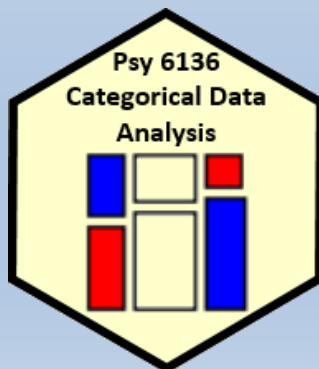
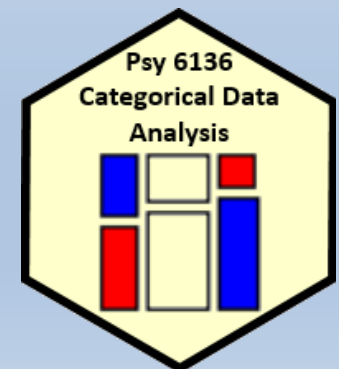


Loglinear models & mosaic displays



Michael Friendly
Psych 6136

<http://friendly.github.io/psy6136>



Today's topics

- Mosaic displays: basic ideas
- Models for count data
 - Fitting loglinear models
- Two-way tables
- Three-way tables: different kinds of independence
- Sequential plots & models
- Marginal & partial displays

Mosaic displays

- Similar to sieve plot, tile plot, using area \sim frequency
- Mosaic plots generalize more readily to n -way tables (subject to resolution of the display)
- Intimately connected to loglinear & generalized linear models
 - Can fit **sequential** models as variables are entered
 - Show the **pattern** of association not accounted for in a given model

Mosaic displays: basic ideas

Mosaic displays theory: Hartigan & Kleiner (1981); Friendly (1994, 1999)

UCB Admissions: Gender frequencies

Area proportional display for an n-way table

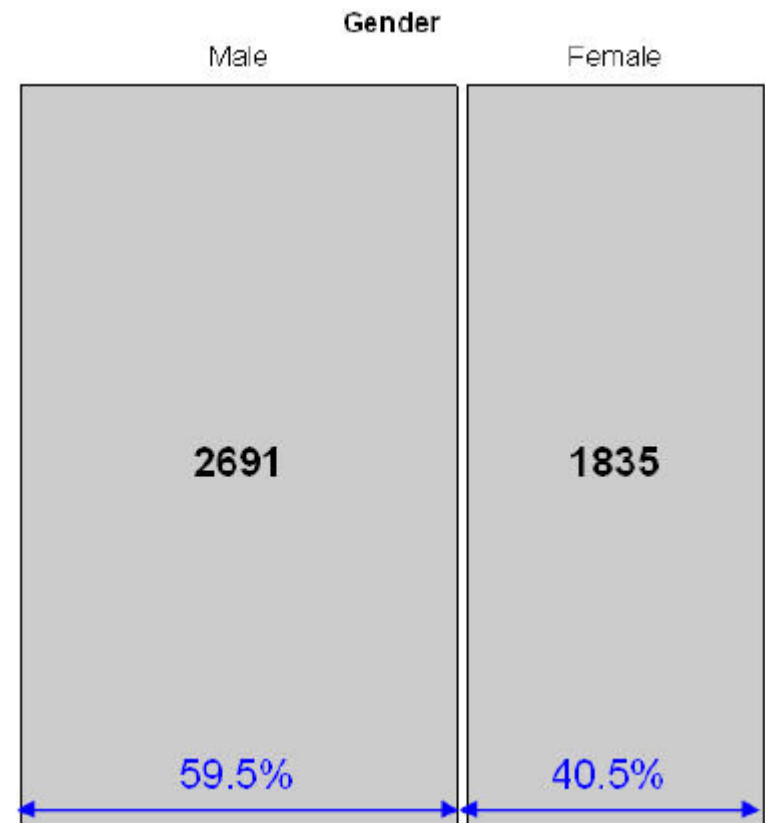
Tiles: **recursive** splits of a unit square, alternating H, V

V_1 : width \sim marginal frequencies, n_{i++}

V_2 : height \sim cond freq: $V_2 \mid V_1 = n_{ij} / n_{i++}$

V_3 : width \sim cond freq: $V_3 \mid V_1, V_2 = n_{ijk} / n_{ij+}$

\rightarrow Area \sim cell frequency, n_{ijk}



Mosaic displays: basic ideas

UCB Admissions: Gender x Admit

Area proportional display for an n-way table

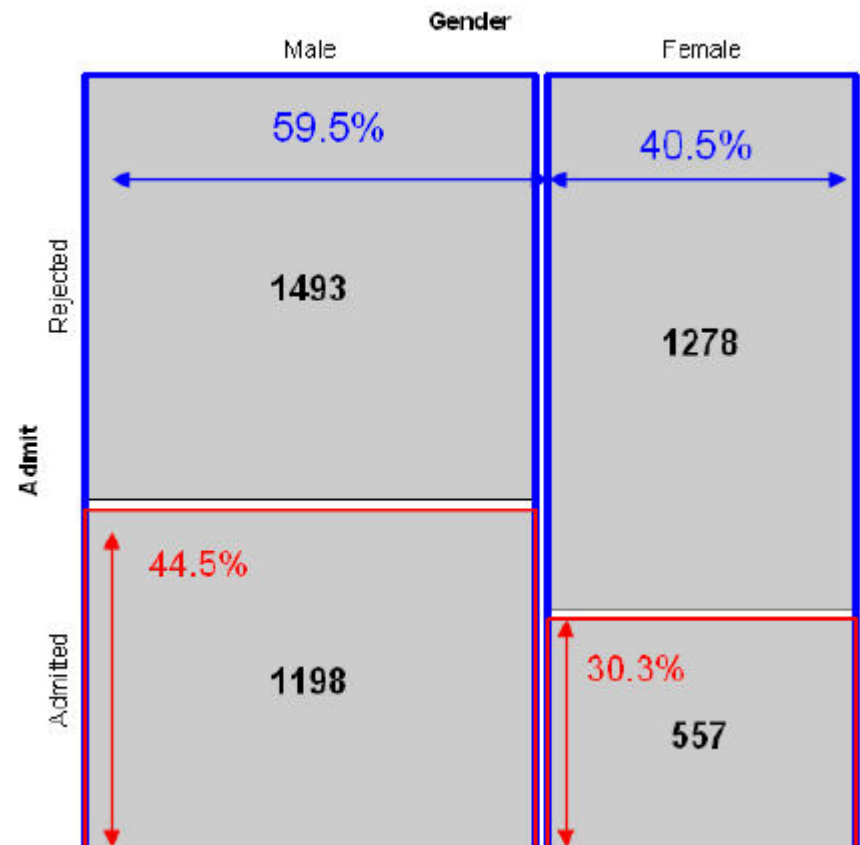
Tiles: recursive splits of a unit square, alternating H, V

V_1 : width \sim marginal frequencies, n_{i++}

V_2 : height \sim cond freq: $V_2 | V_1 = n_{ij} / n_{i++}$

V_3 : width \sim cond freq: $V_3 | V_1, V_2 = n_{ijk} / n_{ij+}$

\rightarrow Area \sim cell frequency, n_{ijk}



Mosaic displays: basic ideas

Gender x Admit x Dept frequencies

Area proportional display for an n-way table

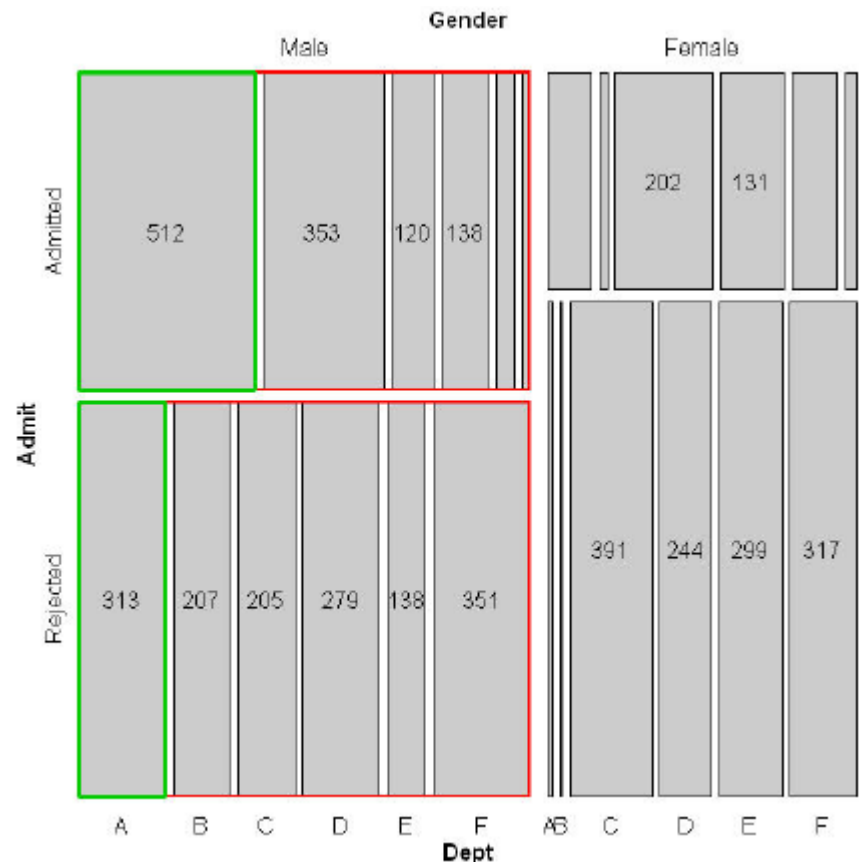
Tiles: recursive splits of a unit square, alternating H, V

V_1 : width \sim marginal frequencies, n_{i++}

V_2 : height \sim cond freq: $V_2 | V_1 = n_{ij} / n_{i++}$

V_3 : width \sim cond freq: $V_3 | V_1, V_2 = n_{ijk} / n_{ij+}$

\rightarrow Area \sim cell frequency, n_{ijk}



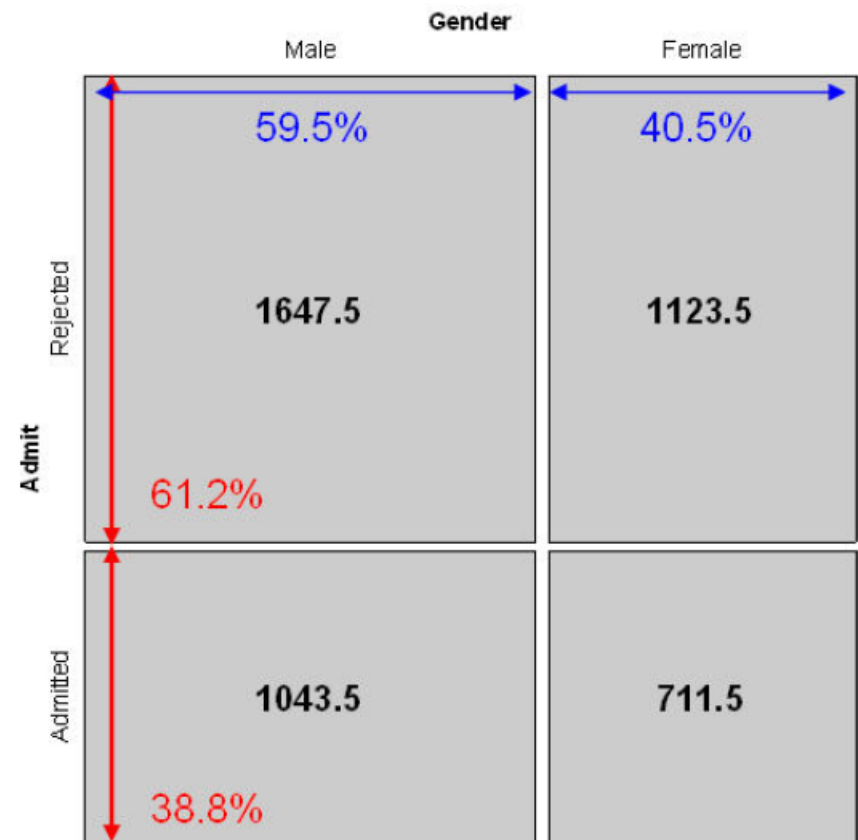
Mosaic displays: Independence

Expected frequencies if $\text{Admit} \perp \text{Gender}$

Expected frequencies under independence are products of the row / col margins

$$\hat{m}_{ij} = \frac{n_{i+}n_{+j}}{n_{++}} = n_{++} \text{row \%col \%}$$

→ Row and col tiles align when variables are independent

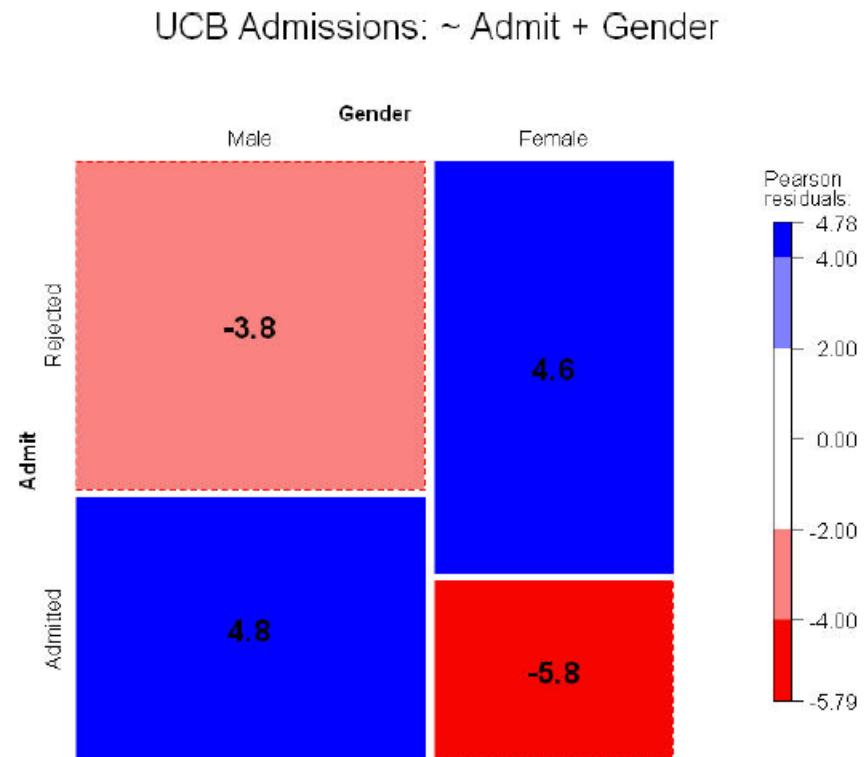


Mosaic displays: Residuals & shading

- Pearson residuals:

$$d_{ij} = \frac{n_{ij} - \hat{m}_{ij}}{\sqrt{\hat{m}_{ij}}}$$

- Pearson $\chi^2 = \sum \sum d_{ij}^2 = \sum \sum \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$
- Other residuals: deviance (LR), Freeman-Tukey (FT), adjusted (ADJ), ...
- Shading:
 - Sign: **— negative in red; + positive in blue**
 - Magnitude: intensity of shading: $|d_{ij}| > 0, 2, 4, \dots$
- \Rightarrow Independence: rows align, or cells are empty!



Loglinear models: Perspectives

Loglinear models grew up and developed from three different ideas and ways of thinking about notions of independence in frequency data

- **Loglinear approach**: analog of ANOVA; associations are interactions
- **glm() approach**: analog of general regression model, for $\log(\text{Freq})$, with Poisson dist^n of errors
- **Logit models**: Loglinear, simplified for a binary response

Loglinear approach

First developed as analog of classical ANOVA models, where multiplicative relations are re-expressed in additive form as models for $\log(\text{Freq})$

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B \equiv [A][B] \equiv \sim A + B$$

- This expresses the independence model for a 2-way table as no A*B association
- Short-hand notations: $[A][B] = A \perp B = \sim A + B$
- Fit by simple **iterative proportional scaling**: **MASS::loglm()**

```
loglm(Freq ~ A + B + C)      # [A] [B] [C]
loglm(Freq ~ A * B + C )     # [A B] [C]
loglm(Freq ~ A * B * C)      # [A B C]
```

glm() approach

Extension of classical linear models recognized loglinear models as a model for $\log(\text{Freq})$, with Poisson dist^n for cell counts

$$\log \mathbf{m} = \mathbf{X} \boldsymbol{\beta}$$

- Looks like std ANOVA/regression model, but for $\log(\text{Freq})$
- This allows **quantitative** predictors and special ways to treat **ordinal** factors
- Fit by maximum likelihood using `glm(..., family=poisson)`
- Standard diagnostic methods available

```
glm( Freq ~ A + B + C, family = poisson)      # [A] [B] [C]
glm( Freq ~ A * B + C, family = poisson)      # [A B] [C]
glm( Freq ~ A * (B+C), family = poisson)      # [A B] [A C]
```

Logit models

When one variable is a **binary** response, a logit model is a simpler way to specify a loglinear model

$$\log(m_{1jk}/m_{2jk}) = \alpha + \beta_j^B + \beta_k^C \equiv [AB][AC][BC]$$

- $\log(m_{1jk}/m_{2jk})$ is the log odds of response 1 vs 2
- The model only includes terms for the effect of A on B & C
- Equivalent loglinear model: $[AB][AC][BC]$
- The logit models **assumes** the $[BC]$ association;

$$[AB] \rightarrow \beta_j^B \quad [AC] \rightarrow \beta_k^C$$

- Fit using **family=binomial**

```
glm(outcome=="survived" ~ B + C, family = binomial)
```

Two-way tables: loglinear approach

For two discrete variables, A and B , suppose a multinomial sample of total size n over the IJ cells of a two-way $I \times J$ contingency table, with cell frequencies n_{ij} , and cell probabilities $\pi_{ij} = n_{ij}/n$.

- The table variables are **statistically independent** when the cell (joint) probability equals the product of the marginal probabilities, $\Pr(A = i \ \& \ B = j) = \Pr(A = i) \times \Pr(B = j)$, or,

$$\pi_{ij} = \pi_{i+} \pi_{+j} \ .$$

- An equivalent model in terms of expected frequencies, $m_{ij} = n\pi_{ij}$ is

$$m_{ij} = (1/n) \ m_{i+} \ m_{+j} \ .$$

- This multiplicative model can be expressed in additive form as a model for $\log m_{ij}$,

$$\log m_{ij} = -\log n + \log m_{i+} + \log m_{+j} \ . \quad (1)$$

Two-way tables: loglinear approach

Independence model

By analogy with ANOVA models, the independence model (1) can be expressed as

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B, \quad (2)$$

- μ is the grand mean of $\log m_{ij}$
- the parameters λ_i^A and λ_j^B express the marginal frequencies of variables A and B — “main effects”
- typically defined so that $\sum_i \lambda_i^A = \sum_j \lambda_j^B = 0$ as in ANOVA

Two-way tables: loglinear approach

Saturated model

Dependence between the table variables is expressed by adding association parameters, λ_{ij}^{AB} , giving the *saturated model*,

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \equiv [AB] \equiv \sim A * B . \quad (3)$$

- The saturated model fits the table perfectly ($\hat{m}_{ij} = n_{ij}$): there are as many parameters as cell frequencies. Residual df = 0.
- A global test for association tests $H_0 : \lambda_{ij}^{AB} = \mathbf{0}$.
- If reject H_0 , which $\lambda_{ij}^{AB} \neq 0$?
- For *ordinal* variables, the λ_{ij}^{AB} may be structured more simply, giving tests for ordinal association.

Example: Independence

Generate a table of Education by Party preference, strictly independent

```
> educ <- c(50, 100, 50) # marginal frequencies
> names(educ) <- c("Low", "Med", "High")
> party <- c(20, 50, 30) # marginal frequencies
> names(party) <- c("NDP", "Liberal", "Cons")
> table <- outer(educ, party) / sum(party) # cell = row * col / n
> names(dimnames(table)) <- c("Education", "Party")
> table
```

	Party		
Education	NDP	Liberal	Cons
Low	10	25	15
Med	20	50	30
High	10	25	15

Perfect fit:

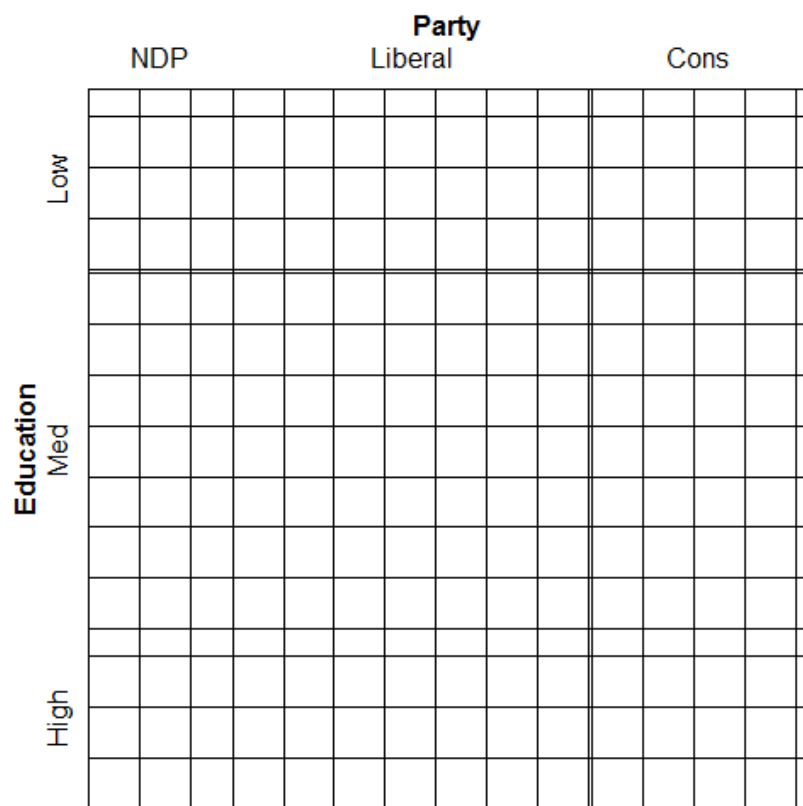
```
> MASS::loglm(~ Education + Party, table)
Call:
MASS::loglm(formula = ~Education + Party, data = table)
```

Statistics:

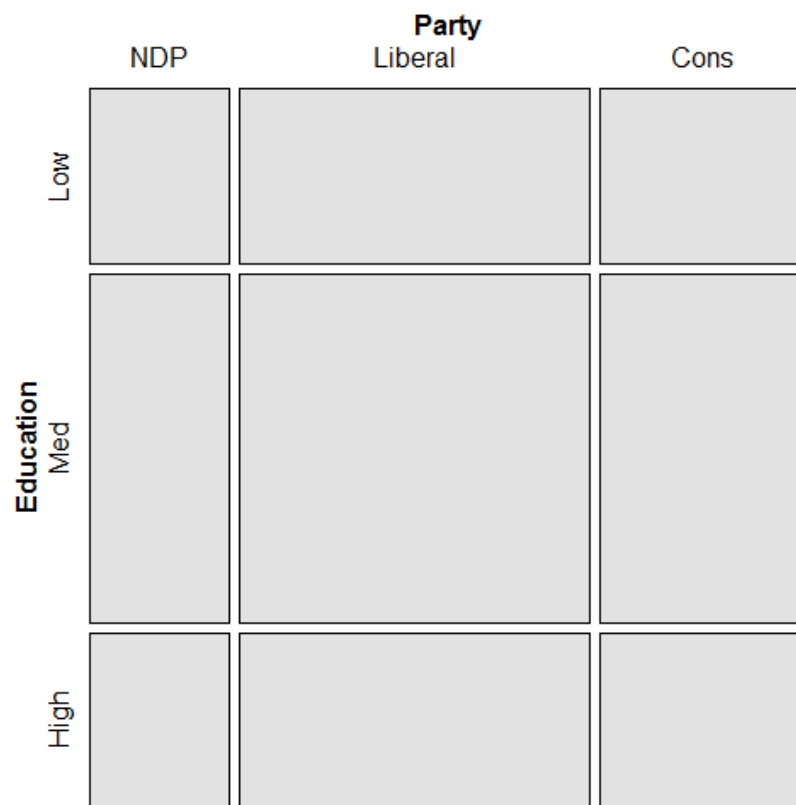
	X^2	df	P(> X^2)
Likelihood Ratio	0	4	1
Pearson	0	4	1

Both sieve diagrams and mosaic plots show what independence “looks like”

```
> sieve(table, shade=TRUE)
```



```
> mosaic(table, shade=TRUE)
```



Two-way tables: glm approach

In the GLM approach, the vector of cell frequencies, $\mathbf{n} = \{n_{ij}\}$ is specified to have a **Poisson** distribution with means $\mathbf{m} = \{m_{ij}\}$ given by

$$\log \mathbf{m} = \mathbf{X}\beta$$

- \mathbf{X} is a known design (model) matrix, expressing the table factors
 - β is a column vector containing the unknown λ parameters.
-
- This is the same as the familiar matrix formulation of ANOVA/regression, except that
 - The response, $\log \mathbf{m}$ makes multiplicative relations additive
 - The distribution is taken as Poisson rather than Gaussian (normal)

Example: 2 x 2 table

For a 2×2 table, the saturated model (3) with the usual zero-sum constraints can be represented as

$$\log \begin{pmatrix} m_{11} \\ m_{12} \\ m_{21} \\ m_{22} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{pmatrix} \mu \\ \lambda_1^A \\ \lambda_1^B \\ \lambda_{11}^{AB} \end{pmatrix} \begin{array}{l} \text{total } n \\ \text{margin A} \\ \text{margin B} \\ \text{association} \end{array}$$

- only the linearly independent parameters are represented. $\lambda_2^A = -\lambda_1^A$, because $\lambda_1^A + \lambda_2^A = 0$, and so forth.
- **association** is represented by the parameter λ_{11}^{AB}
- can show that $\lambda_{11}^{AB} = \frac{1}{4} \log(\theta)$ (log odds ratio)

- Advantages of the GLM formulation: easier to express models with ordinal or quantitative variables, special terms, etc. Can also allow for *over-dispersion*.

Assessing goodness of fit

Goodness of fit of a specified model may be tested by the likelihood ratio G^2 ,

$$G^2 = 2 \sum_i n_i \log \left(\frac{n_i}{\hat{m}_i} \right) , \quad (4)$$

or the Pearson X^2 ,

$$X^2 = \sum_i \frac{(n_i - \hat{m}_i)^2}{\hat{m}_i} , \quad (5)$$

with degrees of freedom $df = \# \text{ cells} - \# \text{ estimated parameters}$.

- E.g., for the model of independence, $[A][B]$, $df = IJ - [(I - 1) + (J - 1)] = (I - 1)(J - 1)$
- The terms summed in (4) and (5) are the squared *cell residuals*
- Other measures of balance goodness of fit against parsimony, e.g., *Akaike's Information Criterion* (smaller is better)

$$AIC = G^2 - 2df \text{ or } AIC = G^2 + 2 \# \text{ parameters}$$

R functions for loglinear models

- **vcd::assocstats()** – only χ^2 tests for two-way tables; not a model (no parameters; no residuals)

- **MASS::loglm()** – general loglinear models for n -way tables

`loglm(formula, data, subset, na.action, ...)`

- **glm()** – all generalized linear models; loglinear with **family = poisson**

`glm(formula, data, weights, subset, ...)`

- Formulas have the form:

- **table** form: $\sim A + B + \dots$ (independence);
- $\sim A * B + C$ (allow $A*B$ association)
- **frequency** data frame: $\text{Freq} \sim A * B + C$

R functions

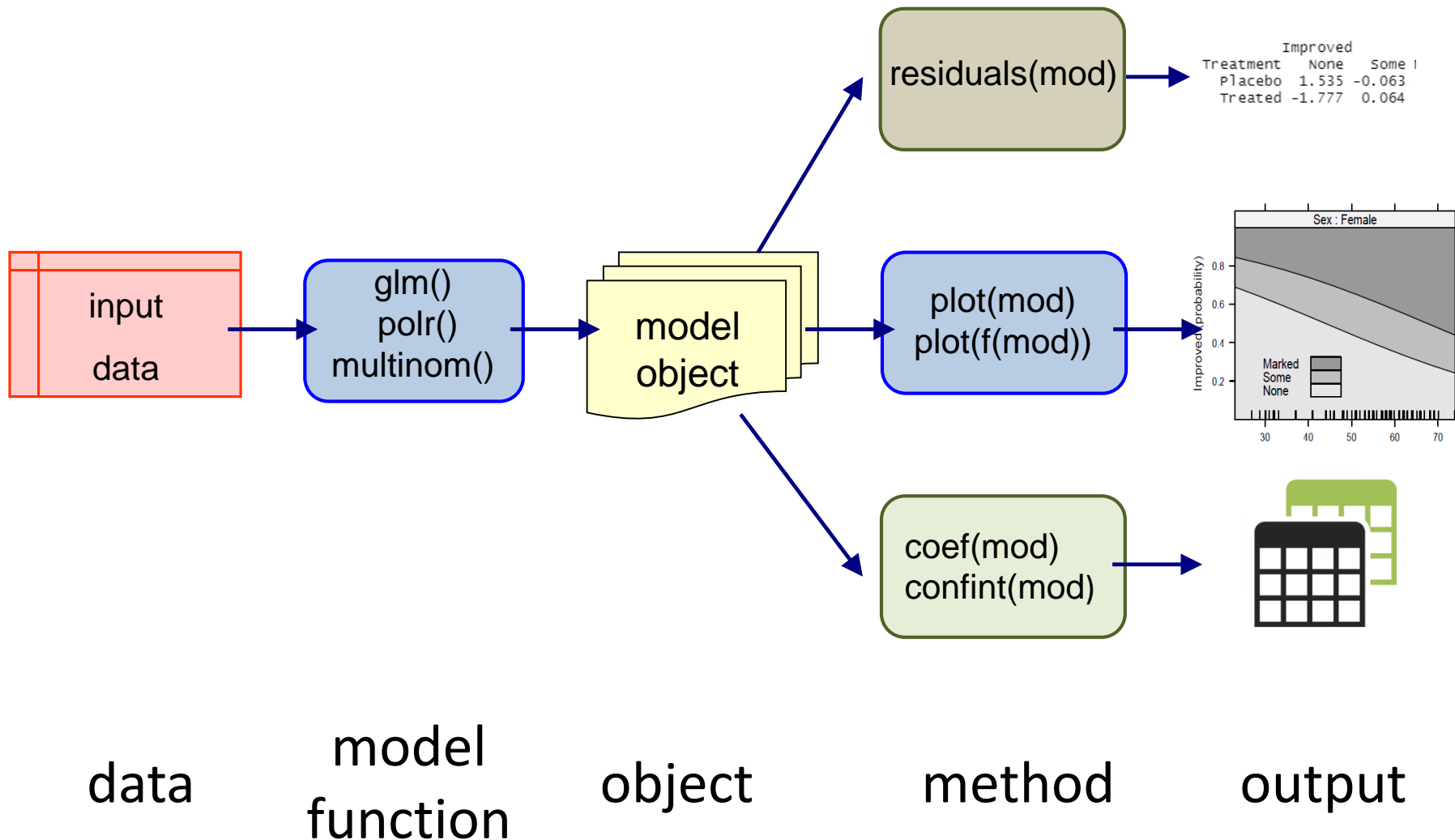
- `loglm()` and `glm()` return an R object with named components and with a `class()`

```
> arth.mod <- loglm(~Treatment+Improved, data=arth.tab, fitted=TRUE)
> names(arth.mod)
> names(arth.mod)
[1] "lrt"          "pearson"      "df"           "margin"       "fitted"       "param"
[7] "call"         "formula"      "frequencies"  "deviance"     "nobs"         "terms"

class(arth.mod)
[1] "loglm"
```

- They have methods: `print()`, `summary`, `coef()`, `residuals()`, `plot()` and other methods
 - Methods are specific to the class of the object
 - E.g., `residuals(arth.mod) → residuals.loglm(arth.mod)`

Model-based methods: Fitting & graphing



Example: Arthritis treatment

Data on effects of treatment for rheumatoid arthritis (in [case form](#))

```
> data(Arthritis, package="vcd")
> str(Arthritis)
'data.frame':      84 obs. of  5 variables:
 $ ID      : int  57 46 77 17 36 23 75 39 33 55 ...
 $ Treatment: Factor w/ 2 levels "Placebo","Treated": 2 2 2 2 2 2 2 2 2 2 ...
 $ Sex      : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
 $ Age      : int  27 29 30 32 46 58 59 59 63 63 ...
 $ Improved : Ord.factor w/ 3 levels "None"<"Some"<...: 2 1 1 3 3 3 1 3 1 1 ...
```

For now, ignore Age; consider the 2 x 3 table of Treatment x Improved

```
> arth.tab <- with(Arthritis, table(Treatment, Improved))
> arth.tab
```

	Improved		
Treatment	None	Some	Marked
Placebo	29	7	7
Treated	13	7	21

Arthritis treatment

Fit the independence model, \sim Treatment + Improved

```
> (arth.mod <- loglm(~Treatment + Improved, data = arth.tab, fitted=TRUE))
Call:
loglm(formula = ~Treatment + Improved, data = arth.tab, fitted = TRUE)
```

Statistics:

	X^2	df	P(> X^2)
Likelihood Ratio	13.53	2	0.001154
Pearson	13.06	2	0.001463

Some methods:

```
> round(residuals(arth.mod), 3)
      Improved
Treatment  None  Some Marked
  Placebo  1.535 -0.063 -2.152
   Treated -1.777  0.064  1.837

# Likelihood ratio chisquare
> deviance(arth.mod)
[1] 13.53
```

```
> coef(arth.mod)
$`(Intercept)`
[1] 2.543

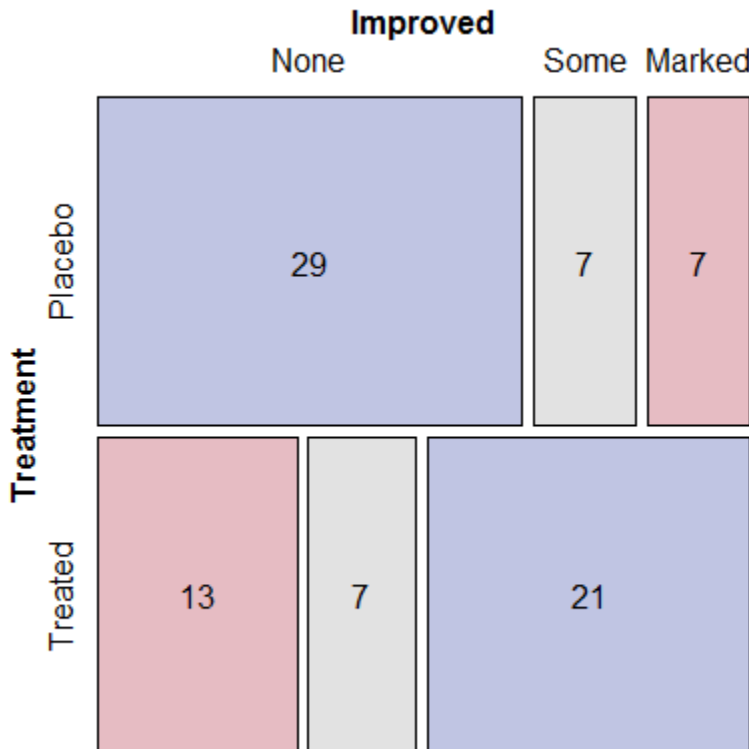
$Treatment
  Placebo  Treated
0.02381 -0.02381

$Improved
      None      Some      Marked
0.50136 -0.59725  0.09589
```

Arthritis treatment: Plots

Visualization: `mosaic()` or `plot()` the model or table

```
> mosaic(arth.mod, shade=TRUE, gp_args=list(interpolate=1:4),  
         labeling = labeling_values)
```



Pearson
residuals:
2.0
1.0
0.0
-1.0
-1.9
p-value =
0.0015

Splits by the response,
Treatment first

Custom scheme for
shading levels; normally
`c(2, 4)` for `|residual|`

Cells can be labeled by
freq, residual, ...

Arthritis treatment: glm()

glm() for loglinear models easiest with the data as a data.frame in **frequency** form

```
> arth.df <- as.data.frame(xtabs(~ Treatment + Improved,  
                               data=Arthritis))
```

```
> arth.df  
  Treatment Improved Freq  
1  Placebo      None   29  
2  Treated      None   13  
3  Placebo      Some    7  
4  Treated      Some    7  
5  Placebo    Marked    7  
6  Treated    Marked   21
```

```
> arth.glm <- glm(Freq ~ Treatment + Improved, data = arth.df,  
                  family = poisson)
```

More on **glm()** models later

Example: Hair color & Eye color

```
> haireye <- margin.table(HairEyeColor, 1:2)
> (HE.mod <- loglm(~ Hair + Eye, data=haireye))
Call:
loglm(formula = ~Hair + Eye, data = haireye)
```

Statistics:

	X^2	df	P(> X^2)
Likelihood Ratio	146.4	9	0
Pearson	138.3	9	0

```
> round(residuals(HE.mod), 2)
```

Re-fitting to get frequencies and fitted values

	Eye			
Hair	Brown	Blue	Hazel	Green
Black	4.00	-3.39	-0.49	-2.21
Brown	1.21	-2.02	1.31	-0.35
Red	-0.08	-1.85	0.82	2.04
Blond	-7.33	6.17	-2.47	0.60

Mosaic displays: Seeing patterns

- In two-way models, residuals contain the info on lack of independence
 - Equivalently: help to understand the **pattern** of association
 - Effect ordering: permuting the rows / cols often makes the pattern more apparent
- Correspondence analysis: → reorder by scores on Dim 1
 - `seriation::permute(order="CA")` does this for two-way tables

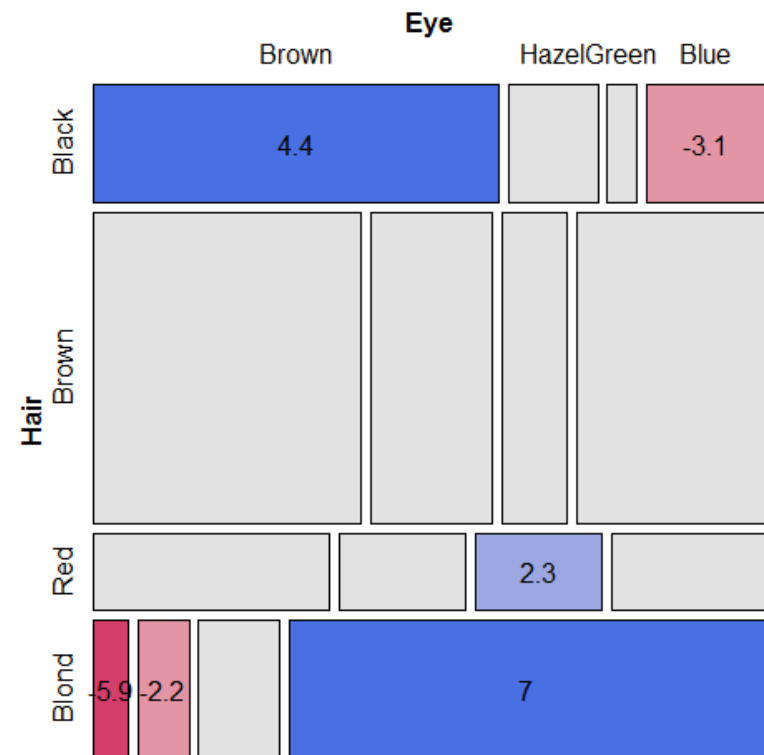
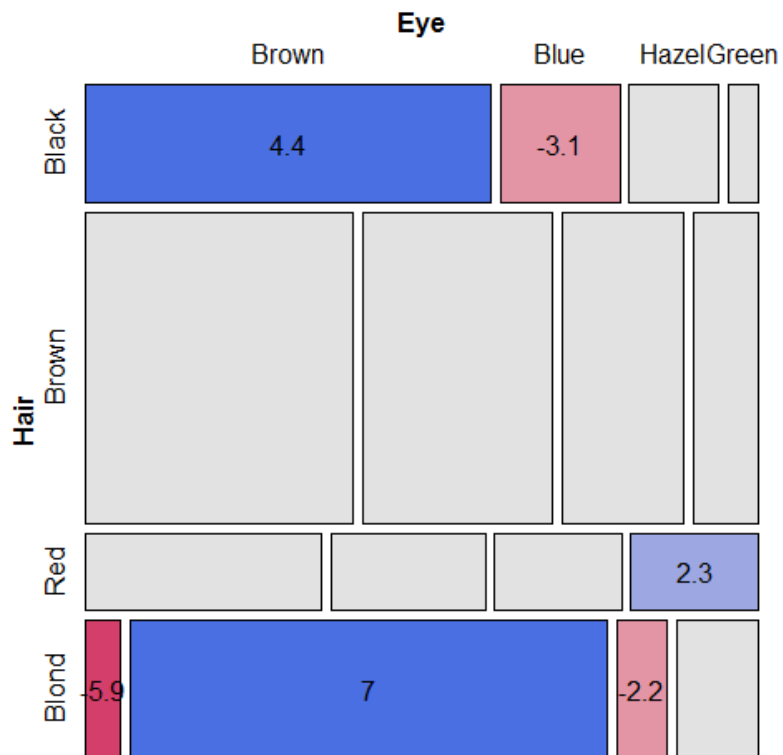
```
> haireye
```

	Eye			
Hair	Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16

```
> library(seriation)
> permute(haireye, "CA")
```

	Eye			
Hair	Brown	Hazel	Green	Blue
Black	68	15	5	20
Brown	119	54	29	84
Red	26	14	14	17
Blond	7	10	16	94

```
mosaic(haireye, shade=TRUE, labeling=labeling_residuals)
mosaic(permute(haireye, "CA"), shade=TRUE, labeling=labeling_residuals)
```



Bee abundance data

A study by Taylor Kerekes examined the abundance of bee species in Ontario over three periods of time.

Q: Does relative abundance of species differ over years?

A: Do a chi-square test

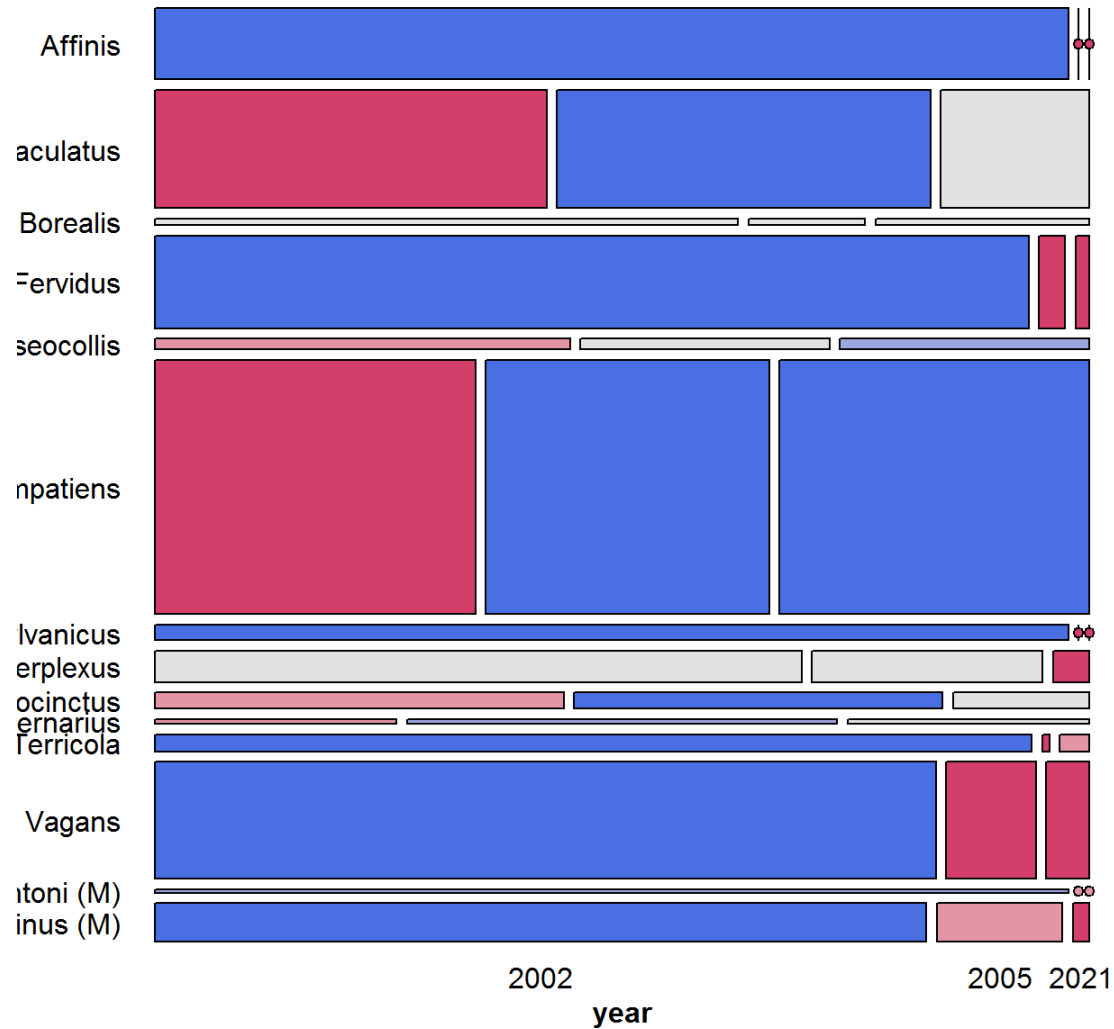
```
chisq.test(bees[, -1])  
Pearson's Chi-squared test  
  
data:  bees[, -1]  
X-squared = 1981, df = 26, p-value <2e-16
```

```
# A tibble: 14 x 4  
  species      `2002` `2005` `2021`  
  <chr>      <dbl> <dbl> <dbl>  
1 Affinis      508     0     0  
2 Bimaculatus  362    345    137  
3 Borealis     30     6     11  
4 Fervidus    634    19     10  
5 Griseocollis  35    21     21  
6 Impatiens   638   564   616  
7 Pensylvanicus 112     0     0  
8 Perplexus   160    57     9  
9 Rufocinctus  51    46    17  
10 Ternarius    9    16     9  
11 Terricola   119     1     4  
12 Vagans      713    82    39  
13 Ashtoni (M)  27     0     0  
14 Citrinus (M) 234    38     5
```

How to understand the pattern of association?

```
mosaic(bees.mat, shade=TRUE, ...)
```

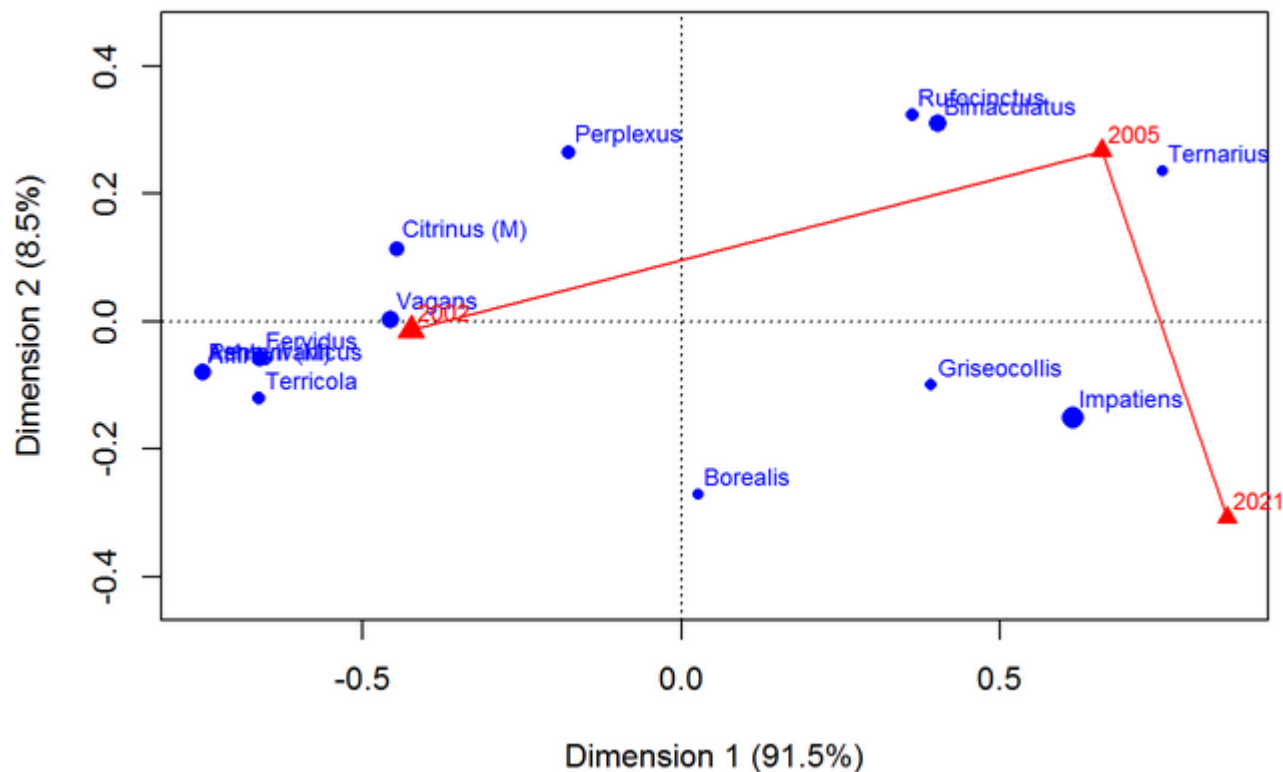
Bees Abundance Data



Alphabetic order of
species:
No clear pattern

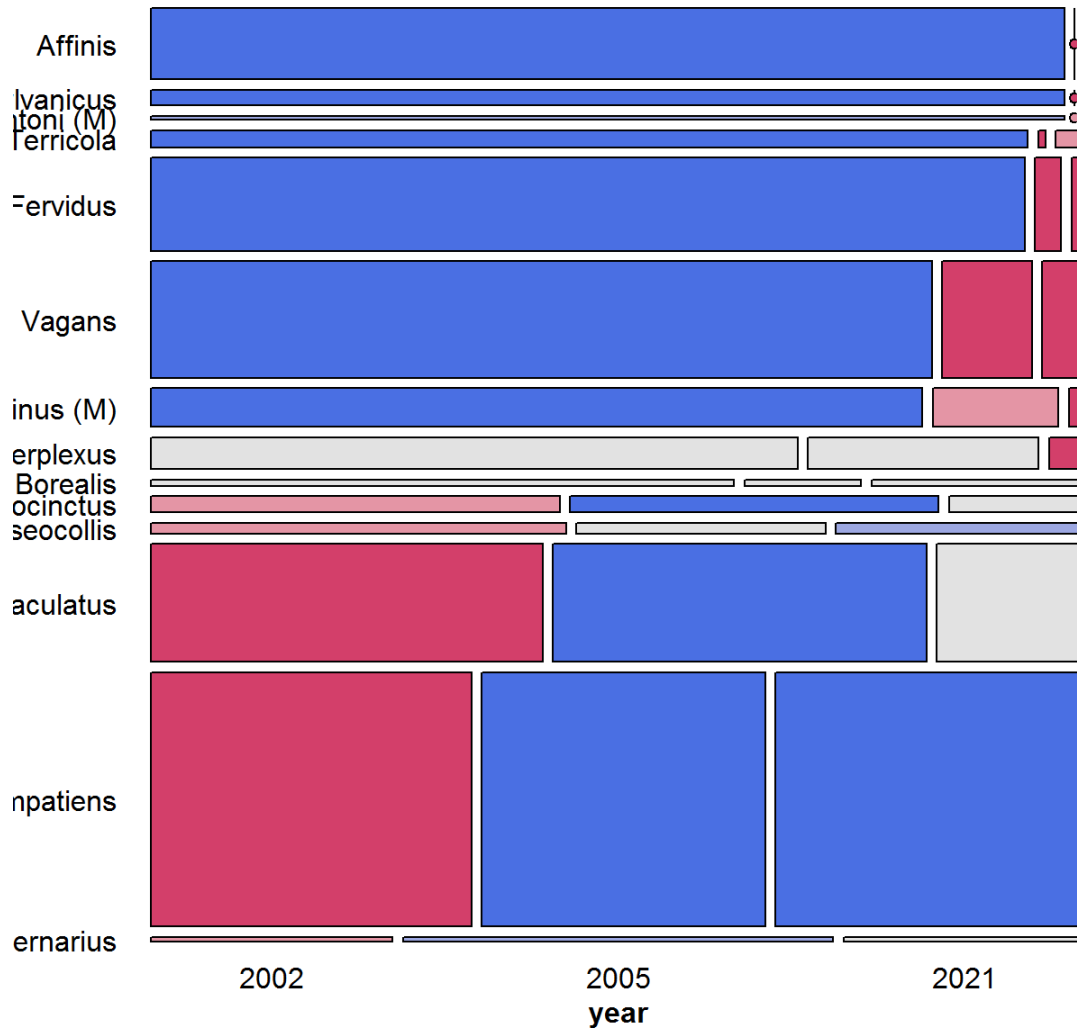
Correspondence analysis finds scores for the row & col categories to account for maximum χ^2

```
bees.ca <- ca(bees.mat)
plot(bees.ca,
     lines=c(FALSE, TRUE),      # join years with lines
     mass = c(TRUE, TRUE))     # symbol size ~ marginal frequency
```



```
mosaic(permute(bees.mat, "CA"), shade=TRUE, ...)
```

Bees Abundance Data



One main cluster was very prevalent in 2002

A few species became prominent in later years

Three-way tables

Saturated model

For a 3-way table, of size $I \times J \times K$ for variables A, B, C , the **saturated** loglinear model includes associations between all pairs of variables, as well as a 3-way association term, λ_{ijk}^{ABC}

$$\log m_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}. \quad (6)$$

- One-way terms ($\lambda_i^A, \lambda_j^B, \lambda_k^C$): differences in the *marginal frequencies* of the table variables.
- Two-way terms ($\lambda_{ij}^{AB}, \lambda_{ik}^{AC}, \lambda_{jk}^{BC}$) pertain to the *partial association* for each pair of variables, *controlling* for the remaining variable.
- The three-way term, λ_{ijk}^{ABC} allows the partial association between any pair of variables to vary over the categories of the third variable.
- Fits perfectly, but doesn't *explain* anything, so we hope for a simpler model!

Reduced models

- Goal: fit the smallest model sufficient to explain/describe the observed frequencies
 - Similar to Anova models, $\sim(A + B + C)^3$ with all interactions
- Hierarchical models
 - A high-order term, like $\lambda_{ijk}^{ABC} \rightarrow$ all lower order terms included
 - E.g. $[ABC] \rightarrow A + B + C + AB + AC + BC$
 - $[AB][AC] \rightarrow A + B + C + AB + AC$
- Thus, a shorthand notation for a loglinear model lists only the high-order terms

Reduced models

- For a three-way table there is a range of models between mutual independence, $[A][B][C]$, and the saturated model, $[ABC]$
- Each model has an independence interpretation:

$$[A][B] \equiv A \perp B \equiv A \text{ independent of } B$$

- Special names for various submodels

Table: Log-linear Models for Three-Way Tables

Model	Model symbol	Interpretation
Mutual independence	$[A][B][C]$	$A \perp B \perp C$
Joint independence	$[AB][C]$	$(A \ B) \perp C$
Conditional independence	$[AC][BC]$	$(A \perp B) \mid C$
All two-way associations	$[AB][AC][BC]$	homogeneous assoc.
Saturated model	$[ABC]$	ABC interaction

Model types

- **Joint independence:** $(AB) \perp C$, allows $A*B$ association, but asserts no $A*C$ and $B*C$ associations

$$[AB][C] \equiv \log m_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB}$$

- **Conditional independence:** $A \perp B$, controlling for C

$$[AC][BC] \equiv \log m_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$$

- **Homogeneous association:** All two-way, but each two-way is the *same* over the other factor

$$[AB][AC][BC] \equiv \log m_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$$

Model types: loglm()

Each of these have simple translations into the model formulae for loglm()

<code>loglm(~ A + B + C)</code>	<code># mutual independence</code>	<code>[A][B][C]</code>
<code>loglm(~ A * B + C)</code>	<code># joint independence</code>	<code>[AB][C]</code>
<code>loglm(~ A*C + B*C)</code>	<code># conditional independence</code>	<code>[AC][BC]</code>
<code>loglm(~ (A + B + C)^2)</code>	<code># homogeneous, all 2-way</code>	<code>[AB][AC][BC]</code>
<code>loglm(~ A * B * C)</code>	<code># saturated model</code>	<code>[ABC]</code>

Collapsibility: Marginal & conditional associations

- Q: When can we legitimately collapse a table, ABC over some variable (C)?
- A: When the **marginal** association of AB is the same as the **conditional** association, $AB \mid C$
- Recall the Berkeley data
 - Margin of Admit, Gender ignoring Dept showed strong association
 - The partial assoc. within Dept were mostly NS
 - This is an example of Simpson's paradox
- Three-way tables: The AB marginal and $AB \mid C$ conditional associations are the same, if either:
 - A & C are conditionally independent, $A \perp C \mid B = [AB][CB]$
 - B & C are conditionally independent, $B \perp C \mid A = [AB][AC]$
 - \rightarrow no three-way association

Response vs. Association models

- In **association models**, the interest is just on *which* variables are associated, and *how*
 - Hair-eye data: [Hair Eye]? [Hair Sex]? [Eye Sex]
 - \implies fit the homogeneous association model (or the saturated model)
 - Test the individual terms, delete those which are NS
- In **response models**, the interest is on which predictors are associated with the response
 - The minimal (null or baseline) model is the model of joint independence of the response (say, A) from all predictors, [A] [B C D ...]
 - Associations among the predictors are fitted exactly (not analyzed)
 - Similar to regression, where predictors can be arbitrarily correlated
 - e.g., Berkeley data: fit the baseline model [Admit] [Gender Dept]
 - lack-of-fit \implies associations [Admit Gender] and/or [Admit Dept]

Goodness of fit tests

As noted earlier, **overall** goodness of fit of a **specified** model may be tested by the likelihood ratio G^2 , or the Pearson X^2 ,

$$G^2 = 2 \sum_i n_i \log \left(\frac{n_i}{\hat{m}_i} \right) \quad X^2 = \sum_i \frac{(n_i - \hat{m}_i)^2}{\hat{m}_i} ,$$

with residual degrees of freedom $\nu = \# \text{ cells} - \# \text{ estimated parameters}$.

- These measure the **lack of fit** of a given model— a large value \implies a poor model
- Both are distributed as $\chi^2(\nu)$ (in large samples: all $\hat{m}_i > 5$)
- $\mathcal{E}(\chi^2(\nu)) = \nu$, so G^2/ν (or X^2/ν) measures lack of fit per degree of freedom (**overdispersion**)
- But: how to compare or test competing models?

Nested models & ANOVA-type tests

Two models, M_1 and M_2 are **nested** when one (say, M_2) is a special case of the other

- Model M_2 (w/ v_2 df) fits a subset of the parameters of M_1 (w/ v_1 df)
- M_2 is more restrictive – cannot fit better than M_1 : $G^2(M_2) \geq G^2(M_1)$
- The least restrictive model is the saturated model [ABC ...], w/ $G^2 = 0$

Therefore, we can test the **difference in G^2** as a specific test of the added restrictions in M_2 compared to M_1 .

- This test has a χ^2 distribution with $df = v_2 - v_1$

$$\begin{aligned}\Delta G^2 \equiv G^2(M_2 | M_1) &= G^2(M_2) - G^2(M_1) \\ &= 2 \sum n_i \log(\hat{m}_{i1} / \hat{m}_{i2})\end{aligned}\tag{7}$$

Example: Berkeley admissions

For the UC Berkeley data, with table variables [A]dmit, [D]ept and [G]ender the following models form a nested chain

$$[A][D][G] \subset [A][DG] \subset [AD][AG][DG] \subset [ADG]$$

Table: Hierarchical G^2 tests for loglinear models fit to the UC Berkeley data

Type	LLM terms	G^2	df	$\Delta(G^2)$	$\Delta(df)$	$\Pr(> \Delta(G^2))$
Mutual ind	[A][D][G]	2097.67	16			
Joint	[A][DG]	877.06	11	1220.62	5	0.0000
All 2-way	[AD][AG][DG]	20.20	5	1128.70	5	0.0000
Saturated	[ADG]	0.0	0	20.20	5	0.0011

- Only testing the **decrease** in G^2 from one model to the next
- Here, each model is significantly better than the previous
- Joint vs. all two-way: Does Admit depend on Dept and/or Gender?
- Absolute fit of all 2-way model is not terrible. Investigate this further!

Fitting these in R

`loglm()` - data in contingency table form (MASS package)

```
1 data(UCBAdmissions)
2   ## conditional independence (AD, DG) in Berkeley data
3 mod.1 <- loglm(~ (Admit + Gender) * Dept, data=UCBAdmissions)
4   ## all two-way model (AD, DG, AG)
5 mod.2 <- loglm(~ (Admit + Gender + Dept)^2, data=UCBAdmissions)
```

`glm()` - data in frequency form

```
1 berkeley <- as.data.frame(UCBAdmissions)
2 mod.3 <- glm(Freq ~ (Admit + Gender) * Dept, data=berkeley,
3             family='poisson')
```

- `loglm()` simpler for nominal variables
- `glm()` allows a wider class of models and quantitative predictors (covariates)
- `gnm()` fits models for structured association and generalized *non-linear* models
- `vcdExtra` package provides visualizations for all.

Example: Berkeley admissions

Fit the model of mutual independence, using loglm()

```
> berk.loglm0 <- loglm(~ Admit + Dept + Gender, data=UCBAdmissions)
> berk.loglm0
Call:
loglm(formula = ~Admit + Dept + Gender, data = UCBAdmissions)

Statistics:

```

	X^2	df	P(> X^2)
Likelihood Ratio	2097.7	16	0
Pearson	2000.3	16	0

Conditional independence [AD] [AG]

```
> berk.loglm1 <- loglm(~ Admit * (Dept + Gender), data=UCBAdmissions)
> berk.loglm1
Call:
loglm(formula = ~Admit * (Dept + Gender), data = UCBAdmissions)

Statistics:

```

	X^2	df	P(> X^2)
Likelihood Ratio	1148.9	10	0
Pearson	1015.7	10	0

Conditional independence, [AD] [AG]

```
> berk.loglm2 <- loglm(~ Admit + (Dept * Gender), data=UCBAdmissions)
> berk.loglm2
```

Call:

```
loglm(formula = ~Admit + (Dept * Gender), data = UCBAdmissions)
```

Statistics:

	X ²	df	P(> X ²)
Likelihood Ratio	877.06	11	0
Pearson	797.70	11	0

All two-way model, [AD] [AG] [DG]

```
> berk.loglm3 <-loglm(~(Admit+Dept+Gender)^2, data=UCBAdmissions)
> berk.loglm3
```

Call:

```
loglm(formula = ~(Admit + Dept + Gender)^2, data = UCBAdmissions)
```

Statistics:

	X ²	df	P(> X ²)
Likelihood Ratio	20.204	5	0.0011441
Pearson	18.823	5	0.0020740

ANOVA tests

These are nested. Compare with **anova()**

```
> aov1 <- anova(berk.loglm0, berk.loglm1, berk.loglm3, test="Chisq")
> aov1
LR tests for hierarchical log-linear models
```

Model 1:

~Admit + Dept + Gender

Model 2:

~Admit * (Dept + Gender)

Model 3:

~(Admit + Dept + Gender)^2

	Deviance	df	Delta(Dev)	Delta(df)	P(> Delta(Dev))
Model 1	2097.671	16			
Model 2	1148.901	10	948.770	6	0.00000
Model 3	20.204	5	1128.697	5	0.00000
Saturated	0.000	0	20.204	5	0.00114

These are tests of **relative fit**, $\Delta G^2 = G^2(M_i | M_{i-1})$

LRstats

`vcdExtra::LRstats()` gives one-line summaries of a collection of models
These are tests of **absolute** goodness of fit

```
> LRstats(berk.loglm0, berk.loglm1, berk.loglm2, berk.loglm3)
```

Likelihood summary table:

	AIC	BIC	LR	Chisq	Df	Pr(>Chisq)	
berk.loglm0	2273	2282		2098	16	<2e-16	***
berk.loglm1	1336	1352		1149	10	<2e-16	***
berk.loglm2	1062	1077		877	11	<2e-16	***
berk.loglm3	217	240		20	5	0.0011	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- AIC and BIC are GOF measures adjusted for model **parsimony**
- Not not significance tests, but **smaller is better**
- Also apply to **non-nested** models

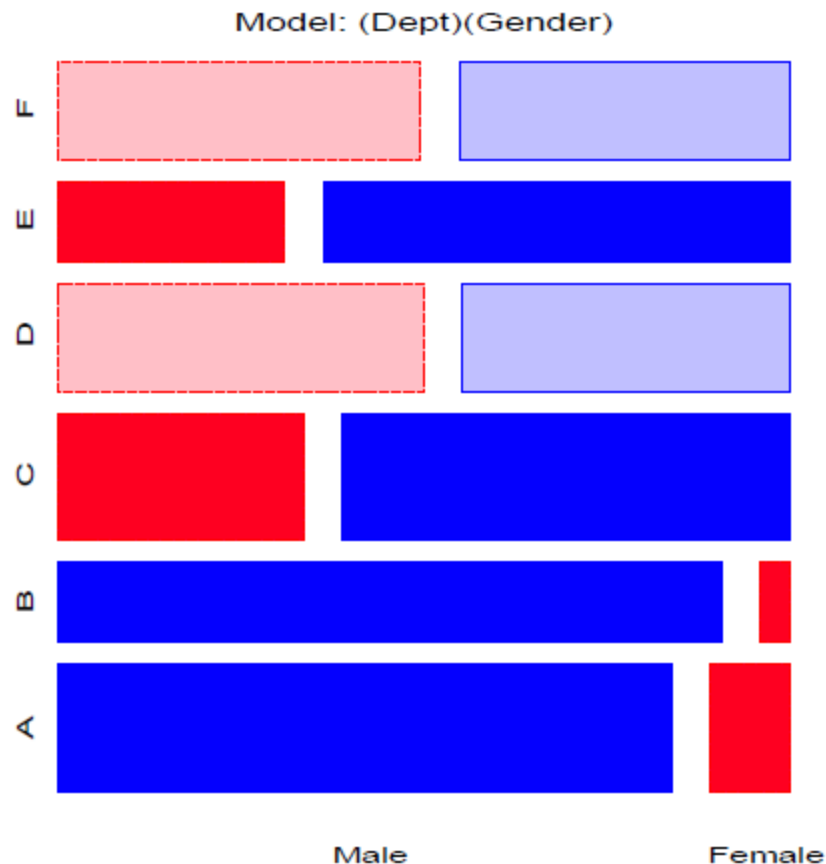
$$AIC = G^2 + 2 \times \# \text{ parameters}$$

$$BIC = G^2 + 2 \log(n) \times \# \text{ parameters}$$

Mosaic displays: Predictor variables

Berkeley data: Departments \times Gender (ignoring Admit):

- Did departments differ in the total number of applicants?
- Did men and women apply differentially to departments?



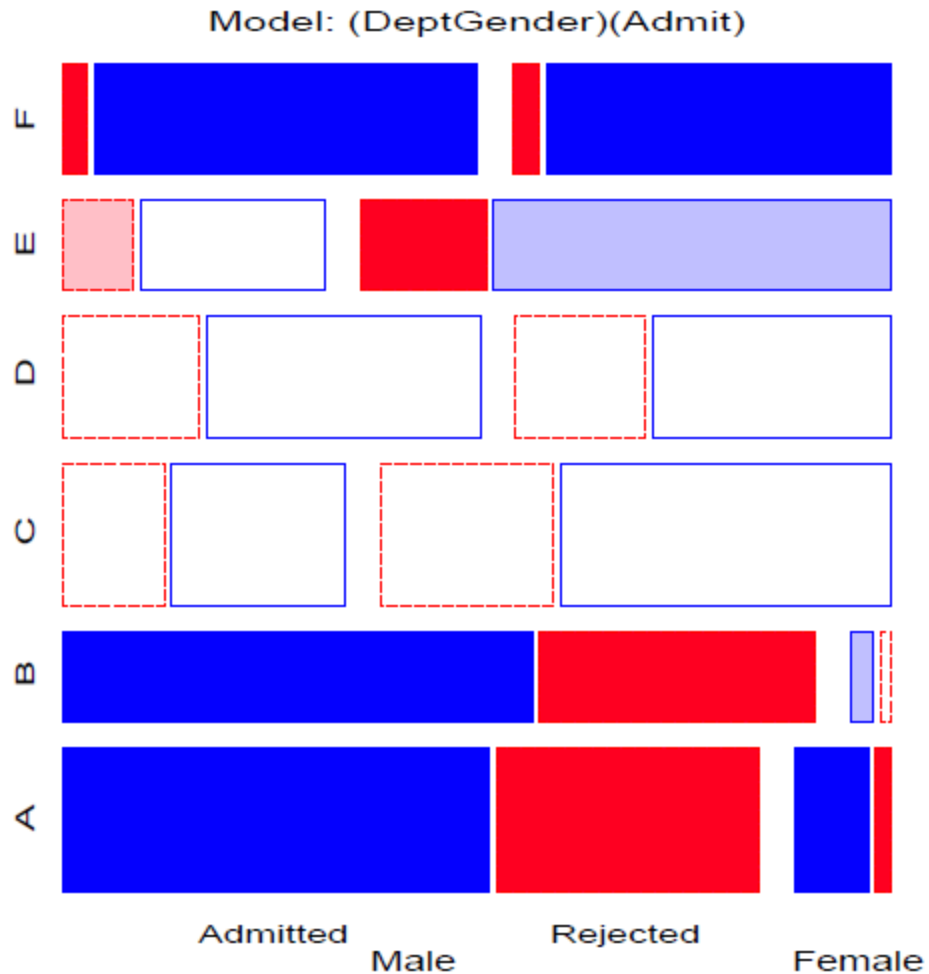
In response models, the mosaic of the predictors gives a graphic summary of background variables

- Model [Dept] [Gender]: $G^2_{(5)} = 1220.6$.
- *Note*: Departments ordered A–F by overall rate of admission.
- Men more likely to apply to departments A,B; women more likely in depts C–F

Mosaic displays: Visual fitting

- Each mosaic shows:
 - The DATA – size of tiles
 - (some) marginal frequencies – initial splits (visual grouping)
 - RESIDUALS (shading) – what associations have been omitted?
- Visual fitting
 - Start with a simple model: mutual independence or joint independence for response models
 - Pattern of residuals: suggest a better model → smaller residuals
 - Add terms: → smaller residuals, less shading: “cleaning the mosaic”
 - Good fitting model will have mostly unshaded tiles

For the Berkeley data, start with the model of [joint independence](#), [A][DG]
 Fits badly: $G^2_{(11)} = 877.1$



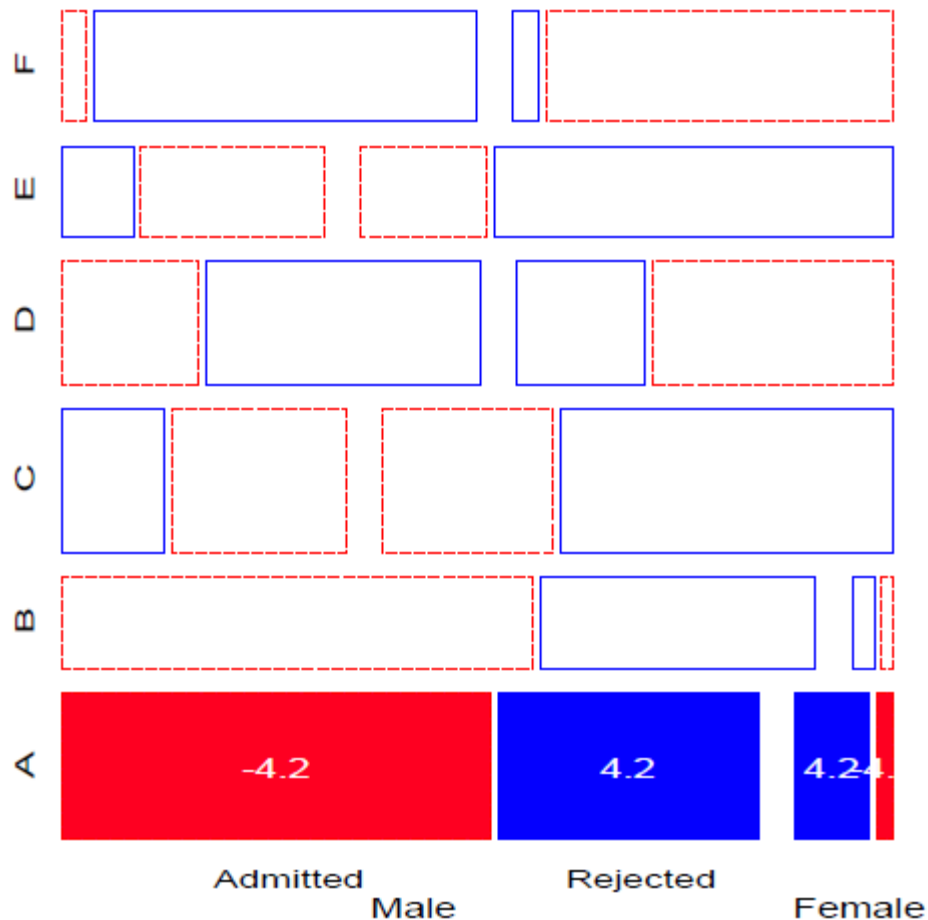
This is the [null](#), or [baseline](#) model when Admit is the response variable.

Allows assoc. of [Dept Gender], not shown in shading

Remaining shading suggests:
 [AD] : Admit varies w/ Dept
 [AG] : Admit varies w/ Gender

Conditional independence, [AD] [DG]:

Model: (DeptGender)(DeptAdmit)

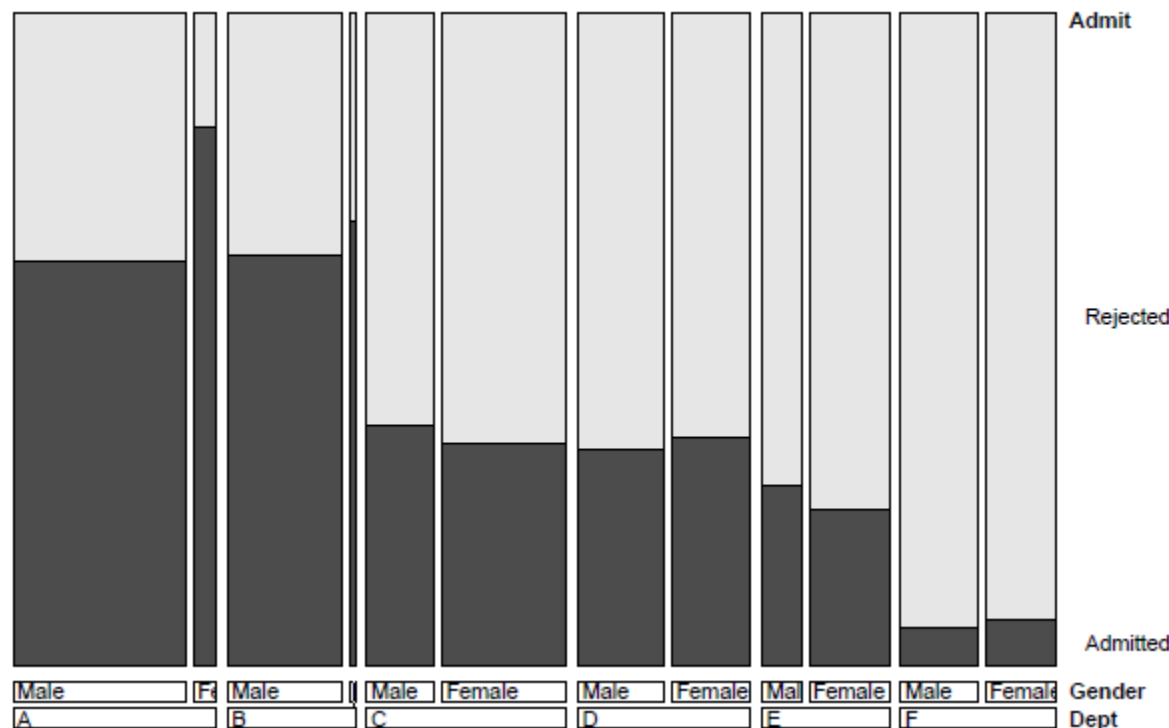


- E.g., Add [Admit Dept] association → Conditional independence:
 - Fits poorly: ($G^2_{(6)} = 21.74$)
 - But, only in Department A!
- GLM approach allows fitting a special term for Dept. A
- Note: These displays use *standardized residuals*: better statistical properties.

Double decker plots

Visualize dependence of one **response** variable (typically binary) on combinations of predictors
Formally: mosaic plots with vertical splits for all predictors, highlighting the response by shading

```
doubledecker(Admit ~ Dept + Gender, data = UCBAmissions[2:1, ,])
```



An exploratory plot

Highlights the M-F
diff^{ce} in Admit for
Dept A

DDAR Fig 5.34, p 211



Survival on the *Titanic*

4-way tables: Survival on the *Titanic*

Data on the fate of passengers & crew on the HMS Titanic: a $4 \times 2 \times 2 \times 2$ table

```
> data(Titanic, package="datasets")
> str(Titanic)
'table' num [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 ...
- attr(*, "dimnames")=List of 4
 ..$ Class      : chr [1:4] "1st" "2nd" "3rd" "Crew"
 ..$ Sex        : chr [1:2] "Male" "Female"
 ..$ Age        : chr [1:2] "Child" "Adult"
 ..$ Survived: chr [1:2] "No" "Yes"
```

What proportion survived? Ans: $711/2201 = 32.3\%$

```
> addmargins(margin.table(Titanic, 4))
Survived
  No   Yes  Sum
1490  711 2201
> margin.table(Titanic, 4) / sum(Titanic)
Survived
  No   Yes
0.677 0.323
```


Zero cells

```
> structable(Titanic)
```

		Sex		Male		Female	
		Survived		No	Yes	No	Yes
Class	Age						
1st	Child		0	5		0	1
	Adult		118	57		4	140
2nd	Child		0	11		0	13
	Adult		154	14		13	80
3rd	Child		35	13		17	14
	Adult		387	75		89	76
Crew	Child		0	0		0	0
	Adult		670	192		3	20

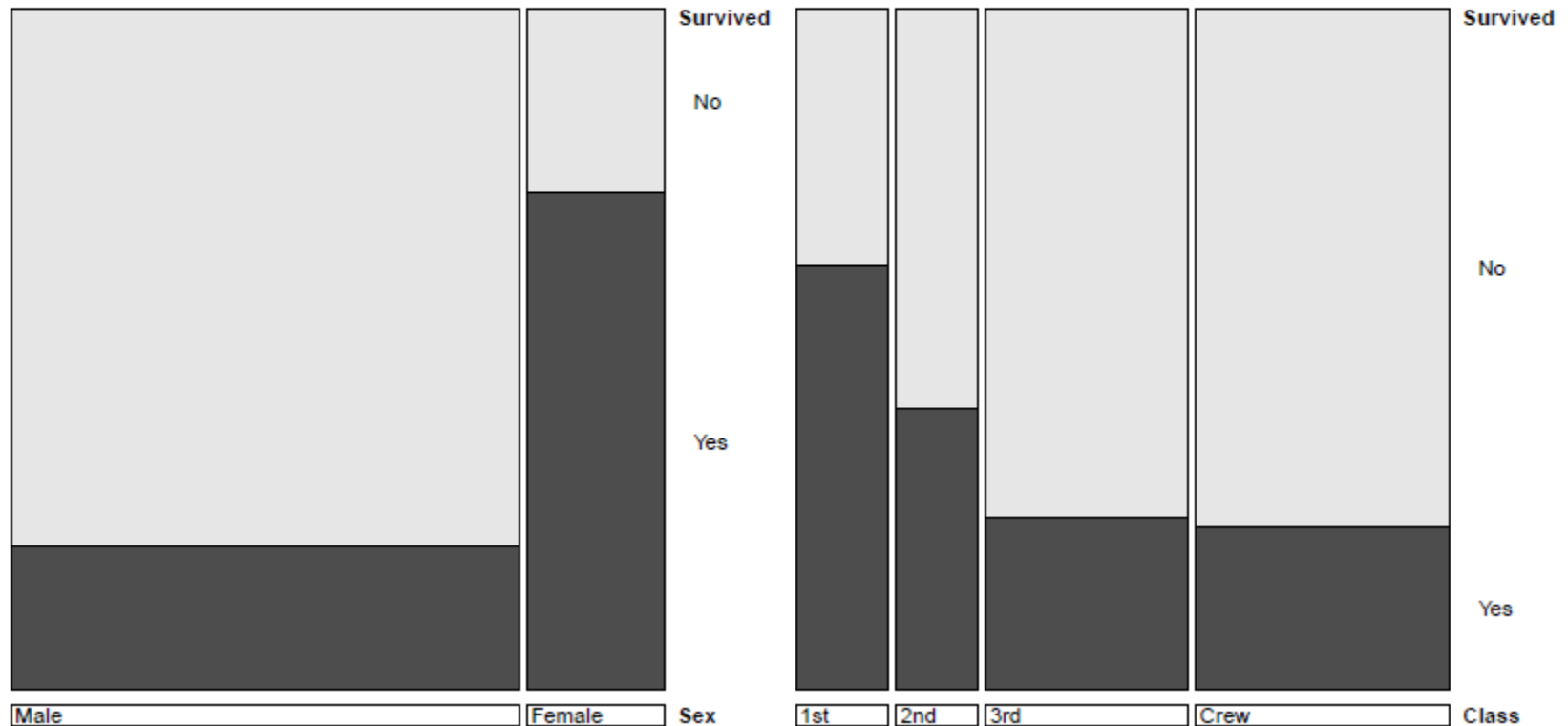
Two types of zero cells:

- **Structural zeros**: could not occur (children in crew)
- **Sampling zeros**: did not happen to occur (children in 1st & 2nd who died)
- Beware: zeros can cause problems:
 - Loss of df
 - $0/0 \rightarrow \text{NaN}$ in χ^2 tests

Exploratory plots

One-way doubledecker plots against survival show what might be expected:

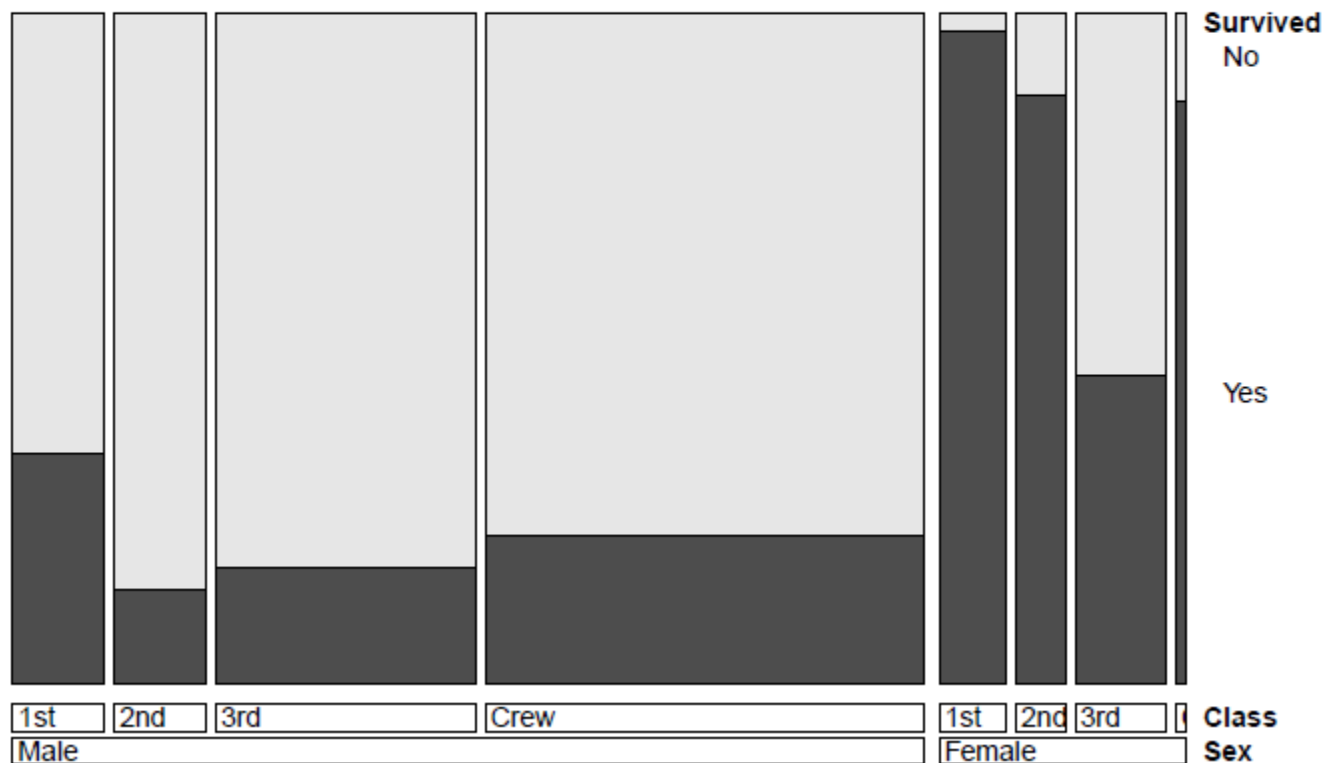
```
doubledecker(Survived ~ Sex, data=Titanic)
doubledecker(Survived ~ Class, data=Titanic)
```



Exploratory plots

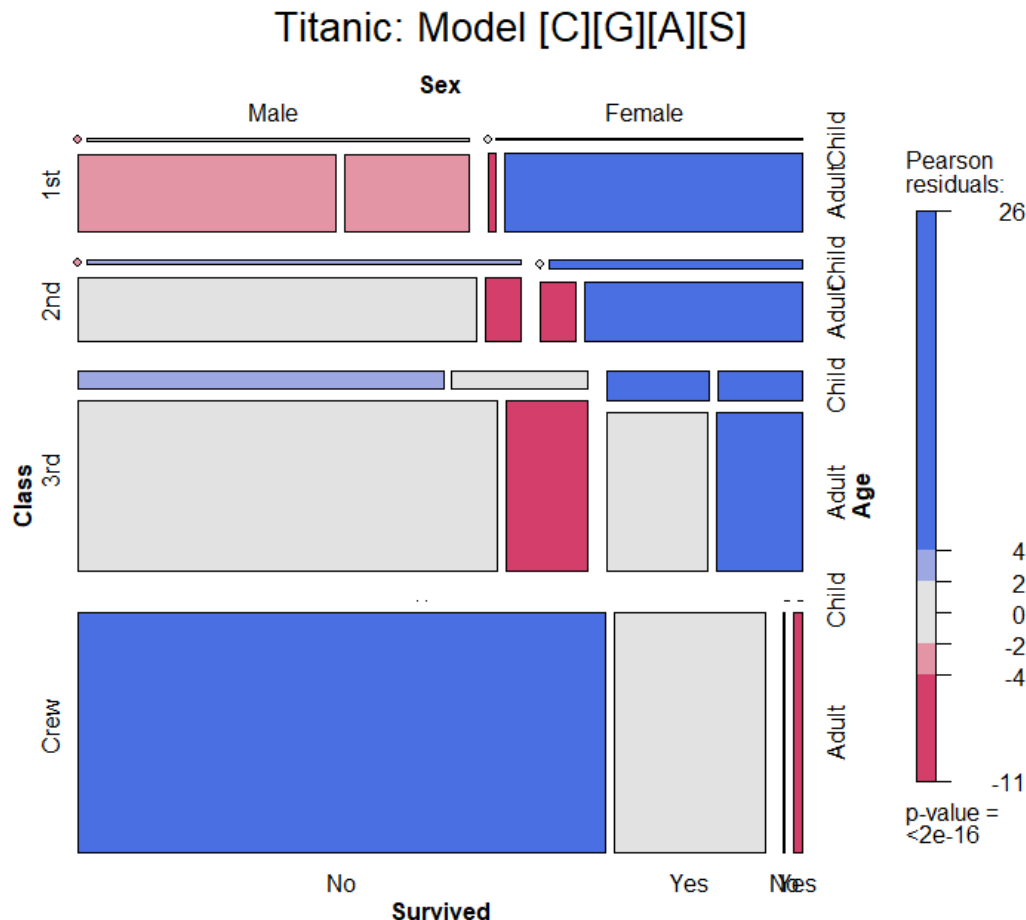
Two-way doubledecker plot against survival shows different effects of `Class` for men and women:

```
doubledecker(Survived ~ Sex + Class, data=Titanic)
```



Fitting & visualizing models

```
mod0 <- loglm(~ 1 + 2 + 3 + 4, data=Titanic)
mosaic(mod0, main="Titanic: Model [C][G][A][S]")
```



In the model formulas, I'm using variable numbers 1-4 for **C**lass, **G**ender, **A**ge and **S**urvived

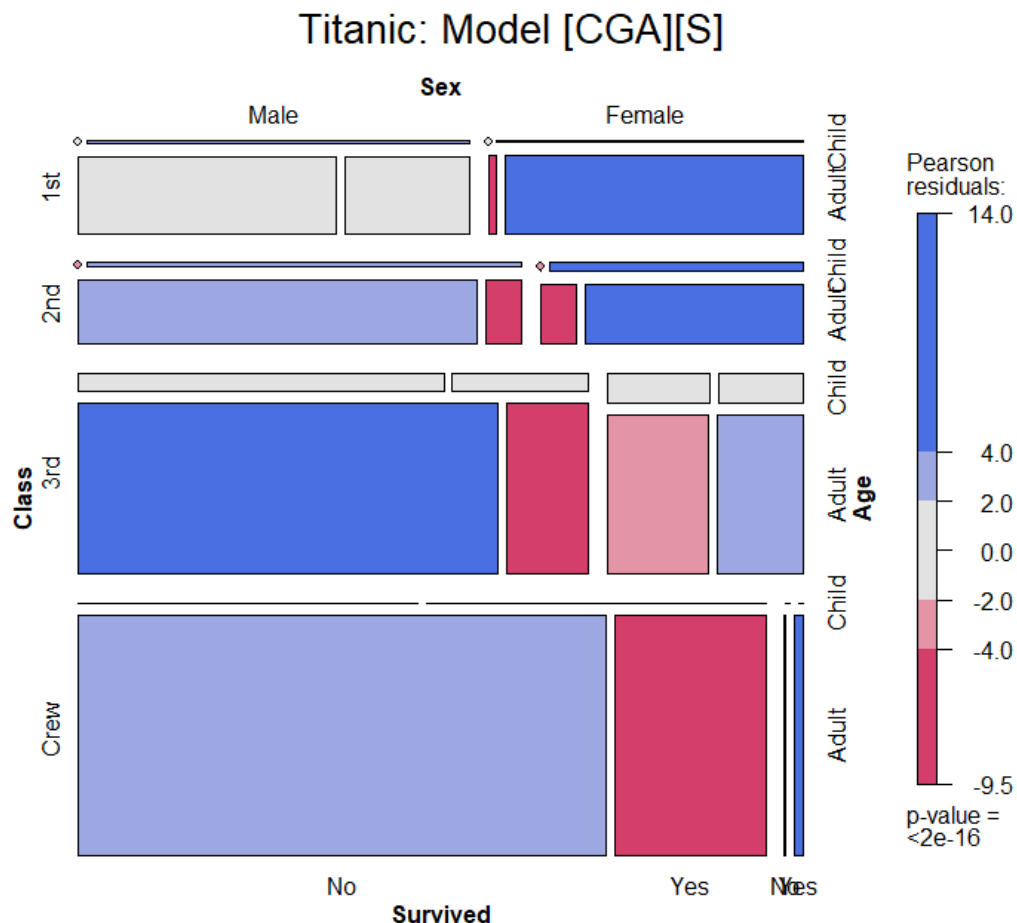
The independence model serves only as a background for the total associations in the table

Let's clean this mosaic!!

Note the scale of residuals:
+26 -- -11

Baseline model for Survived

```
mod1 <- loglm(~ 1*2*3 + 4, data=Titanic)
mosaic(mod1, main="Titanic: Model [CGA][S]")
```



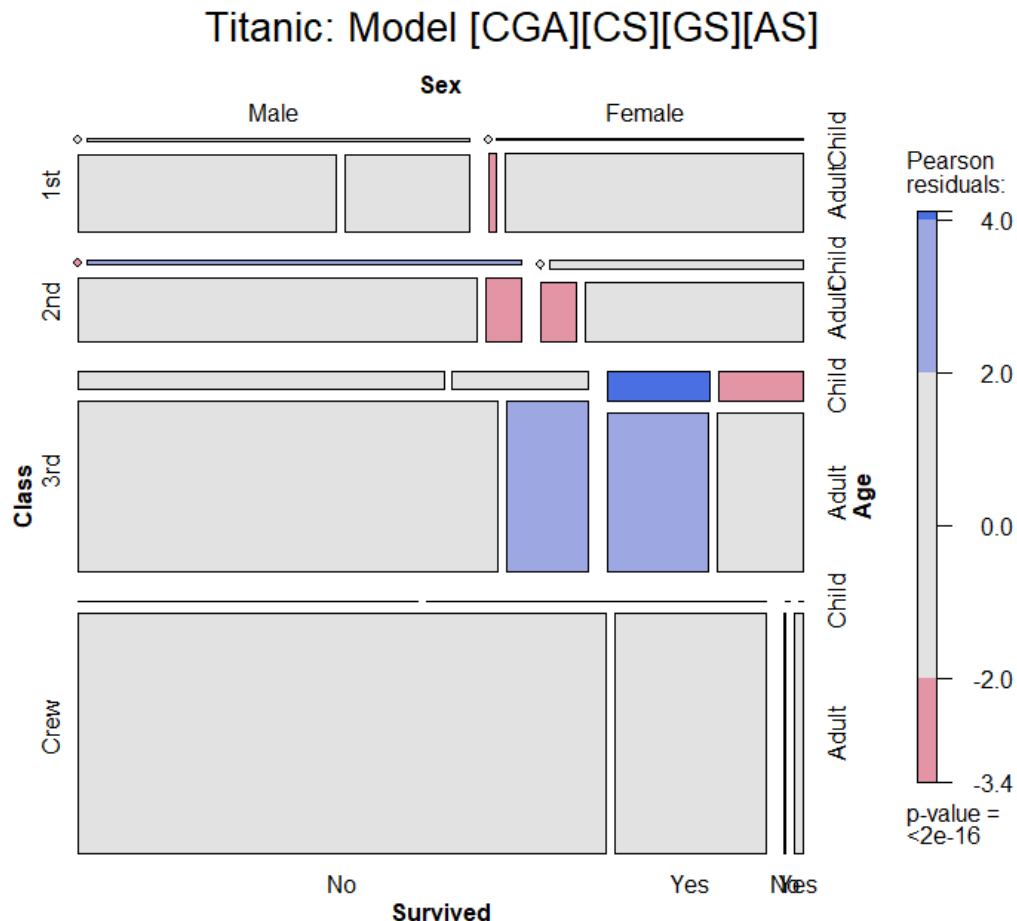
With **S** as response, the **baseline** model includes all associations among [CGA]

But this model asserts survival is independent of all of these

$G^2(15) = 671.96$, a very poor fit

Adding associations: Main effects

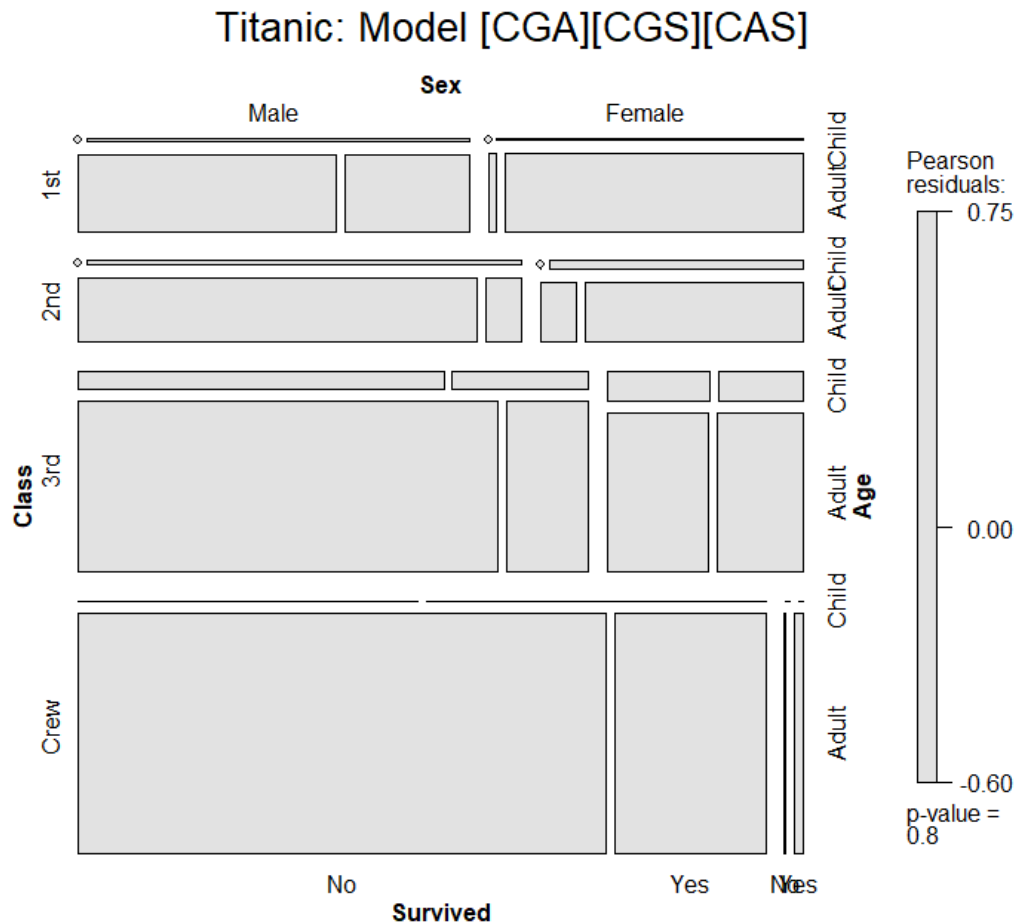
```
mod2 <- loglm(~ 1*2*3 + (1+2+3)*4, data=Titanic)
mosaic(mod2, main="Titanic: Model [CGA][CS][GS][AS]")
```



- This model allows associations of each of C, G, A with Survived
- $G^2(10) = 112.57$, still not good
- Pattern of residuals suggests 2-way interactions (3-way terms):
- “Women & children first”: suggests a term [GAS]
- Allow interactions of Class with Gender [CGS] and Class with Age [CAS]

Final model

```
mod3 <- loglm(~ 1*2*3 + (1*2)*4 + (1*3)*4, data=Titanic)
mosaic(mod3, main="Titanic: Model [CGA][CGS][CAS]")
```



Nice & clean!

$$G^2(4) = 1.69, p=0.79$$

Before accepting this,
should compare models,
and consider

- parsimony
- model explanations

Comparing models

As usual, **anova()** give compact **relative** comparisons of a set of nested models

```
> anova(mod0, mod1, mod2, mod3)
LR tests for hierarchical log-linear models
```

Model 1:

~1 + 2 + 3 + 4

Model 2:

~1 * 2 * 3 + 4

Model 3:

~1 * 2 * 3 + (1 + 2 + 3) * 4

Model 4:

~1 * 2 * 3 + (1 * 2) * 4 + (1 * 3) * 4

	Deviance	df	Delta(Dev)	Delta(df)	P(> Delta(Dev))
Model 1	1243.66	25			
Model 2	671.96	15	571.70	10	0.000
Model 3	112.57	10	559.40	5	0.000
Model 4	1.69	4	110.88	6	0.000
Saturated	0.00	0	1.69	4	0.793

Comparing models

LRstats() gives **absolute** GOF tests; also provides AIC, BIC stats: model parsimony

```
> LRstats(mod0, mod1, mod2, mod3)
Likelihood summary table:
      AIC   BIC LR Chisq Df Pr(>Chisq)
mod0 1385 1395    1244 25    <2e-16 ***
mod1  833  858    672 15    <2e-16 ***
mod2  284  316    113 10    <2e-16 ***
mod3  185  226      2  4      0.79
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

mod3 [CGA] [CGS] [CAS] wins!

- Acceptable G^2
- Looks best by AIC & BIC

Model interpretation

Recall that the goal of analysis is to tell a story

- Greatest impact: lower class → decreased survival, regardless of Gender & Age
- Differences in survival by Class were **moderated** by both Gender & Age
 - Term [CGS]: Women in 3rd class did not have an advantage, while men in 1st class did vs. other classes
 - Term [CAS]: No children in 1st or 2nd class died, but nearly 2/3 in 3rd class did
- Summary:
 - Not so much “women & children first”, rather
 - Women & children, ordered by class, and 1st class men!

Sequential plots & models

- Mosaic for an n-way table → hierarchical decomposition of association
- Joint cell probabilities are decomposed as:

$$p_{ijk\ell\ldots} = \overbrace{p_i \times p_{j|i}}^{\{v_1 v_2\}} \times p_{k|ij} \times p_{\ell|ijk} \times \cdots \times p_{n|ijk\ldots}$$

$\{v_1 v_2 v_3\}$

- First 2 terms: → mosaic for v_1, v_2
- First 3 terms: → mosaic for v_1, v_2, v_3
- ... and so on
- Roughly analogous to sequential fitting in regression: $X_1 ; X_2 | X_1 ; X_3 | X_1, X_2$
- Order of variables matters for interpretation
 - Mosaics: 1st split: easiest to see the marginal proportions
 - Mosaics: 2nd variable seen as conditional proportions, given the 1st

Sequential plots & models

- Sequential models of joint independence
 - Give an **additive** decomposition of total association – mutual independence $[v_1][v_2] \dots [v_p]$

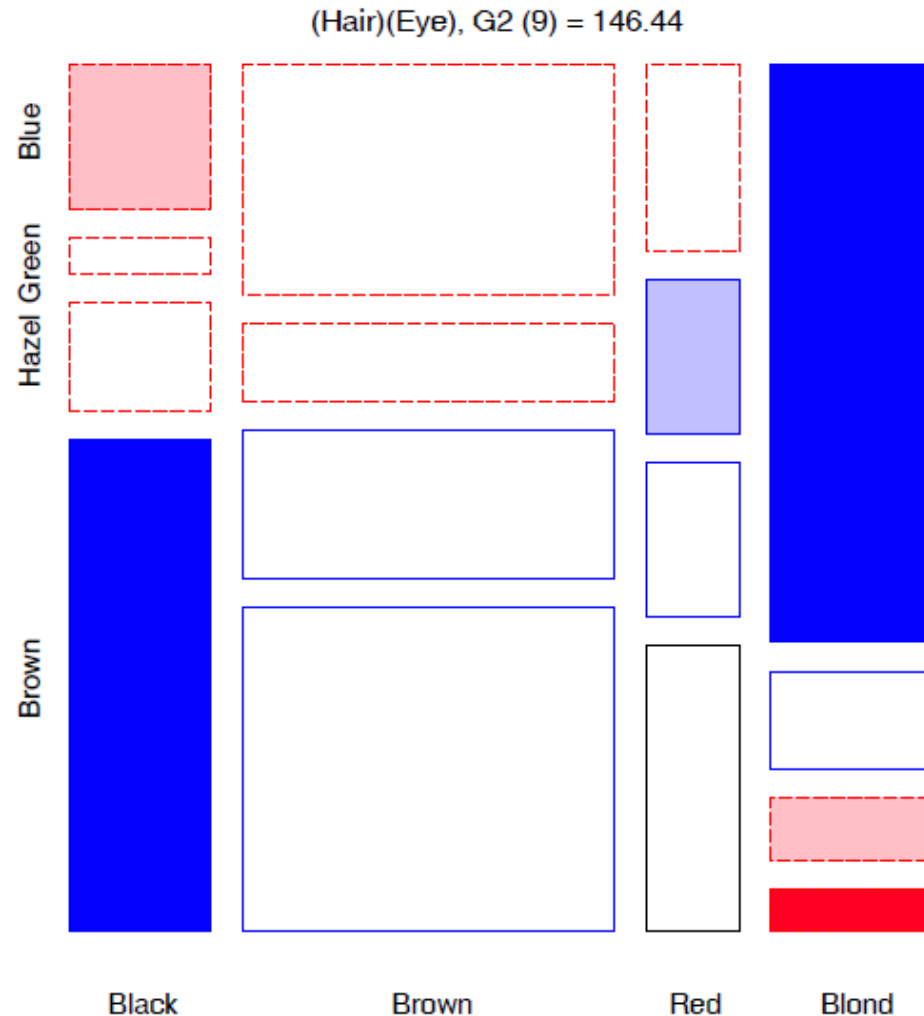
$$G^2_{[v_1][v_2]\dots[v_p]} = G^2_{[v_1][v_2]} + G^2_{[v_1 v_2][v_3]} + G^2_{[v_1 v_2 v_3][v_4]} + \dots + G^2_{[v_1 \dots v_{p-1}][v_p]}$$

- E.g., for Hair Eye color data

Model	Model symbol	df	G^2
Marginal	[Hair] [Eye]	9	146.44
Joint	[Hair, Eye] [Sex]	15	19.86
Mutual	[Hair] [Eye] [Sex]	24	166.30

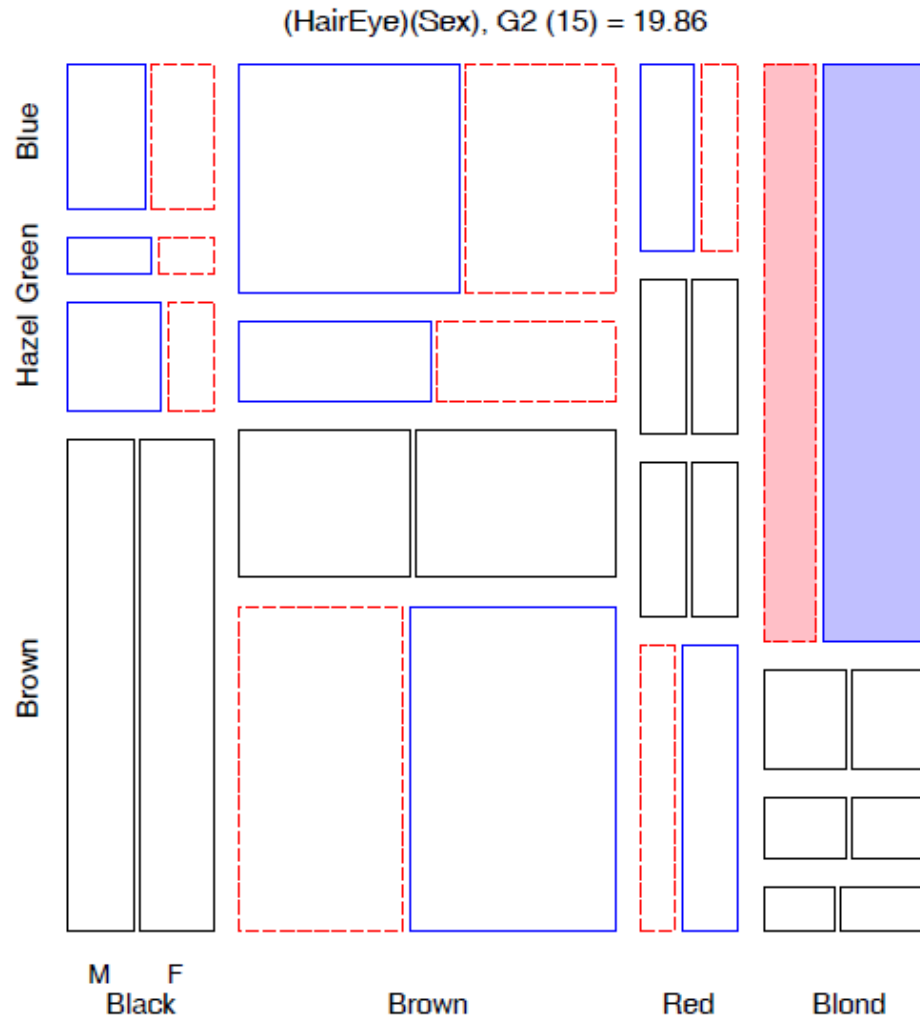
Sequential plots & models

Hair color × Eye color marginal table (ignoring Sex)



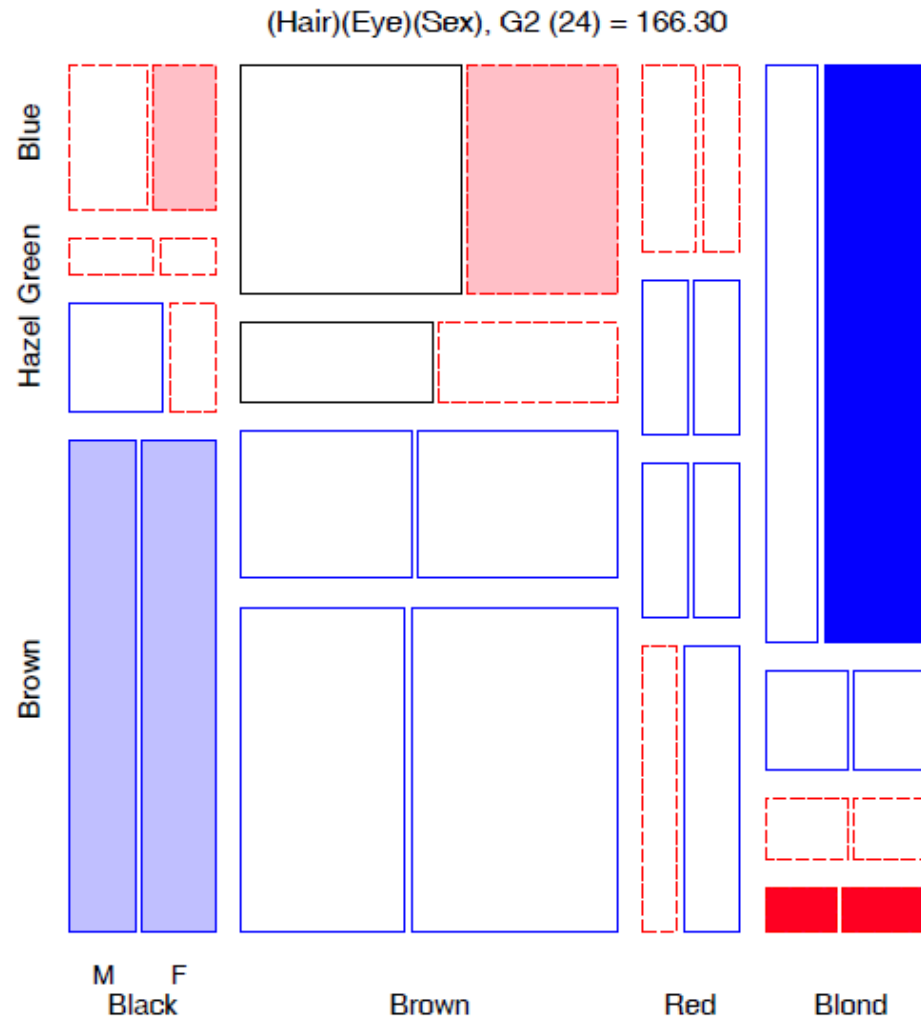
Sequential plots & models

3-way table, Joint independence model [Hair Eye][Sex]



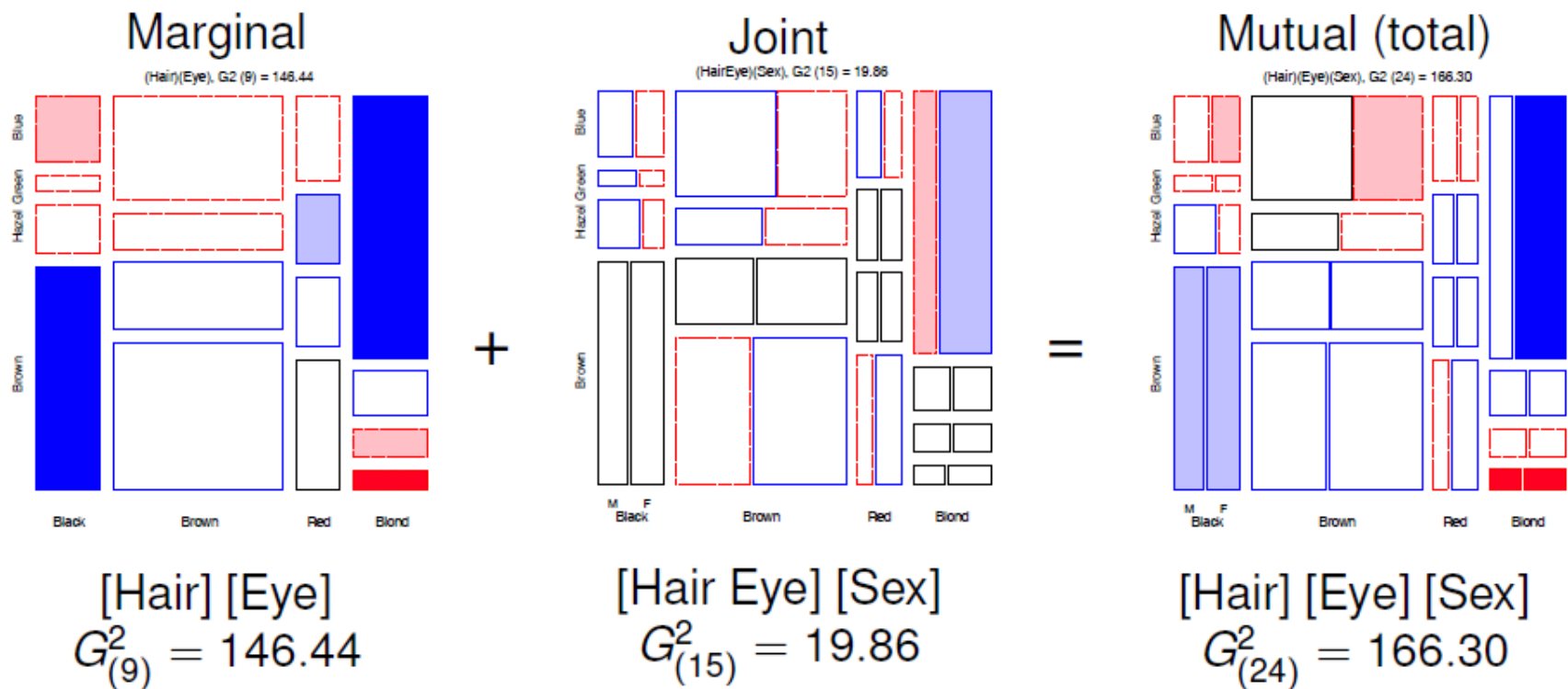
Sequential plots & models

3-way table, Mutual independence [Hair] [Eye][Sex]



Sequential plots & models

Putting these together:



Sequential models: Applications

Response models

- When one variable, R , is a response and E_1, E_2, \dots are explanatory, the baseline model is the model of joint independence, $[E_1, E_2, \dots][R]$
- Sequential mosaics then show the associations among the predictors
- The last mosaic shows all associations with R
- Better-fitting models will need to add associations of the form $[E_i R], [E_i E_j R] \dots$

Causal models

- Sometimes there is an assumed **causal ordering** of variables:

$$A \rightarrow B \rightarrow C \rightarrow D$$

- Each path of arrows: $A \rightarrow B$, $A \rightarrow B \rightarrow C$ is a sequential model of joint independence: $[A][B]$, $[AB][C]$, $[ABC][D]$.
- Testing these decomposes all joint probabilities

Example: Marital status, pre- & extra-marital sex

Thornes and Collard (1979) studied divorce patterns in relation to premarital and extramarital sex, a 2^4 table, **PreSex** in vcd ($G \times P \times E \times M$)

```
> data("PreSex", package="vcd")
> structable(Gender + PremaritalSex + ExtramaritalSex ~
             MaritalStatus, data = PreSex)
```

	Gender	Women				Men			
	PremaritalSex	Yes		No		Yes		No	
	ExtramaritalSex	Yes	No	Yes	No	Yes	No	Yes	No
MaritalStatus									
Divorced		17	54	36	214	28	60	17	68
Married		4	25	4	322	11	42	4	130

Submodels:

- ❖ $[G][P]$: Do men & women differ by pre-marital sex?
- ❖ $[GP][E]$: Given G & P, are there differences in extra-marital sex?
- ❖ $[GPE][M]$: Are there differences in divorce among the G, P, E groups?

Example: Marital status, pre- & extra-marital sex

Order the table variables as $G \rightarrow P \rightarrow E \rightarrow M$

```
> names(dimnames(PreSex))      # table variable names
[1] "MaritalStatus"  "ExtramaritalSex" "PremaritalSex"  "Gender"

> PreSex <- aperm(PreSex, 4:1)  # order variables G, P, E, M
```

Fit each sequential model to the marginal sub-table. **vcdExtra::seq_loglm()** generates these models of joint independence

```
PreSex.mods <- seq_loglm(PreSex,
                        type="joint",
                        marginals = 2:4)

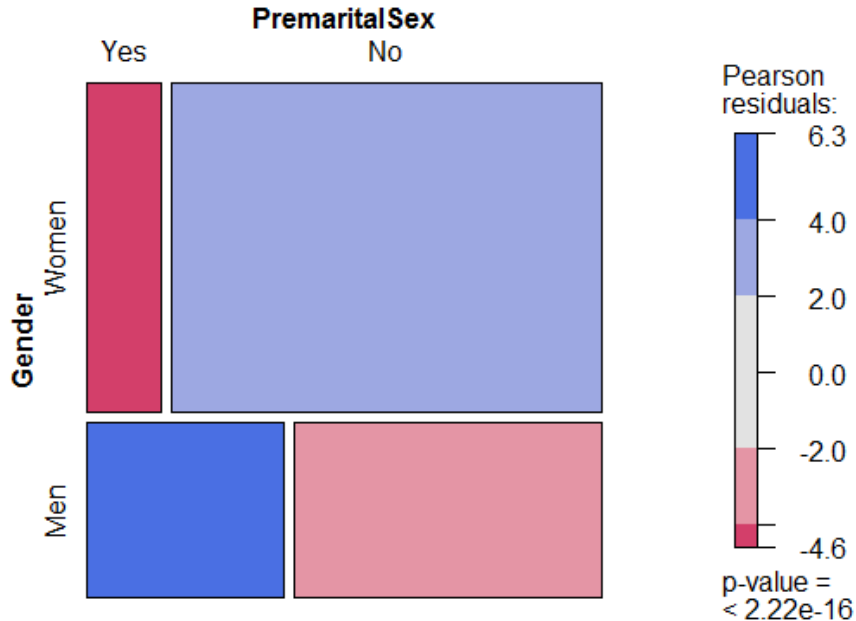
LRstats(PreSex.mods)
```

Model	df	G^2
[G] [P]	1	75.259
[GP] [E]	3	48.929
[GPE] [M]	7	107.956
[G] [P] [E] [M]	11	232.142

Mosaic plots

```
# (Gender Pre)
mosaic(margin.table(PreSex, 1:2), shade=TRUE,
       main = "Gender and Premarital Sex")
```

Gender and Premarital Sex

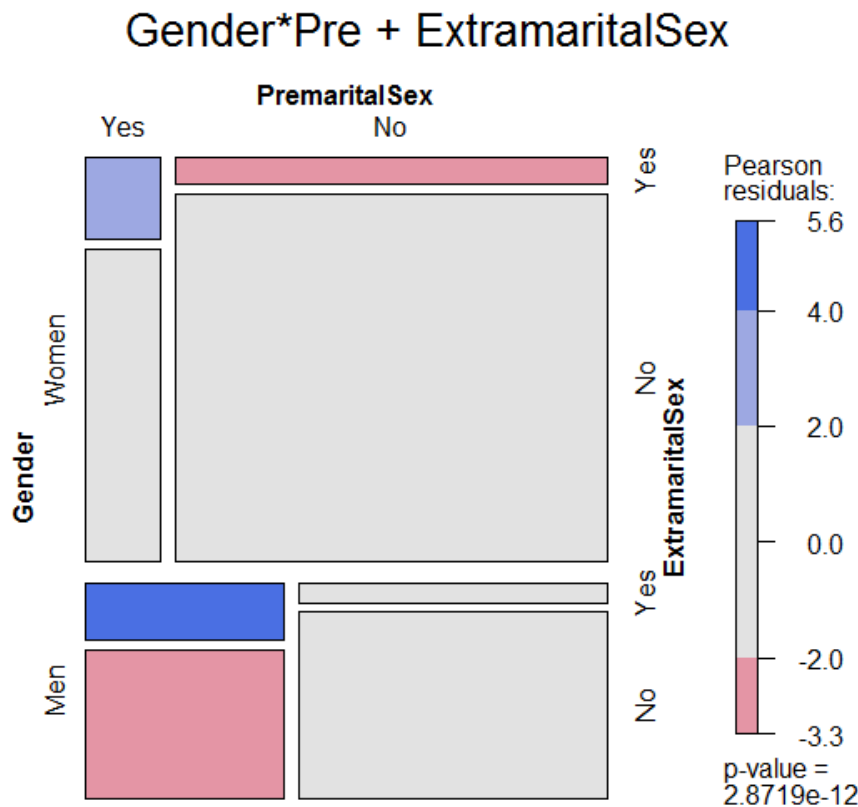


Twice as many women in this sample

Men far more likely to report pre-marital sex than women (odds ratio = 3.7)

Mosaic plots

```
# (Gender Pre) (Extra)
mosaic(margin.table(PreSex, 1:3),
       expected = ~Gender * PremaritalSex + ExtramaritalSex,
       main = "Gender*Pre + ExtramaritalSex")
```

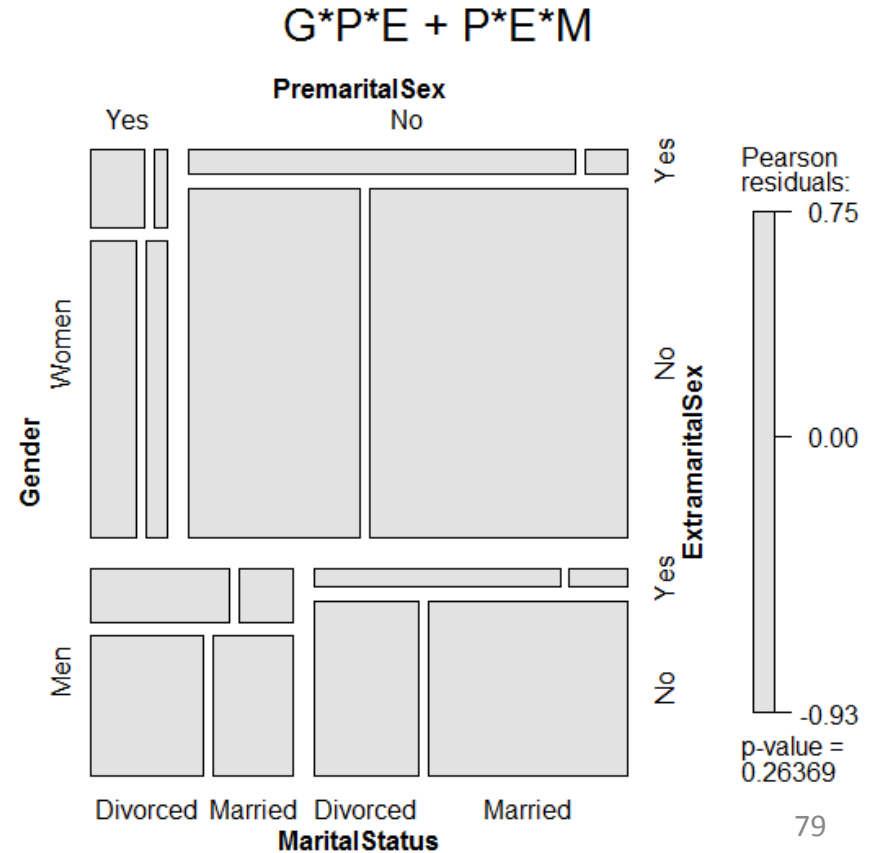
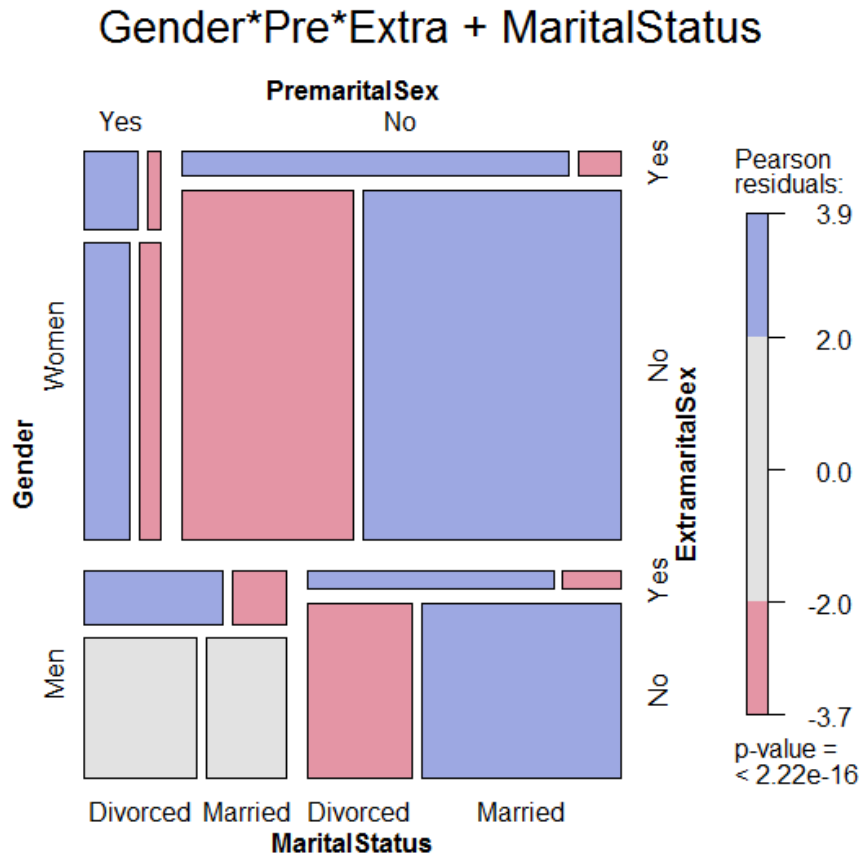


Men & women who reported Pre-
far more likely to report Extra- sex

Odds ratio of Extra- given Pre-
about the same for men & women
(3.61 vs. 3.56)

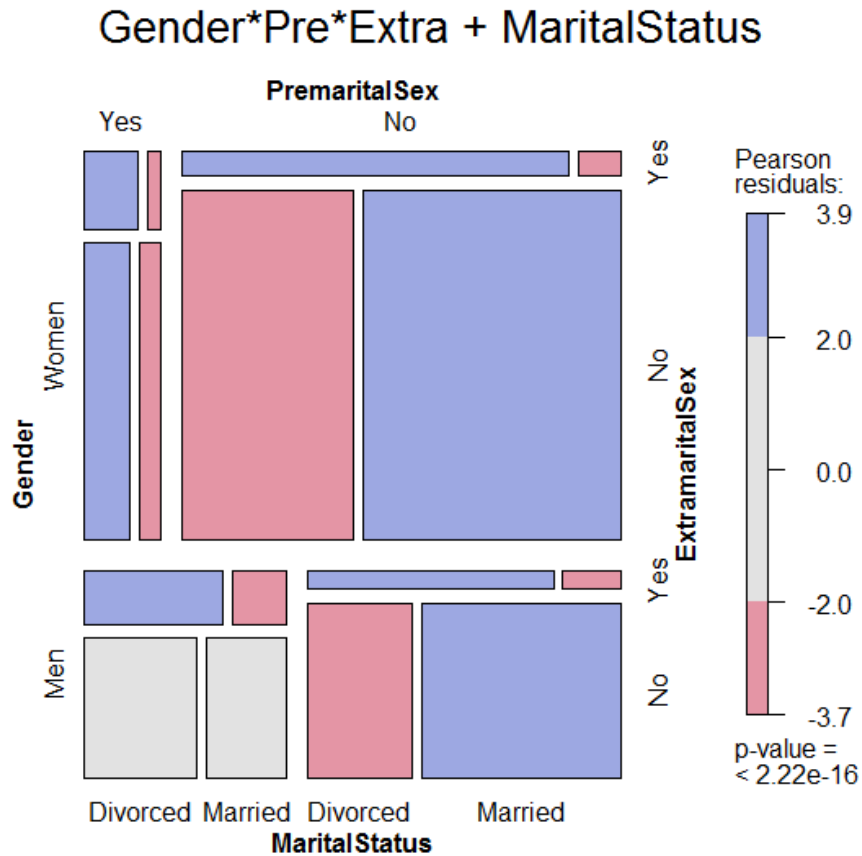
Mosaic plots

```
mosaic(PreSex,
      expected = ~Gender * PremaritalSex * ExtramaritalSex
      + MaritalStatus,
      main = "Gender*Pre*Extra + MaritalStatus")
# (GPE) (PEM)
mosaic(PreSex,
      expected = ~ Gender * PremaritalSex * ExtramaritalSex
      + MaritalStatus * PremaritalSex * ExtramaritalSex,
      main = "G*P*E + P*E*M")
```



Mosaic plots

```
mosaic(PreSex,  
       expected = ~Gender * PremaritalSex * ExtramaritalSex  
       + MaritalStatus,  
       main = "Gender*Pre*Extra + MaritalStatus")
```



In the model [GPE][M], marital status depends in a complex way

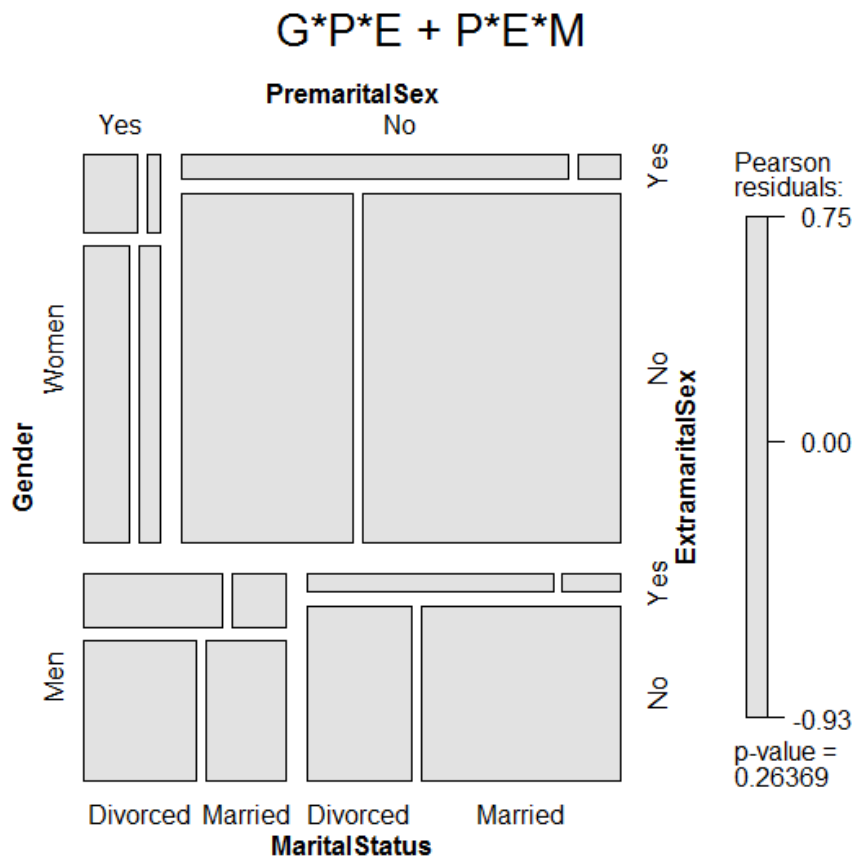
Among women, those reporting Pre-
more likely to be divorced

Among men, those reporting Pre-
only more likely to be divorced if Extra-

This suggests adding associations of
M with P and E: [PEM] term

Mosaic plots

```
# (GPE) (PEM)
mosaic(PreSex,
      expected = ~ Gender * PremaritalSex * ExtramaritalSex
      + MaritalStatus * PremaritalSex * ExtramaritalSex,
      main = "G*P*E + P*E*M")
```



This model fits well, $G^2(4) = 5.26$, $p=0.26$

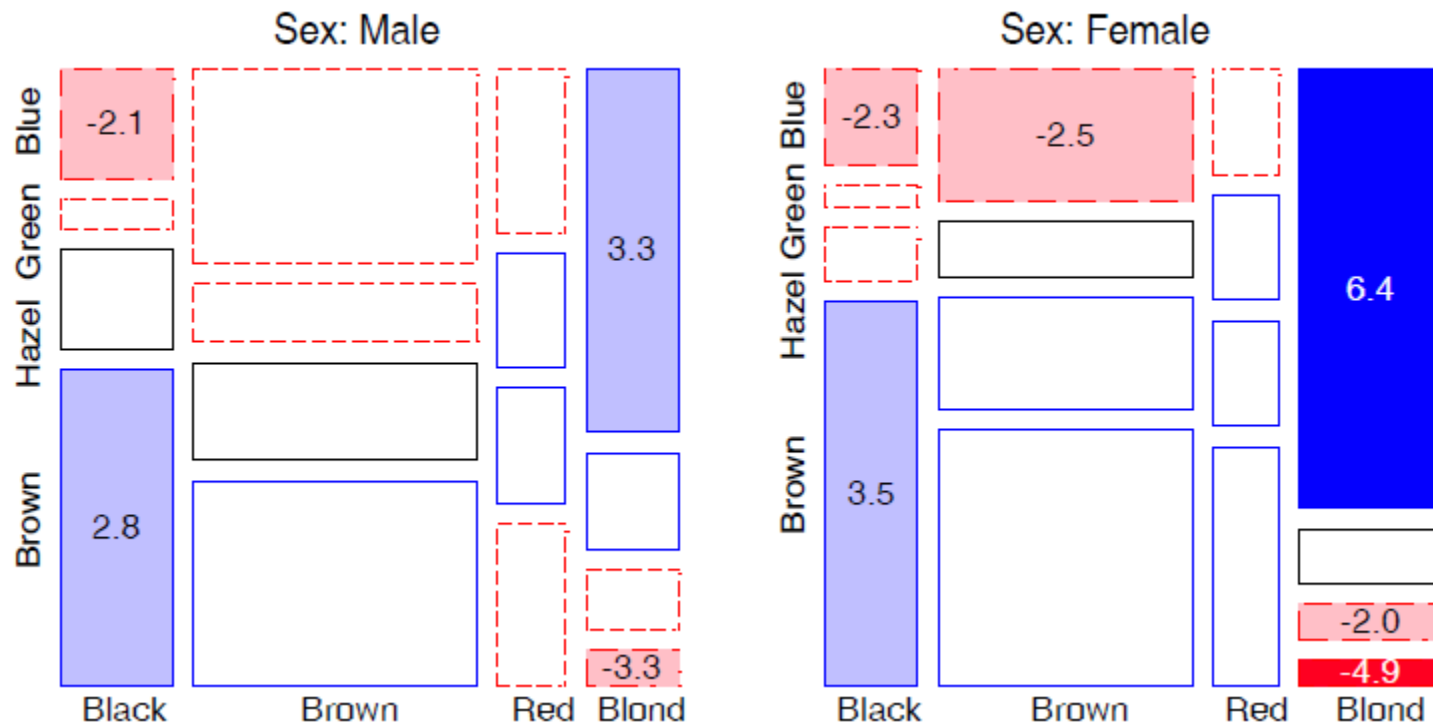
Loglinear thinking: once we take GPE into account, are there simpler models for association with M?

Looking forward: logit models for MaritalStatus often provide an easier path

Partial association, partial mosaics

Sometimes useful to do a [stratified analysis](#)

- How does association between two (or more) variables vary over levels of other variables?
- Mosaic plots for main variables show [partial association](#) at each level of others
- E.g., Hair color, Eye color, subset by Sex



Partial association, partial mosaics

Stratified analysis: conditional decomposition of G^2

- Fit models of partial (conditional) independence, $A \perp B \mid C_k$ at each level of (controlling for) C .
- \Rightarrow partial G^2 's add to the overall G^2 for conditional independence, $A \perp B \mid C$

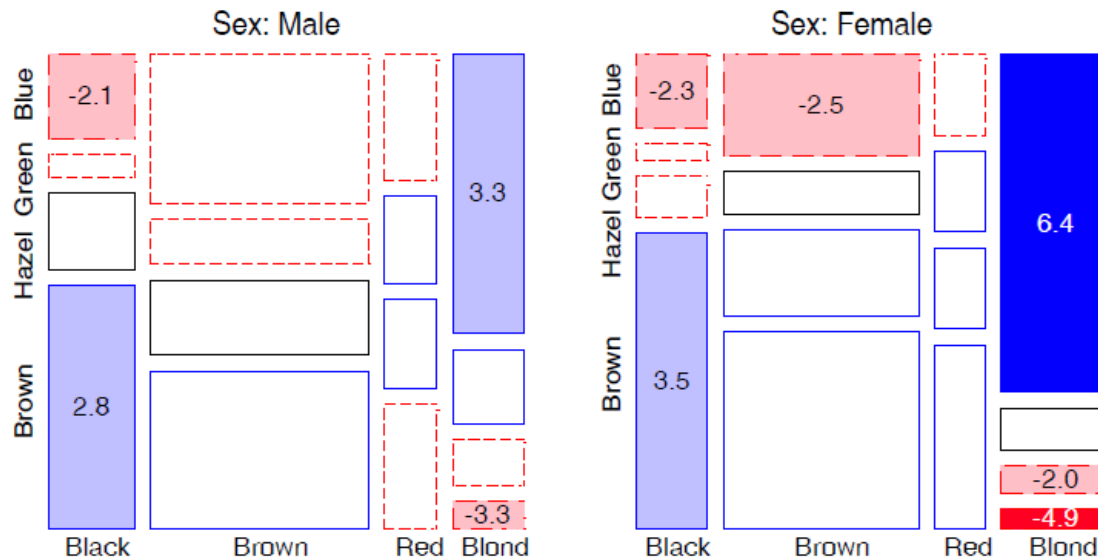
$$G^2_{A \perp B \mid C} = \sum_k G^2_{A \perp B \mid C(k)}$$

Table: Partial and Overall conditional tests, $Hair \perp Eye \mid Sex$

Model	df	G^2	p -value
$[Hair][Eye] \mid \text{Male}$	9	44.445	0.000
$[Hair][Eye] \mid \text{Female}$	9	112.233	0.000
$[Hair][Eye] \mid \text{Sex}$	18	156.668	0.000

Partial association: Summary

- Overall, there is a strong association of hair color and eye color, controlling for sex, $G^2(18) = 156.67$
 - For F, $G^2(9) = 112.23$ accounts for 72% of this association
- The pattern of association is similar for M & F
 - The largest difference is for blue-eyed blonds, much more prevalent among F than M. Is there a hair dye effect?



Summary: What we've learned

- Mosaic plots use sequential splits to show marginal and conditional frequencies in an n-way table
 - Shading: sign and magnitude of residuals, contributions to χ^2
 - Shows the pattern of association not accounted for
- Loglinear models
 - Express associations with ANOVA-like interaction terms: $A*B$, $A*C$
 - Joint independence: $[AB][C] \equiv A * B + C$
 - Conditional independence: $[AC][BC] \equiv A \perp B \mid C$
 - Fitting models \cong “cleaning the mosaic”
 - Response models: include all associations among predictors
- Sequential plots & models: Decompose all associations, $V1$, $V2 \mid V1$, $V3 \mid \{V1, V2\}$, ...