Advances in Visualizing Categorical Data Using the vcd, gnm and vcdExtra Packages in R

Michael Friendly[1] Heather Turner[2] David Firth[2]
Achim Zeileis[3]

[1]Psychology Department
York University

[2]University of Warwick, UK

[3]Department of Statistics
Universität Innsbruck

CARME 2011
Rennes, February 9–11, 2011
Slides: `http://datavis.ca/papers/adv-vcd-4up.pdf`

Co-conspirators

Heather Turner
University of Warwick

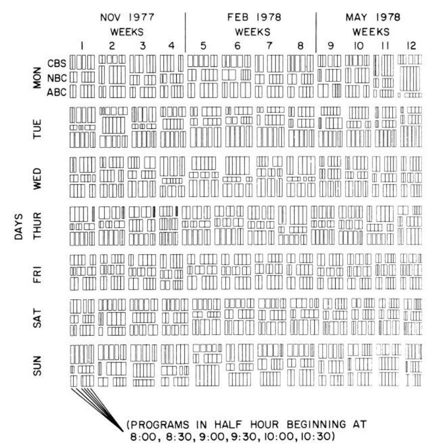
David Firth
University of Warwick

Achim Zeileis
Universität Innsbruck

Outline

Brief History of VCD

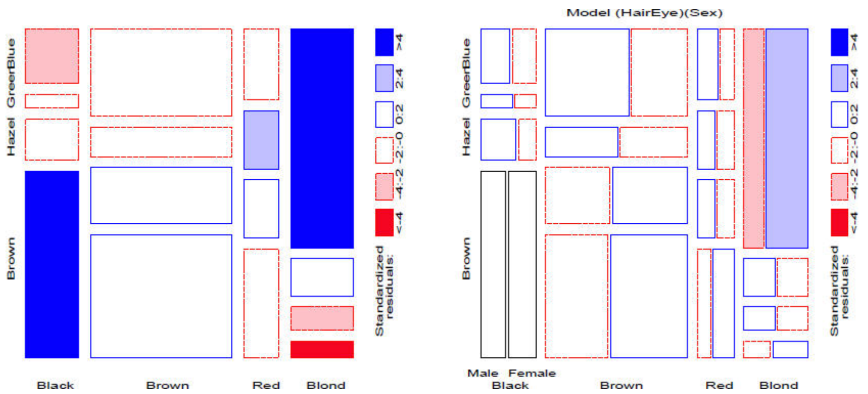
- Hartigan and Kleiner (1981, 1984): representing an n-way contingency table by a "mosaic display," showing a (recursive) decomposition of frequencies by "tiles", area $\sim$ cell frequency.



e.g., a 4-way table of viewing TV programs
`Freq ~Day + Week + Time + Network`

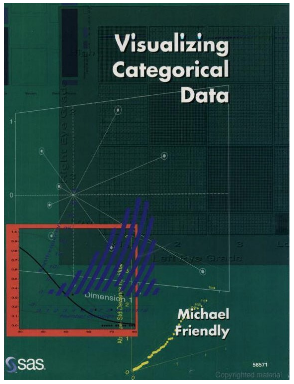
Brief History of VCD

- Friendly (1994): developed the connection between mosaic displays and loglinear models
  - Showed how mosaic displays could be used to visualize both observed frequency (area) and residuals (shading) from some model.
  - 1^{st} presented at CARME 1995 (thx: Michael & Jörg!)

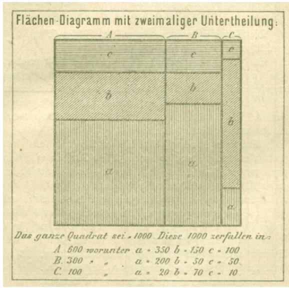


Brief History of VCD

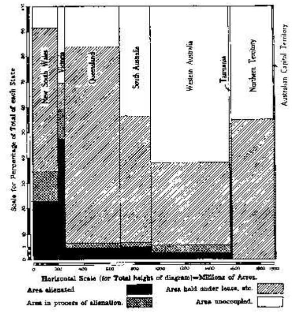
- *Visualizing Categorical Data* (Friendly, 2000)
- But: mosaic-like displays have a long history (Friendly, 2002)!



von Mayr (1877)

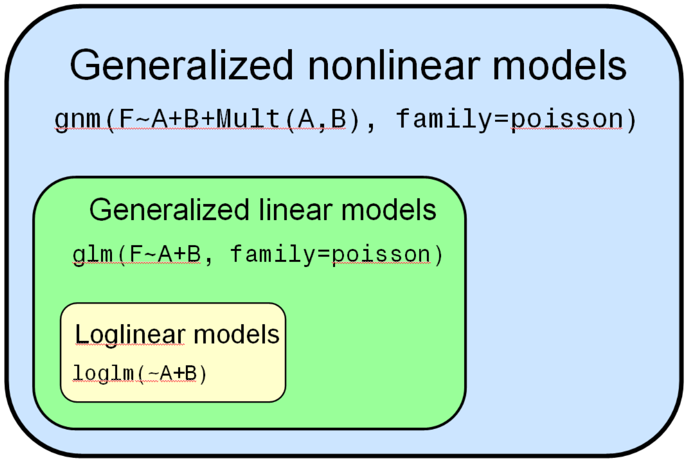


Birch (1964)



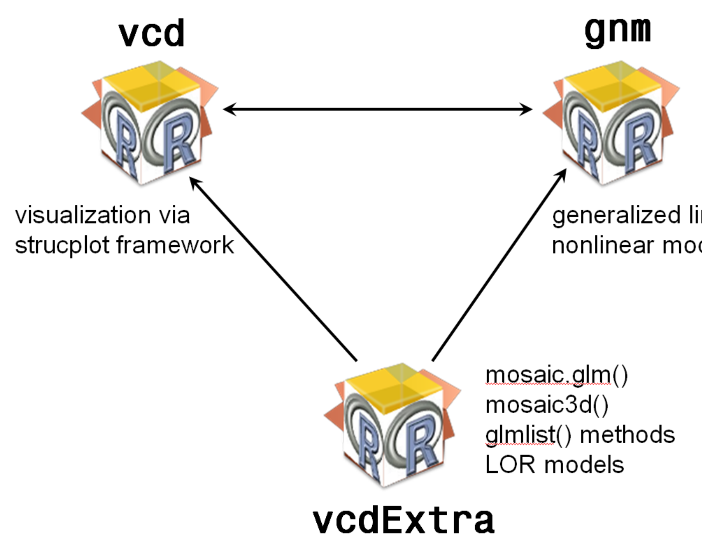
- 2002: vcd project at TU & WU, Vienna (Kurt Hornik, David Meyer, Achim Zeileis) $\mapsto$ **vcd** package

Visual overview: Models for frequency tables



- Related models: logistic regression, polytomous regression, log odds models, ...
- Goals: Connect all with visualization methods

Visual overview: R packages



Extending mosaic-like displays

Initial ideas for mosaic displays were extended in a variety of ways:

- pairs plots and trellis-like layouts for marginal, conditional and partial views (Friendly 1999).
- varying the shape attributes of bar plots and mosaic displays
  - double-decker plots (Hofmann 2001),
  - spine plots and spinograms (Hofmann & Theus 2005)
- residual-based shadings to emphasize pattern of association in log-linear models or to visualize significance (Zeileis et al., 2007).
- dynamic interactive versions (ViSta, MANET, Mondrian):
  - linking of several graphs and models
  - selection and highlighting across graphs and models
  - interactive modification of the visualized models

Generalized mosaic displays

vcd package and the strucplot framework

- Various displays for n-way frequency tables
  - flat (two-way) tables of frequencies
  - fourfold displays
  - mosaic displays
  - sieve diagrams
  - association plots
  - doubledecker plots
  - spine plots and spinograms
- Commonalities
  - All have to deal with representing n-way tables in 2D
  - All graphical methods use area to represent frequency
  - Some are model-based — designed as a visual representation of an underlying statistical model
  - Graphical methods use visual attributes (color, shading, etc.) to highlight relevant statistical aspects

Familiar example: UCB Admissions

Data on admission to graduate programs at UC Berkeley, by Dept, Gender and Admission

```
> structable(Dept ~ Gender + Admit, UCBAdmissions)

                Dept   A   B   C   D   E   F
Gender Admit
Male   Admitted      512 353 120 138  53  22
       Rejected      313 207 205 279 138 351
Female Admitted       89  17 202 131  94  24
       Rejected       19   8 391 244 299 317
```

or, as a two-way table (collapsed over Dept),

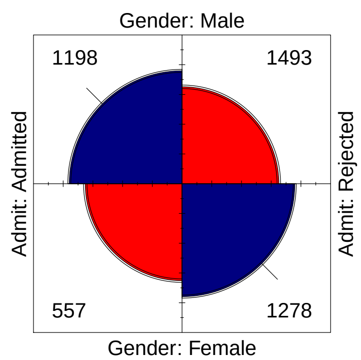
```
> structable(~Gender + Admit, UCBAdmissions)

       Admit Admitted Rejected
Gender
Male             1198     1493
Female            557     1278
```


Fourfold displays for 2 × 2 tables

General ideas:

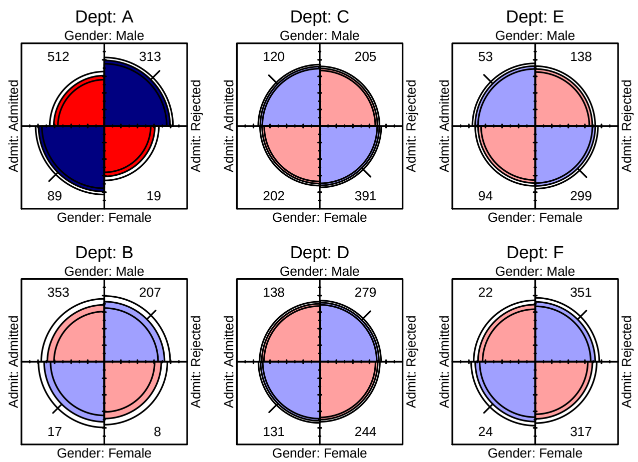
- Model-based graphs can show *both* data and model tests (or other statistical features)
- Visual attributes tuned to support perception of relevant statistical comparisons



- Quarter circles: radius $\sim \sqrt{n_{ij}} \Rightarrow$ **area** $\sim$ **frequency**
- Independence: Adjoining quadrants $\approx$ align
- Odds ratio: ratio of areas of diagonally opposite cells
- Confidence rings: Visual test of $H_0 : \theta = 1 \leftrightarrow$ adjoining rings overlap

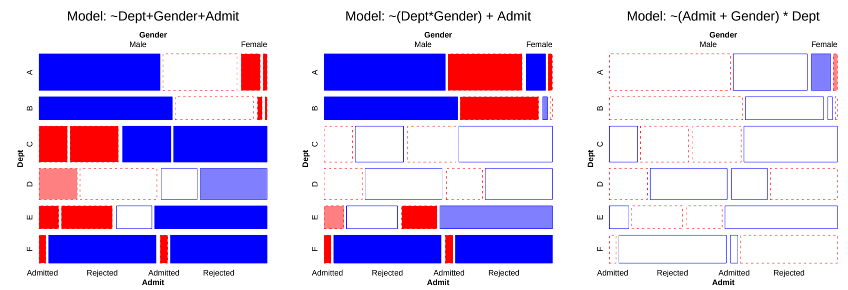
Fourfold displays for 2 $\times$ 2 $\times k$ tables

- Stratified analysis: one fourfold display for each department
- Each 2×2 table standardized to equate marginal frequencies
- Shading: highlight departments for which $H_a : \theta_i \neq 1$



Mosaic displays

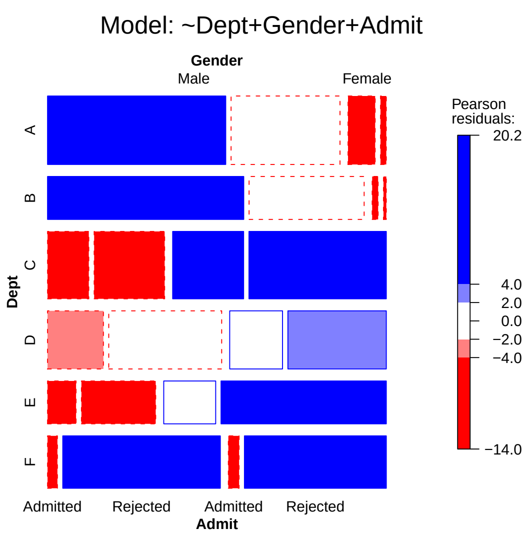
- Tiles: Area $\sim$ observed frequencies, n_{ijk}
- Friendly shading (highlight association *pattern*):
  - Residuals: $r_{ijk} = (n_{ijk} - \hat{m}_{ijk})/\sqrt{(\hat{m}_{ijk})}$
  - Color— blue: $r > 0$, red: $r < 0$
  - Saturation: $|r| < 2$ (none), > 4 (max), else (middle)
- (Other shadings highlight *significance*)
- (Other color schemes: HSV, HCL, ...)



Mosaic displays: Fitting & visualizing models

Mutual independence model: Dept $\perp$ Gender $\perp$ Admit

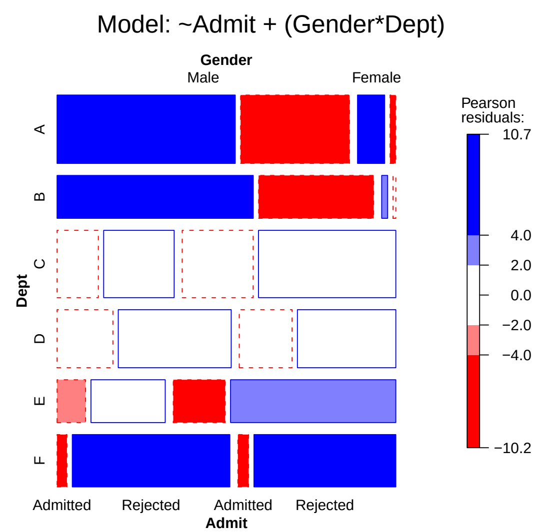
```
> berk.mod0 <- loglm(~Dept + Gender + Admit, data = UCB)
> mosaic(berk.mod0, gp = shading_Friendly, ...)
```



Mosaic displays: Fitting & visualizing models

Joint independence model: Admit $\perp$ (Gender, Dept)

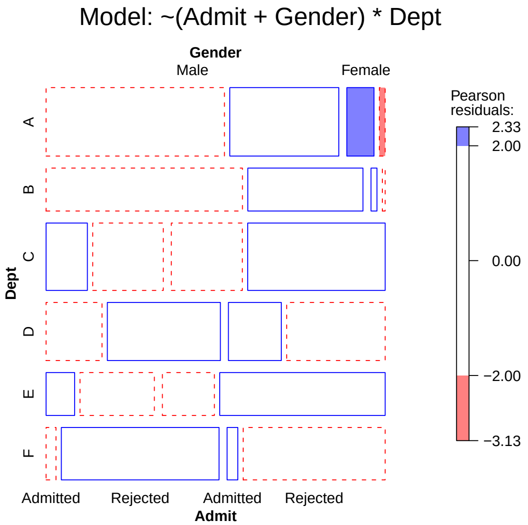
```
> berk.mod1 <- loglm(~Admit + (Gender * Dept), data = UCB)
> mosaic(berk.mod1, gp = shading_Friendly, ...)
```



Mosaic displays: Fitting & visualizing models

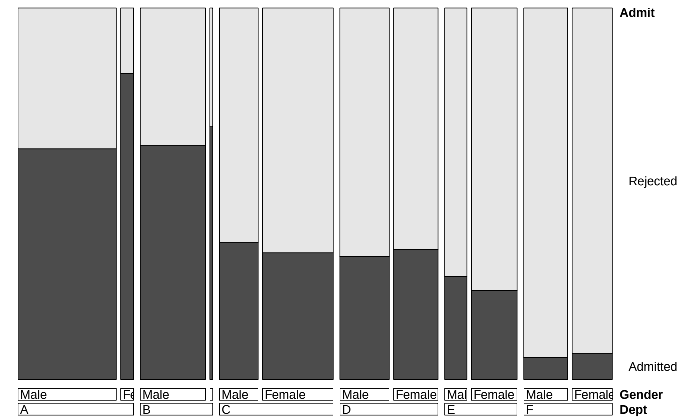
Conditional independence model: Admit $\perp$ Gender | Dept

```
> berk.mod2 <- loglm(~(Admit + Gender) * Dept, data = UCB)
> mosaic(berk.mod2, gp = shading_Friendly, ...)
```



Double decker plots

- Visualize dependence of one categorical (typically binary) variable on predictors
- Formally: mosaic plots with vertical splits for all predictor dimensions, highlighting response



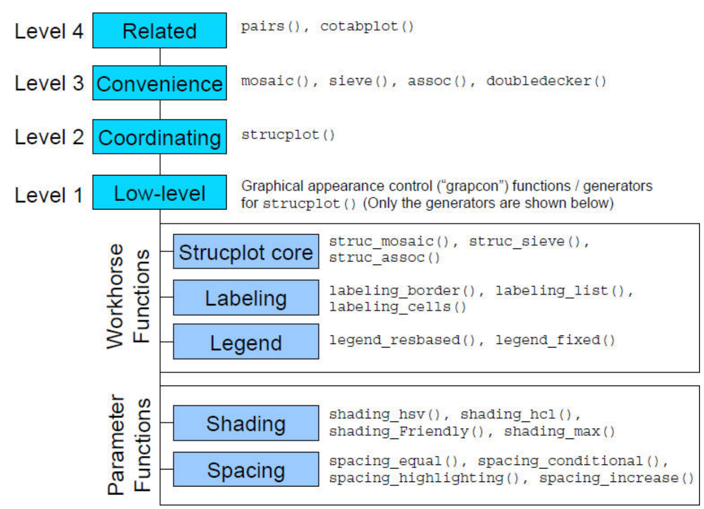
The strucplot framework

A general, flexible system for visualizing n-way frequency tables:

- integrates tabular displays, mosaic displays, association plots, sieve plots, etc. in a common framework.
- n-way tables: variables partitioned into row and column variables in a "flat" 2D display using model formulae
- arguments allow for fitting *any* loglinear model via `loglm()` in the **MASS** package.
- high-level functions for all-pairwise views (`pairs()`), conditional views (`cotabplot()`).
- low-level functions control *all* aspects of labeling, shading, spacing, etc.

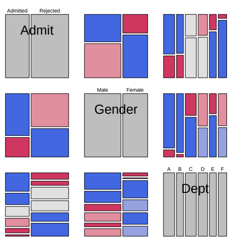
The strucplot framework

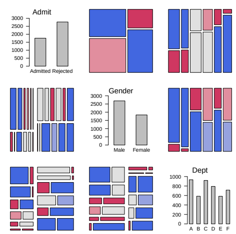
Components of the strucplot framework:

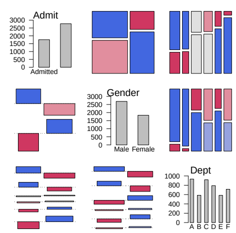


Pairwise bivariate plots

- Visualize all 2-way views of different independence models in n-way tables: `type=`
  - `"pairwise"`: Burt matrix: bivariate, marginal views
  - `"total"`: pairwise plots for mutual independence
  - `"conditional"`: marginal independence, given all others
  - `"joint"`: joint independence of all pairs from other variables
- Panel functions for upper, lower, diagonal panels
  - upper, lower: mosaic, assoc, sieve, ...
  - diagonal: barplot, text, mosaic, ...

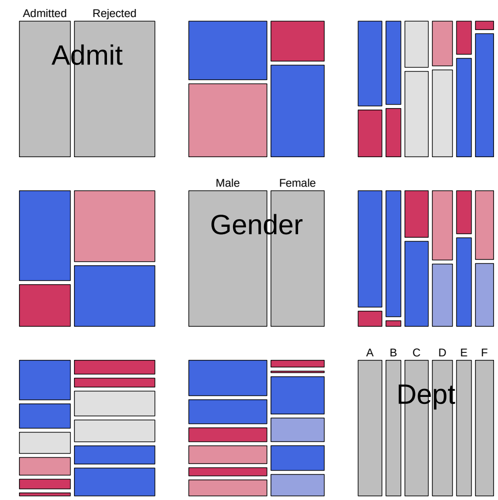






Pairwise bivariate plots

```
> pairs(UCBAdmissions, shade=TRUE, space=0.2,
+   diag_panel = pairs_diagonal_mosaic(offset_varnames=-3, ...))
```



Loglinear models and generalized linear models

- Loglinear models
  - Model fitting in the **vcd** package is based on loglinear models

$$\begin{aligned} \log(m_{ij}) &= \mu + \lambda_i^A + \lambda_j^B \equiv [A][B] \equiv \sim \texttt{A + B} \\ \log(m_{ij}) &= \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \equiv [AB] \equiv \sim \texttt{A * B} \end{aligned}$$

  - Fit using iterative proportional fitting (`loglm()`)
  - $\mapsto$ No standard errors, limited syntax for expressing models
- Generalized linear models
  - Link function:

$$\begin{aligned} E(y \,|\, \boldsymbol{x}) &= g(\mu) = \eta(\boldsymbol{x}) \\ &= \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k \end{aligned}$$

  - Variance function: $\text{Var}(y \,|\, \boldsymbol{x}) = f(\mu)$
  - Loglinear models as special cases with log link, Poisson dist^n $\mapsto$ $\text{Var}(y \,|\, \boldsymbol{x}) = \mu$

Generalized nonlinear models: **gnm** package

- A generalized non-linear model (GNM) is the same as a GLM, except that we allow

$$g(\mu) = \eta(\boldsymbol{x}; \boldsymbol{\beta})$$

  where $\eta(\boldsymbol{x}; \boldsymbol{\beta})$ is nonlinear in the parameters $\boldsymbol{\beta}$.
- GNMs are very general, combining:
  - classical nonlinear models
  - standard link and variance functions for GLM families
- In the context of models for categorical data, GNMs provide:
  - parsimonious models for structured association
  - models for multiplicative association (e.g., Goodman's RC(1) model)
  - multiple instances of multiplicative terms (RC(m) models)
  - user-defined functions for custom models

Generalized nonlinear models: **gnm** package

Some models for structured associations in square tables

- quasi-independence (ignore diagonals)
  ```
  > gnm(Freq ~ row + col + Diag(row, col), family = poisson)
  ```
- symmetry ($\lambda_{ij}^{RC} = \lambda_{ji}^{RC}$)
  ```
  > gnm(Freq ~ Symm(row, col), family = poisson)
  ```
- quasi-symmetry = quasi + symmetry
  ```
  > gnm(Freq ~ row + col + Symm(row, col), family = poisson)
  ```
- fully-specified "topological" association patterns
  ```
  > gnm(Freq ~ row + col + Topo(row, col, spec = RCmatrix), ...)
  ```

All of these are actually GLMs, but the **gnm** package provides convienence functions `Diag`, `Symm`, and `Topo` to facilitate model specification.

Nonlinear models

- Nonlinear terms are specified in model formulae by functions of class `"nonlin"`
- Basic nonlinear functions: `Exp()`, `Inv()`, `Mult()`
- Nonlinear terms can be nested. e.g. for a UNIDIFF model:

$$\log \mu_{ijk} = \alpha_{ik} + \beta_{jk} + \exp(\gamma_k)\delta_{ij}$$

  the exponentiated multiplier is specified as `Mult(Exp(C), A:B)`
- Multiple instances. e.g., Goodman's RC(2) model:

$$\log \mu_{rc} = \alpha_r + \beta_c + \gamma_{r1}\delta_{c1} + \gamma_{r2}\delta_{c2}$$

  specified using: `instances(Mult(A,B), 2)`
- user-defined functions of class `"nonlin"` allow further extensions

All of these are fully general, providing residuals, fitted values, etc.

Generalized nonlinear models: **vcdExtra** package

Provides glue, extending the **vcd** package visualization methods for glm and gnm models

- `mosaic.glm()` $\mapsto$ mosaic methods for class `"glm"` and class `"gnm"` objects
- `sieve.glm()`, `assoc.glm()` $\mapsto$ sieve diagrams and association plots
- Generalized residual types:
  - Pearson
  - deviance
  - standard (adjusted) — unit asymptotic variance
- Model lists:
  - `glmlist()` — methods for collecting, summarizing and visualizing a list of related models
  - `Kway()` — generate & fit models of form `~(A+B+...)`^{k}.

Models for ordered categories

Consider an $R \times C$ table having ordered categories

- In many cases, the RC association may be described more simply by assigning numeric scores to the row & column categories.
- For simplicity, we consider only integer scores, 1, 2, ... here
- These models are easily extended to stratified tables

R:C model	μ_{ij}^{RC}	**df**	**Formula**
Uniform association	$i \times j \times \gamma$	1	`i:j`
Row effects	$\alpha_i \times j$	$(I-1)$	`R:j`
Col effects	$i \times \beta_j$	$(J-1)$	`i:C`
Row+Col eff	$j\alpha_i + i\beta_j$	$I+J-3$	`R:j + i:C`
RC(1)	$\phi_i\psi_j \times \gamma$	$I+J-3$	`Mult(R, C)`
Unstructured (R:C)	μ_{ij}^{RC}	$(I-1)(J-1)$	`R:C`

Example: Social mobility in US, UK & Japan

Data from Yamaguchi (1987): Cross-national comparison of occupational mobility in the U.S., U.K. and Japan. Re-analysis by Xie (1992).

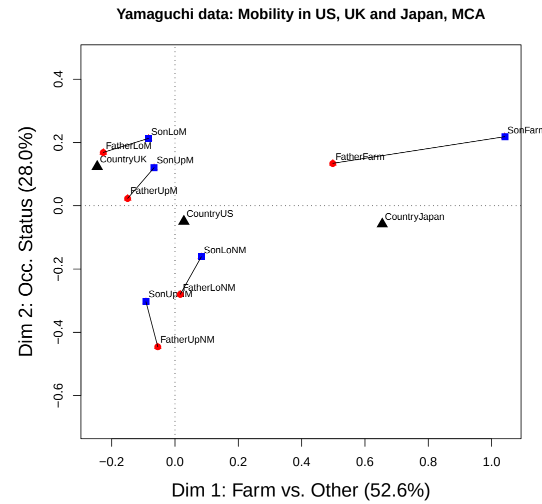
```
> Yama.tab <- xtabs(Freq ~ Father + Son + Country, data = Yamaguchi87)
> structable(Country + Son ~ Father, Yama.tab[, , 1:2])
```

```
       Country   US                             UK
       Son      UpNM LoNM  UpM  LoM Farm UpNM LoNM  UpM  LoM Farm
Father
UpNM            1275  364  274  272   17  474  129   87  124   11
LoNM            1055  597  394  443   31  300  218  171  220    8
UpM             1043  587 1045  951   47  438  254  669  703   16
LoM             1159  791 1323 2046   52  601  388  932 1789   37
Farm             666  496 1031 1632  646   76   56  125  295  191
```

See: `demo("yamaguchi-xie", package="vcdExtra")`

First thought: try MCA

```
> library(ca)
> Yama.dft <- expand.dft(Yamaguchi87)
> yama.mjca <- mjca(Yama.dft)
> plot(yama.mjca, what = c("none", "all"))
```



- Dimensions seem to have reasonable interpretations
- 2^{nd} glance: do they?
- How do they relate to theories of social mobility?
- How to understand Country effects?

Models for stratified mobility tables

Baseline models:

- Perfect mobility: `Freq ~(R+C)*L`
- Quasi-perfect mobility: `Freq ~(R+C)*L + Diag(R, C)`

Layer models:

- Homogeneous: no layer effects
- Heterogeneous: e.g., $\mu_{ijk}^{RCL} = \delta_{ij}^{RC} \exp(\gamma_k^L)$

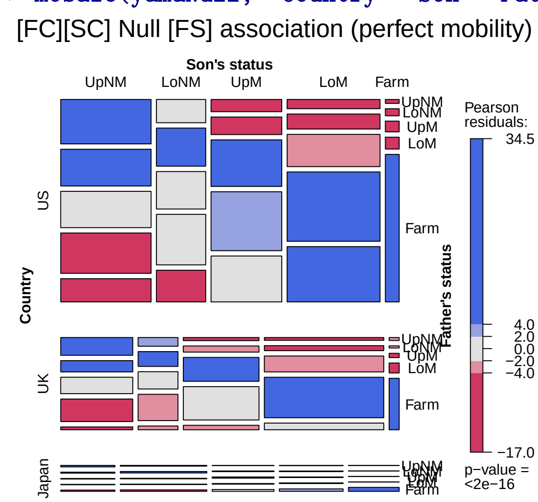
Extended models: Baseline $\oplus$ Layer model(R:C model)

	Layer model	
R:C model	Homogeneous	log multiplicative
Row effects	`~.+ R:j`	`~.+ Mult(R:j, Exp(L))`
Col effects	`~.+ i:C`	`~.+ Mult(i:C, Exp(L))`
Row+Col eff	`~.+ R:j + i:C`	`~.+ Mult(R:j + i:C, Exp(L))`
RC(1)	`~.+ Mult(R, C)`	`~.+ Mult(R, C, Exp(L))`
Full R:C	`~.+ R:C`	`~.+ Mult(R:C, Exp(L)`

Yamaguchi data: Baseline models

Minimal, null model asserts Father $\perp$ Son $|$ Country

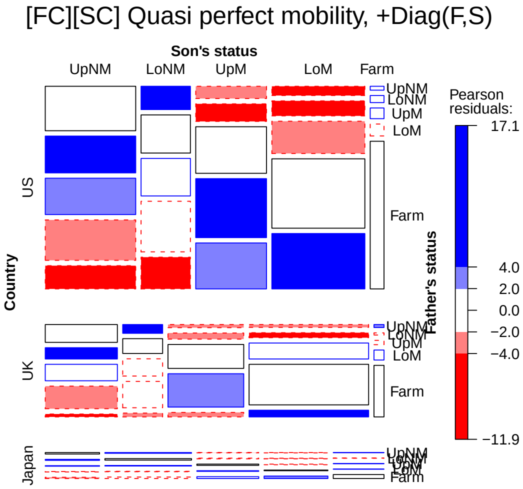
```
> yamaNull <- gnm(Freq ~ (Father + Son) * Country, data = Yamaguchi87,
+     family = poisson)
> mosaic(yamaNull, ~Country + Son + Father, condvars = "Country", ...)
```



Yamaguchi data: Baseline models

But, theory $\mapsto$ ignore diagonal cells

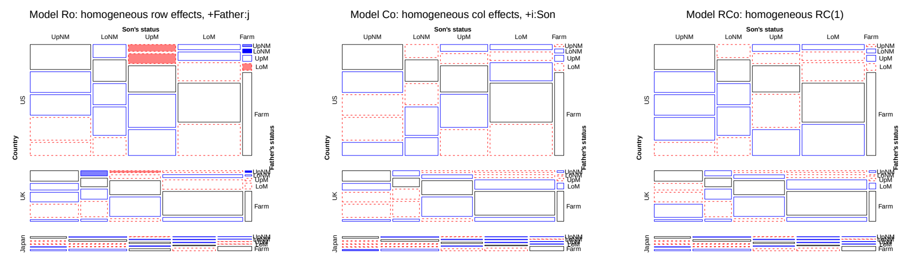
```
> yamaDiag <- update(yamaNull, ~. + Diag(Father, Son):Country)
> mosaic(yamaDiag, ~Country + Son + Father, condvars = "Country", ...)
```



Yamaguchi data: Fit models for homogeneous association

gnm package makes it easy to fit collections of models, with simple `update()` methods

```
> Rscore <- as.numeric(Yamaguchi87$Father)
> Cscore <- as.numeric(Yamaguchi87$Son)
> yamaRo <- update(yamaDiag, ~. + Father:Cscore)
> yamaCo <- update(yamaDiag, ~. + Rscore:Son)
> yamaRpCo <- update(yamaDiag, ~. + Father:Cscore + Rscore:Son)
> yamaRCo <- update(yamaDiag, ~. + Mult(Father, Son))
> yamaFIo <- update(yamaDiag, ~. + Father:Son)
```



Yamaguchi data: Models for heterogeneous association

Log-multiplicative (UNIDIFF) models:

```
> yamaRx <- update(yamaDiag, ~ . + Mult(Father:Cscore, Exp(Country)))
> yamaCx <- update(yamaDiag, ~ . + Mult(Rscore:Son, Exp(Country)))
> yamaRpCx <- update(yamaDiag, ~ . + Mult(Father:Cscore +
+                                      Rscore:Son, Exp(Country)))
> yamaRCx <- update(yamaDiag, ~ . + Mult(Father,Son, Exp(Country)))
> yamaFIx <- update(yamaDiag, ~ . + Mult(Father:Son, Exp(Country)))
```

GNM model methods:

- Summary methods: `print(model)`, `summary(model)`, ...
- Extractor methods: `coef(model)`, `residuals(model)`, ...

Visualization:

- Diagnostics: `plot(model)`
- Mosaics, etc: `mosaic(model)`

Yamaguchi data: Comparing models

`glmlist()` and related methods facilitate model comparison

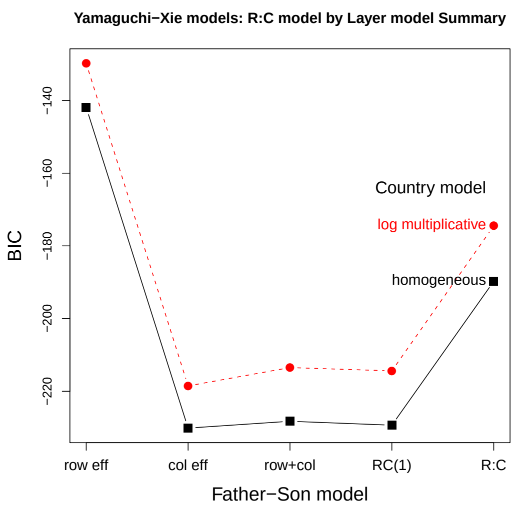
```
> models <- glmlist(yamaNull, yamaDiag,
+                   yamaRo, yamaRx, yamaCo, yamaCx, yamaRpCo,
+                   yamaRpCx, yamaRCo, yamaRCx, yamaFIo, yamaFIx)
> summarise(models)
```

```
Model Summary:
         LR Chisq Df Pr(>Chisq)    AIC    BIC
yamaNull   5591.5 48   0.000000 5495.5 5098.5
yamaDiag   1336.2 33   0.000000 1270.2  997.3
yamaRo      156.0 29   0.000000   98.0 -141.9
yamaRx      147.5 27   0.000000   93.5 -129.8
yamaCo       67.7 29   0.000061    9.7 -230.1
yamaCx       58.8 27   0.000378    4.8 -218.5
yamaRpCo     38.8 26   0.050895  -13.2 -228.2
yamaRpCx     33.0 24   0.103405  -15.0 -213.5
yamaRCo      37.7 26   0.064227  -14.3 -229.3
yamaRCx      32.1 24   0.123995  -15.9 -214.4
yamaFIo      36.2 22   0.028784   -7.8 -189.7
yamaFIx      30.9 20   0.055991   -9.1 -174.5
```

Yamaguchi data: Comparing models

`glmlist()` and related methods facilitate model comparison

```
> BIC <- matrix(summarise(models)$BIC[-(1:2)], 5, 2, byrow = TRUE)
```

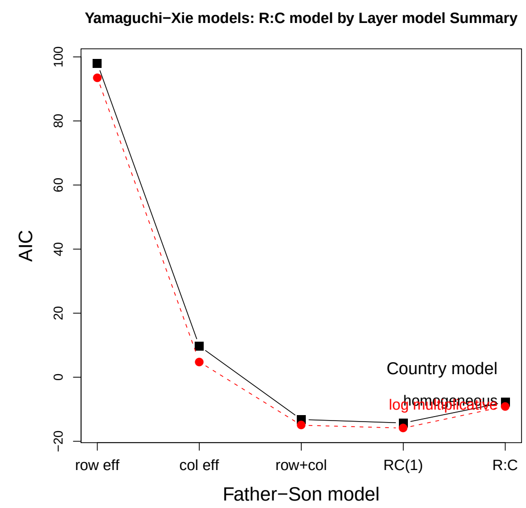


- Homogeneous models all preferred by BIC
- (Xie preferred heterogeneous models)
- Little diff^{ce} among Col, Row+Col and RC(1) models
- $\mapsto$ `R:C` association $\sim$ Row scores (Father's status)

Yamaguchi data: Comparing models

`glmlist()` and related methods facilitate model comparison

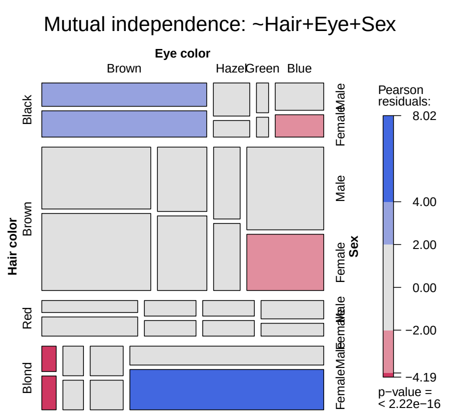
```
> AIC <- matrix(summarise(models)$AIC[-(1:2)], 5, 2, byrow = TRUE)
```

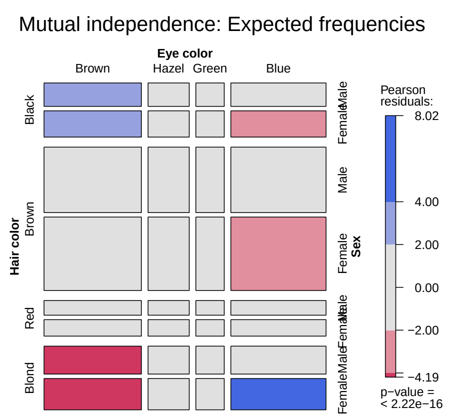


- AIC prefers heterogeneous models
- Row+Col and RC(1) model fit best
- $\mapsto$ `R:C` association $\sim$ Father's status, not just scores
- Model summary plots provide sensitive comparisons!

3D mosaic displays

- Loglinear models rely on $\log(n_{ijk}) \sim$ linear model
  - $\mapsto n_{ijk} \sim$ multiplicative model
- Mosaic displays rely on (nested) use of Area = Height $\times$ Width to represent frequencies in n-way tables
- How to take this to 3D?

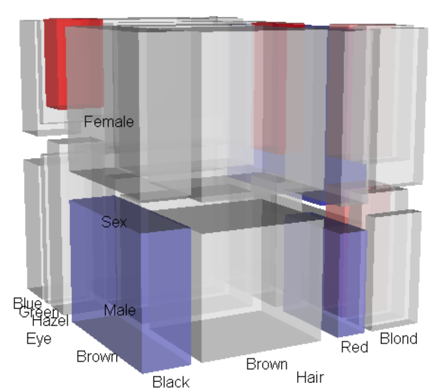




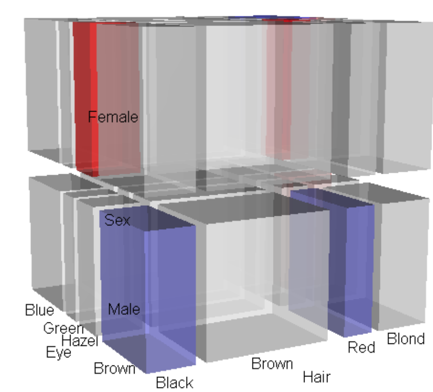
3D mosaic displays

- mosaic3d() in the vcdExtra package
- partitition unit cube $\mapsto$ nested set of 3D tiles, Volume $\sim$ frequency
- uses rgl package: interactive, 3D graphs

```
> mosaic3d(HEC)
```




```
> mosaic3d(HEC, type="expected")
```

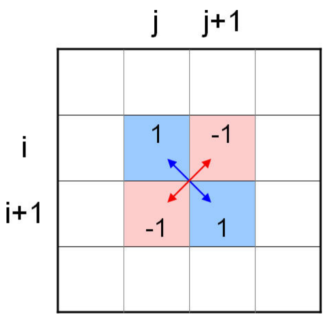


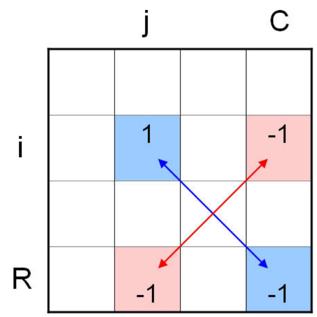
Log odds ratios

- In any two-way, $R \times C$ table, *all* associations can be represented by a set of $(R-1) \times (C-1)$ odds ratios,

$$\theta_{ij} = \frac{n_{ij}/n_{i+1,j}}{n_{i,j+1}/n_{i+1,j+1}} = \frac{n_{ij} \times n_{i+1,j+1}}{n_{i+1,j} \times n_{i,j+1}}$$

$$\ln(\theta_{ij}) = \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \ln \begin{pmatrix} n_{ij} & n_{i+1,j} & n_{i,j+1} & n_{i+1,j+1} \end{pmatrix}^{\mathsf{T}}$$





Log odds ratios

- $\ln \theta_{ij} \sim \mathcal{N}(0, \sigma^2)$, with estimated asymptotic standard error:

$$\widehat{\sigma}(\ln \theta_{ij}) = (n_{ij}^{-1} + n_{i+1,j}^{-1} + n_{i,j+1}^{-1} + n_{i+1,j+1}^{-1})^{1/2}$$

- This extends naturally to $\theta_{ij\,|\,k}$ in higher-way tables, stratified by one or more "control" variables.
- Many models have a simpler form expressed in terms of $\ln(\theta_{ij})$.
  - e.g., Uniform association model

$$\ln(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \gamma \boldsymbol{a}_i \boldsymbol{b}_j \equiv \ln(\theta_{ij}) = \gamma$$

- Direct visualization of log odds ratios permits more sensitive comparisons than area-based displays.

Models for log odds ratios: Computation

- Consider an $R \times C \times K_1 \times K_2 \times \ldots$ frequency table $n_{ij\ldots}$, with factors $K_1, K_2 \ldots$ considered as strata.
- Let $\boldsymbol{n} = \text{vec}(n_{ij\ldots})$ be the $N \times 1$ vectorization of the table.
- Then, all log odds ratios and their asymptotic covariance matrix can be calculated as:
  - $\ln(\widehat{\boldsymbol{\theta}}) = \boldsymbol{C} \ln(\boldsymbol{n})$
  - $\boldsymbol{S} = \text{Var}[\ln(\boldsymbol{\theta})] = \boldsymbol{C}\, \text{diag}(\boldsymbol{n})^{-1}\, \boldsymbol{C}^{\mathsf{T}}$

  where $\boldsymbol{C}$ is an N-column matrix containing all zeros, except for two $+1$ elements and two -1 elements in each row.
- e.g., for a 2×2 table, $\boldsymbol{C} = \begin{bmatrix} 1 & -1 & -1 & 1 \end{bmatrix}$
- With strata, $\boldsymbol{C}$ can be calculated as $\boldsymbol{C} = \boldsymbol{C}_{RC} \otimes \boldsymbol{I}_{K_1} \otimes \boldsymbol{I}_{K_2} \otimes \cdots$
- `loddsratio()` in **vcdExtra** package provides generic methods (`coef()`, `vcov()`, `confint()`, ...)

Models for log odds ratios: Estimation

- A log odds ratio linear model for the $\ln(\boldsymbol{\theta})$ is

$$\ln(\boldsymbol{\theta}) = \boldsymbol{X}\boldsymbol{\beta}$$

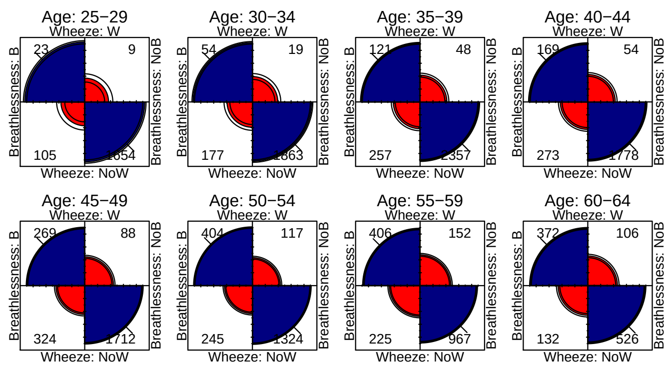
  where $\boldsymbol{X}$ is the design matrix of covariates
- The (asymptotic) ML estimates $\widehat{\boldsymbol{\beta}}$ are obtained by GLS via

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{S}^{-1}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{S}^{-1} \ln \widehat{\boldsymbol{\theta}}$$

  where $\boldsymbol{S} = \text{Var}[\ln(\boldsymbol{\theta})]$ is the estimated covariance matrix
- $\mapsto$ Standard diagnostic and graphical methods can be adapted to this case.
  - diagnostics: influence plots, added-variable plots, ...
  - visualization: effect plots, ...

Example: Breathlessness & Wheeze in Coal Miners

```
> fourfold(CoalMiners, mfcol = c(2, 4), fontsize = 18)
```



- There is a strong + association at all ages
- But can you see the trend?

Example: Breathlessness & Wheeze in Coal Miners

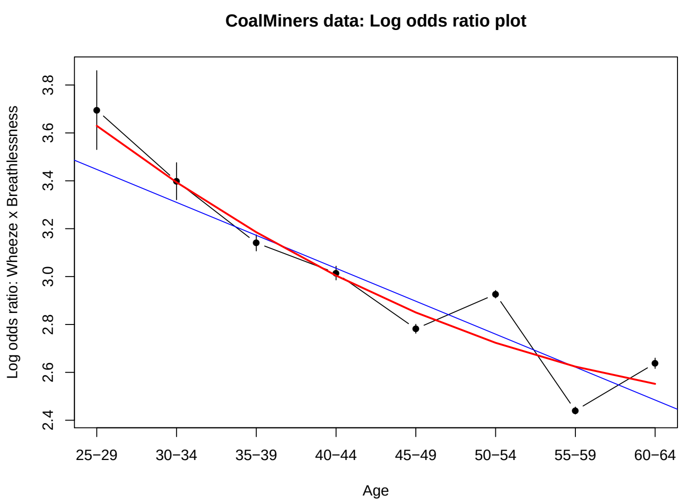
```
> (lor.CM <- loddsratio(CoalMiners))
log odds ratios for Wheeze and Breathlessness by Age
25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64
3.695 3.398 3.141 3.015 2.782 2.926 2.441 2.638
```

Fit linear and quadratic models in Age using WLS:

```
> lor.CM.df <- as.data.frame(lor.CM)
> age <- seq(25, 60, by = 5)
> CM.mod1 <- lm(LOR ~ age, weights=1/ASE^2, data=lor.CM.df)
> CM.mod2 <- lm(LOR ~ poly(age,2), weights=1/ASE^2, data=lor.CM.df)
> anova(CM.mod1, CM.mod2)
Analysis of Variance Table
Model 1: LOR ~ age
Model 2: LOR ~ poly(age, 2)
  Res.Df RSS Df Sum of Sq   F Pr(>F)
1      6 356
2      5 349  1      6.85 0.1   0.77
```


Example: Breathlessness & Wheeze in Coal Miners

Plot log odds ratios and fitted regressions: The trend is now clear!



Attitudes toward corporal punishment

A four-way table, classifying 1,456 persons in Denmark (`Punishment` data in **vcd** package).

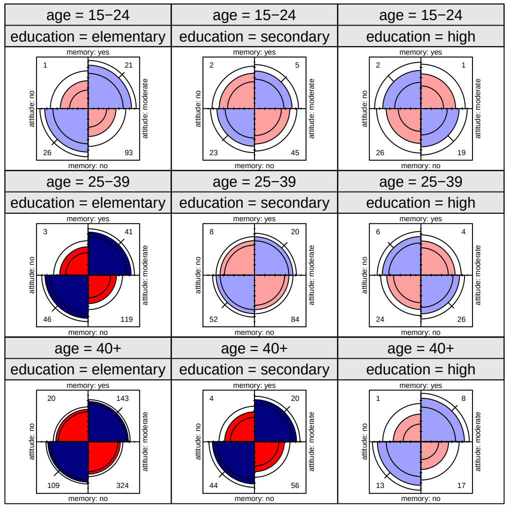
- Attitude: approves moderate punishment of children (moderate), or refuses any punishment (no)
- Memory: Person recalls having been punished as a child?
- Education: highest level (elementary, secondary, high)
- Age group: (15-24, 25-39, 40+)

		Age	15–24		25–39		40+	
Education	**Attitude**	**Memory**	Yes	No	Yes	No	Yes	No
Elementary	No		1	26	3	46	20	109
	Moderate		21	93	41	119	143	324
Secondary	No		2	23	8	52	4	44
	Moderate		5	45	20	84	20	56
High	No		2	26	6	24	1	13
	Moderate		1	19	4	26	8	17

Attitudes toward corporal punishment

Fourfold plots: Association of Attitude with Memory

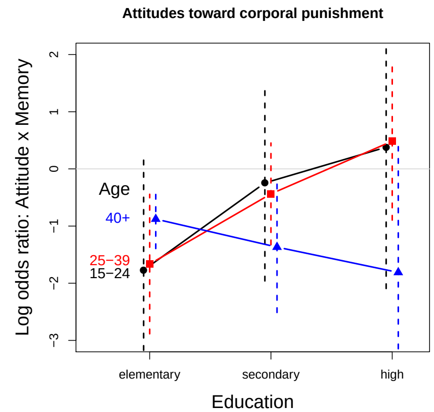
```
> cotabplot(punish, panel = cotab_fourfold)
```



Log odds ratio plot

```
> (lor.pun <- loddsratio(punish))

log odds ratios for memory and attitude by age, education
       education
age     elementary secondary    high
  15-24    -1.7700   -0.2451  0.3795
  25-39    -1.6645   -0.4367  0.4855
  40+      -0.8777   -1.3683 -1.8112
```



- Structure now completely clear
- Little diff^{ce} between younger groups
- Opposite pattern for the $40+$
- Need to fit an LOR model to confirm appearences (SEs large)
- (These methods are under development)

Summary

- Effective data analysis for categorical data depends on:
  - Flexible models, with syntax to specify possibly complex models — *easily*
  - Flexible visualization tools to help understand data, models, lack of fit, etc. — *easily*
- The **vcd** package provides very general visualization methods via the strucplot framework
- The **gnm** package extends the class of applicable models for contingency tables considerably
  - Parsimonious models for structured associations
  - Multiplicative and other nonlinear terms
- The **vcdExtra** package provides glue, and a testbed for new visualization methods

Further information

vcd Zeileis A, Meyer D & Hornik K (2006). The Strucplot Framework: Visualizing Multi-Way Contingency Tables with **vcd**. *Journal of Statistical Software*, **17**(3), 1–48. http://www.jstatsoft.org/v17/i03/ `vignette("strucplot", package="vcd")`.

gnm Turner H & Firth D (2010). Generalized nonlinear models in R: An overview of the **gnm** package. http://CRAN.R-project.org/package=gnm `vignette("gnmOverview", package="gnm")`.

vcdExtra Friendly M & others (2010). **vcdExtra**: vcd additions. http://CRAN.R-project.org/package=vcdExtra. `vignette("vcd-tutorial")`.

References I

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89, 190–200. URL `http://www.jstor.org/stable/2291215`.

Friendly, M. (2000). *Visualizing Categorical Data*. Cary, NC: SAS Institute.

Friendly, M. (2002). A brief history of the mosaic display. *Journal of Computational and Graphical Statistics*, 11(1), 89–107.

Hartigan, J. A. and Kleiner, B. (1981). Mosaics for contingency tables. In W. F. Eddy (Ed.), *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, (pp. 268–273). New York, NY: Springer-Verlag.

Hartigan, J. A. and Kleiner, B. (1984). A mosaic of television ratings. *The American Statistician*, 38, 32–35.

Zeileis, A., Meyer, D., and Hornik, K. (2007). Residual-based shadings for visualizing (conditional) independence. *Journal of Computational and Graphical Statistics*, 16(3), 507–525.