr>
<td></td>
<td>r4</td>
<td>141 142 133</td>
</tr>
<tr>
<td>12</td>
<td>r1</td>
<td>211 212 213</td>
</tr>
<tr>
<td></td>
<td>r2</td>
<td>221 222 223</td>
</tr>
<tr>
<td></td>
<td>r3</td>
<td>231 232 243</td>
</tr>
<tr>
<td></td>
<td>r4</td>
<td>241 242 233</td>
</tr>
</tbody>
</table>
the appropriate definitions would be:

man 3
dim 2 4 3
lab l r c
Linear and log-multiplicative models

- \texttt{LEM} fits many models, described in a 100-page manual available from the website. Among these models are some for linear effects, and for log-multiplicative models. (See sections 2.3 and 11.5.4 of the manual.)

- To fit a model with a single linear effect (as on page 137) we use the \texttt{ass1} subcommand on the \texttt{mod} command:

\begin{verbatim}
man 2
dim 4 5
lab V O
mod \{V,O,ass1(V,O,3a)\}
dum 4 5
dat [123 383 538 703 230
  214 492 732 1172 613
  39 121 237 302 177
  10 28 47 82 60]
\end{verbatim}
Output from this procedure is in `vote-op.lem.out` in the datasets directory.
This is using the \texttt{EVOTE by EOPFAMF} table in unit 5. Fitting this is equivalent to the model on page 143. The fit is reproduced exactly, though the parameterisation is different. In particular the parameters for the \texttt{EVOTE*ELIN} term are set to sum to zero, rather than having a reference category.

The \texttt{ass1(V,O,3a)} phrase means \texttt{V} is the row variable, \texttt{O} the column variable, and the \texttt{3a} means to fit an association with a linear row-effect. \texttt{4a} will give a column effect.
The linear-by-linear model is given by a $2a$ term. Using the \texttt{AOPFAMC*EOPFAMC} table used above, we get the same fit and the same linear-by-linear $\beta$ of 0.6260, though with the other parameters having different values.

\begin{verbatim}
man 2
dim 5 5
lab A E
mod \{A,E,ass1(A,E,2a)\}
dum 5 5
dat [22 63 47 29 4 39 393 549 186 11 40 362 2285 752 46 18 137 801 825 81 3 21 70 127 51]
\end{verbatim}
A log-multiplicative row and column model is like a linear-by-linear model where the row and column indices are replaced by estimated scale values. For this table such a model is fitted with a model statement like this:
\[ \text{mod } A, E, \text{ass2}(A, E, 5a) \]
The results look like this:

```
type 2 association (row=A column=E)
association 5.3907
row  -0.5531 -0.4008 -0.0231  0.3214  0.6555
adj row -1.2841 -0.9305 -0.0535  0.7463  1.5219
column -0.5059 -0.4335 -0.0373  0.2913  0.6855
adj column -1.1747 -1.0065 -0.0866  0.6763  1.5915
```

The $\phi$ is 5.3907, and two sets of the log-multiplicative scales are given: all sum to zero, and the first (row and column, not \textit{adj}) sum to one when squared.

- Log-multiplicative models with only a row scale are fitted by using \textit{e.g.}, \texttt{ass2(A,E,3a)}, or \texttt{ass2(A,E,4a)} for column only.
- \texttt{EM} fits also these models on tables of higher dimension, where for instance the R*C association can differ across a grouping variable.
Some special topics

- Collapsibility: simplifying tables.
- Sparseness and zero cells.
- Dependence and unobserved heterogeneity

Collapsibility

- With tabulated data, it is clear that adding a new variable to the table changes the dataset. The number of cells rises by a factor of $n$, the number
of categories in the new variable, and the counts are spread out over this larger number of cells.

- Because of the risk of sparsity, and loss of parsimony, we may seek ways to exclude variables from the modelling process. But there are two ways of excluding variables: dropping them from the model (e.g., the /DESIGN statement) and not using them in constructing the table.

- The criteria for excluding a variable from a model are clear: its exclusion does not raise $G^2$ significantly. When can we take bigger step of excluding it from the table? That is, when can we collapse the table along one variable?

- In general we can collapse a table where the variable has no significant interaction with any other terms in the model. For instance, in an A*B*C
Loglinear Analysis Unit 9

table, if C has no significant interaction with the A*B association, then the marginal association is not different from the conditional association, and we lose no important information in modelling the marginal A*B table. As demonstrated earlier (page ) if there is association the marginal table may be entirely misleading.

- If one variable is clearly understood as a dependent variable, it may be appropriate to collapse the table along dimensions not related to the dependent variable, even when interactions are present between that dimension and other independent variables.

Brendan Halpin, Sociology Group, Dept of Government & Society, University of Limerick
Swiss Summer School, Lugano August 2002
We use an example from Lindsey (p32ff) which he draws from Fingleton: a survey of shoppers, in particular a 3-way table of shopping trips by size of town visited (small, medium, large), mode of transport (walk, bus, car) and frequency (often, seldom). (See http://teaching.sociology.ul.ie/~brendan/CDA/datasets/travel.sps.)

If we consider frequency as the dependent variable, we can think in terms of a logit model: independence in this framework is given by /DESIGN freq mode*size. The model doesn’t fit, with $G^2$ of 132.9 for 8. Adding a freq*size term drops the $G^2$ by about 50 for 2 df, which is a significant improvement. Adding the freq*mode term instead drops the deviance to 9.02 for 6 df: this is a far more important interaction. The combined model with both interactions has a $G^2$ of 4.14 for 4 df: adding freq*size at this stage is only marginally significant, $p = 0.09$. 
So we ask the question: should we exclude size from both the model statement and the table? First we fit the model without the freq*size interaction on the uncollapsed table:

```
   genlog freq size mode
    /print=est/plot=none
    /design = freq size mode mode*size.
```

and then collapse the table:

```
   genlog freq mode
    /print=est/plot=none
    /design = freq mode.
```

We find the fit changing from 132.9 for 8 df to 123.9 for 2 df, \( p = 0.17 \). Not only does the fit not improve, but we are facing the opposite problem
of sparsity, a table that is too small to fit interesting tables to (only $2 \times 3$). Therefore we retain size as a classifying variable.
Collapsing categories

- Where we can’t exclude an entire variable it may be possible to address sparseness and parsimony by simplifying variables, reducing the number of categories. In this model

```
genlog freq size mode
   /print=est/plot=none
   /design = freq size mode mode*size
       freq*mode freq*size.
```

The two estimated parameters for the `freq*size` interaction are 0.5828 and 0.3895, for the effect of large and medium respectively, versus small, the reference category. Thus the difference between large and medium is approximately half the difference between medium and small. What
happens if we redefine size as large/medium versus small?

recode size (1,2=1) (3=2) into size2.
genlog freq size2 mode
   /print=est/plot=none
   /design = freq size2 mode mode*size2
   freq*mode freq*size2.

Deviance falls from 4.1415 for 4 df to 0.2633 for 2 df. They are both very well fitting models, and the difference between them is slight \( p = 0.14 \), but the simpler model has the virtue of being simpler to interpret.

- These represent two methods for simplifying complex tables: the exclusion of a variable from the table (collapsing the table), or collapsing categories of variables. These strategies can be very useful in larger tables where sparsity or lack of parsimony may be a problem.
• Agresti has an interesting discussion of collapsibility in terms of graph theory (section 7.1).
Sparseness and sampling zeros

- Sparseness occurs when the number of cells in a table is relatively large in comparison with the sample size. It may also occur in parts of a table which exhibits strong association: if many cases are on the diagonal, the bottom left and top right corners may be sparsely populated.

- Sparseness is a general problem in that the properties of $G^2$ and $X^2$ that we depend on for assessing fit are asymptotic large sample properties: we can trust these measures less when there is sparseness. For instance, where there are several cells with fitted values less than about five.

- General sparseness can be dealt with by collapsing the table, or collapsing variable categories (good for localised sparseness), or by getting more data.
Some localised sparseness is not a disaster: models can usually cope with parts of a table which have a low probability.

**Zero cells**

- Zero cells arise for two reasons: either that combination is impossible (structural zero) or it is rare and happens not to occur in the sample.

- The first table shown in this course (page 18) had a structural zero: no men on maternity leave. If you were to tabulate academics’ jobs at both ends of a five-year span, one triangle of the table would be empty: no one gets demoted. Nearly half the table would be occupied by structural zeros.

- Structural zeros are not a problem: weight them out of the analysis (using...
/CSTRUCTURE in SPSS, for instance). We lose a degree of freedom for each cell weighted out, but they do not contribute any information, so this is correct.
Sampling zeros are a different problem: they are an exacerbated case of sparseness. Depending on their prevalence and location in a table they may not be a problem.

But if a model contains a term relating to a marginal that contains a zero total (e.g., if there is an entire row of sampling zeros) the true value of some corresponding parameters will be infinite: the computer will probably estimate a large parameter with a ridiculously large standard error, and the estimate will be very unstable.

Sampling zeros can also cause practical problems for the program, causing it to have difficulty in converging (this is closely related to parameters being infinite).

An easy solution to this problem is to add a small constant to the zero (or
all) cells: Agresti suggests $10^{-8}$; the algorithm has no difficulty with very small numbers but true zeros confuse it.
• Agresti (section 7.4) and Lindsey (Ch. 5) are good on zero cells. Lindsey (in an earlier book) provides a macro which detects which zero cells are causing problems in a model, and refits the model excluding these cells. This can be regarded as a conservative test, as it excludes cells, reducing the degrees of freedom, and is more likely to affect complex models, making it more likely that they will be rejected.

• This automates a proposal Agresti makes, of excluding problem cells and refitting. Lindsey’s macro is written in GLIM, but I have a Stata version which I can make available to anyone interested.
Dependence and unobserved heterogeneity: overdispersion

- The model underlying loglinear models is Poisson regression. A Poisson distribution has
  - a variance equal to its mean, and
  - the expectation that the events being counted are independent.

- If we are missing something important in the table, such as a relevant variable (unobserved heterogeneity) or some form of dependence (e.g., individuals contributing multiple events) the variance of the counts may be greater than their mean. The effect of this is that standard errors will be underestimated by the normal assumptions.
• One strategy is to scale the standard errors by $\sqrt{X^2/df}$ (making the confidence interval wider), or to divide $G^2$ by $X^2/df$.

• Another approach is to use robust standard error estimates: this is a general strategy available in some programs, for instance Stata (lookup robust).

• Examining parameter estimates and their standard errors is an alternative way of deciding whether they contribute to the model's fit: either a conventional z-test ($|\beta/ASE| > 2$), or a Wald test ($((\beta/ASE)^2$ as a $\chi^2$ variable with $df = 1$). ($ASE$ stands for asymptotic standard error.)
Sigmoid curves and binomial distributions

• When modelling a binary dependent variable we want to predict
  – a probability in the 0–1 range, with
  – a binomial distribution

• Under the generalised linear model framework, a number of transformations are possible with a binomially distributed error:
  – Logit: $\log(p/(1 - p))$
  – Probit: inverse of normal CDF
  – Complementary log-log: $\log(-\log(1 - p))$
  – Log-log: $-\log(-\log(p))$
The logit and probit mappings
• The logit and probit mappings are symmetric and give similar results

• The logit transformation is used more often as it is mathematically more tractable

• Under GLMs, the transformation of the dependent variable is a linear function of the independent variables, with a specified error structure (binomial in the case of binary dependent variables):

\[ g(y) = \alpha + \beta X + \varepsilon \]

• In all linear models we interpret the effect of a parameter in the same way: the change in the linear component due to a one-unit change in the independent variable
Under logistic regression, change in the linear component is a change in the log of the odds, and thus the parameter estimates can be interpreted as the log of the odds ratio for cases that differ only by one unit in an explanatory variable:

\[
\log\left(\frac{p_1}{1-p_1}\right) = \alpha + \beta X \\
\log\left(\frac{p_2}{1-p_2}\right) = \alpha + \beta (X + 1) \\
\Rightarrow \log(\Omega_{X,X-1}) = \logit(p_1) - \logit(p_2) \\
= \beta X - \beta (X - 1) = \beta \\
\Omega_{X,X-1} = e^\beta
\]

This is as true for dummy variables as for continuous
But what of the link between parameters and probability?

\[ p = \frac{\omega}{1 + \omega} = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} \]

The effect of a change of \( \beta \) in the linear component depends **strongly** on the starting point: for probabilities in the mid range it is much bigger than for probabilities closer to 0 or 1.

We can present predicted changes in probability for a given starting point – this is sometimes useful in presentation, but generally it is better to get used to (log) odds-ratios.
• The logit and probit transformations are symmetric, so the effect of a one-unit increase of a variable at $p=0.9$ is the same as that of a one-unit decrease at $p=0.1$.

• In most applications this is acceptable, but for situations where it is not, the third and fourth links
  – Complementary log-log: $\log(-\log(1-p))$
  – Log-log: $-\log(-\log(p))$

give mappings that are asymmetric:
Logistic Regression

The logit and two asymmetric mappings
Grouped and individual level logistic

- We have seen that we can fit logistic regressions on individual level or grouped data.

- That with grouped data there is a strong and simple link between logistic and loglinear models.

- And that the parameter estimates are unchanged whether we fit individually or grouped.

- So why fit grouped?
• It can be shown that for poisson, binomial and multinomial models, that as long as the number of “settings” – combinations of values of explanatory variables – is (i) fixed and (ii) small relative to the sample size, $G^2$ has a $\chi^2$ distribution with the number of parameters as degrees of freedom.

• For grouped logistic and loglinear models, these requirements are usually met: we can think of the data set as a table, with a number of cells that is much smaller than the sample size.

• Generally as long as there are very few cells with a fitted value of 5 or less, it is assumed that $G^2$ approximates a $\chi^2$ distribution.

• This allows us to use $G^2$ as a measure of the overall fit: if it is “significant”, then the saturated model is better than this one.
For individual level data, however, there may be as many “settings” as cases, and a new sample will have different “settings”.

Thus, though we can calculate $G^2$, we cannot use it to measure overall fit of individual level models.

However, $\Delta G^2$ derived in comparing pairs of nested models does still have a $\chi^2$ distribution, and should be used to aid in model search.

$\Delta G^2 = \Delta L_m$, since $G^2 = L_m - L_s$

One strategy recommended to assess overall fit is to group all continuous variables and fit a loglinear logistic model.

The Hosmer-Lemeshow test is less drastic but has a similar logic: it groups the data into deciles according to the predicted probabilities, and conducts...
an $X^2$ test on the observed and expected values in the table (decile by dependent variable)

- This test has been shown to be approximately $\chi^2$ distributed, with 8 degrees of freedom

- In SPSS, if the fit is good, the significance p-value should be high (e.g., in excess of 0.2, certainly greater than 0.1)

- Another commonly used test of fit is the classification table: a table of observed by expected values, where a predicted probability $> 0.5$ is considered as 1, otherwise zero. This cut off point is arbitrary, and almost always unsuitable

- The classification plot is more successful, as this shows a histogram of predicted probability values, distinguishing between observed 0s and 1s
• An extension of the classification table approach is the Receiver Operating Characteristic (ROC) curve approach: this draws a plot of the proportion correctly predicted varying the cut-off point from 0 to 100%, against the one minus the proportion wrongly classified as 0s

• The more successful the prediction, the greater the area between the curve and the 45 degree line

• However, it is possible to have the “best” model for a particular data set and yet not have much predictive power

• Model search also involves looking at the contribution of each variable and its parameters. The Wald test uses the fact that standard errors have an asymptotically normal distribution, so for reasonably large samples we can conduct the usual t-test against the null hypothesis that $\beta = 0$

• See Hosmer and Lemeshow, Ch 5, Agresti p112ff.
Multinomial and Ordinal Logistic Regression

- Extension of binary logistic regression to polytomous (multinomial) and ordinal logistic regression is possible (Agresti Ch 8)

- Multinomial logistic proceeds by comparing the effect of covariates on the possibility of being in each of \( n - 1 \) categories compared with a “baseline category”:

\[
\log \left( \frac{\pi_j}{\pi_j} \right) = \alpha_j + \beta_j x
\]

- This is conceptually equivalent to fitting \( n - 1 \) separate binary logistic models, comparing category 1 with category n, category 2 with n, and so on
• In practice the software usually estimates a simultaneous model which is more statistically sophisticated

• For each independent variable, we have $n - 1$ parameter estimates, each estimating the effect of a one-unit change in this variable on the log odds of being in category $j$ rather than $m$
Ordinal Dependent Variables

- When we have an ordinal dependent variable we have a number of options:
  - Treat it as quantitative and linear and use OLS
  - Treat it as nominal and observe the ordinality in the pattern of the parameter estimates
  - Dichotomise it and fit a binary model, or
  - fit an ordinal logistic regression

- All of these may work well but the last is to be preferred
We have already considered a number of formulations of ordinal logistic regression:

- adjacent-category
- continuation-ratio and
- proportional odds

Adjacent-category logistic models effectively fit separate simultaneous logits on subsets of the data, predicting the log odds of being in category $j$ versus $j-1$. The default is to constrain the parameters in these $n-1$ models to be the same: the effect of the independent variable on being in the higher rather than the lower of each adjacent pair.

This assumes a certain (log) linearity in the dependent variable.
However, the possibility of relaxing this constraint means you can check for departures from linearity at the cost of a more complex (imparsimonious) model.

Closely related is the continuation-ratio model:

$$\log\left(\frac{\pi_j}{\sum_{k=1}^{j-1} \pi_k}\right) = \alpha_j + \beta_j$$

If we constrain the parameters to be constant across the comparisons, we have an ordinal model which measures the effect of covariates on the log odds of being "one category higher"
If we reverse the direction of this model:

$$\log\left(\frac{\pi_j}{\sum_{k=j+1}^{J} \pi_k}\right) = \alpha_j + \beta_j$$

we have a model with a nice interpretation for cases of sequential selection such as the educational system: an ordinal variable such as highest qualification is generated by a sequential process of selections

- Staying on after compulsory
- Completing second level
- Entering university
- Completing a degree

At each of these points a different (sequentially smaller) set of individuals is “at risk of” passing the test
• If we were to model the effect of covariates on these progressions it is likely that they would have different effects as each, but it is possible that some effects would be the same

• The (reversed) continuation-ratio model allows us to fit these effects, and test whether they differ by transition

• By fitting the separate logits simultaneously we get a statistically more efficient model, and to the extent that we can constrain certain parameters to be the same across transitions we get a more parsimonous ordinal model
The Proportional Odds Model

- In SPSS and Stata, “Ordinal Logistic Regression” means the proportional odds model:

$$\log\left(\frac{\sum_{j=1}^{j} \pi_k}{1 - \sum_{k=1}^{j} \pi_k}\right) = \alpha + \beta x$$

- This model is attractive for dependent variables whose ordinality can be seen as being based on an underlying continuous dimension, probably unobserved (Agresti, Ch 8.2)

- Thus it would be better suited to a Likert-style dependent variable, than, say, highest educational qualification
The model estimates a set of $n - 1$ “cut-points”, dividing the sample into the $n$ ordinal categories; these can be considered as representing the ordinality of the categories on the underlying linear dimension (effectively in terms of the proportions above and below each binary comparison point).

It then estimates a single $\beta$ for each independent variable, suggesting that the effect of independent variables is the same across the range: