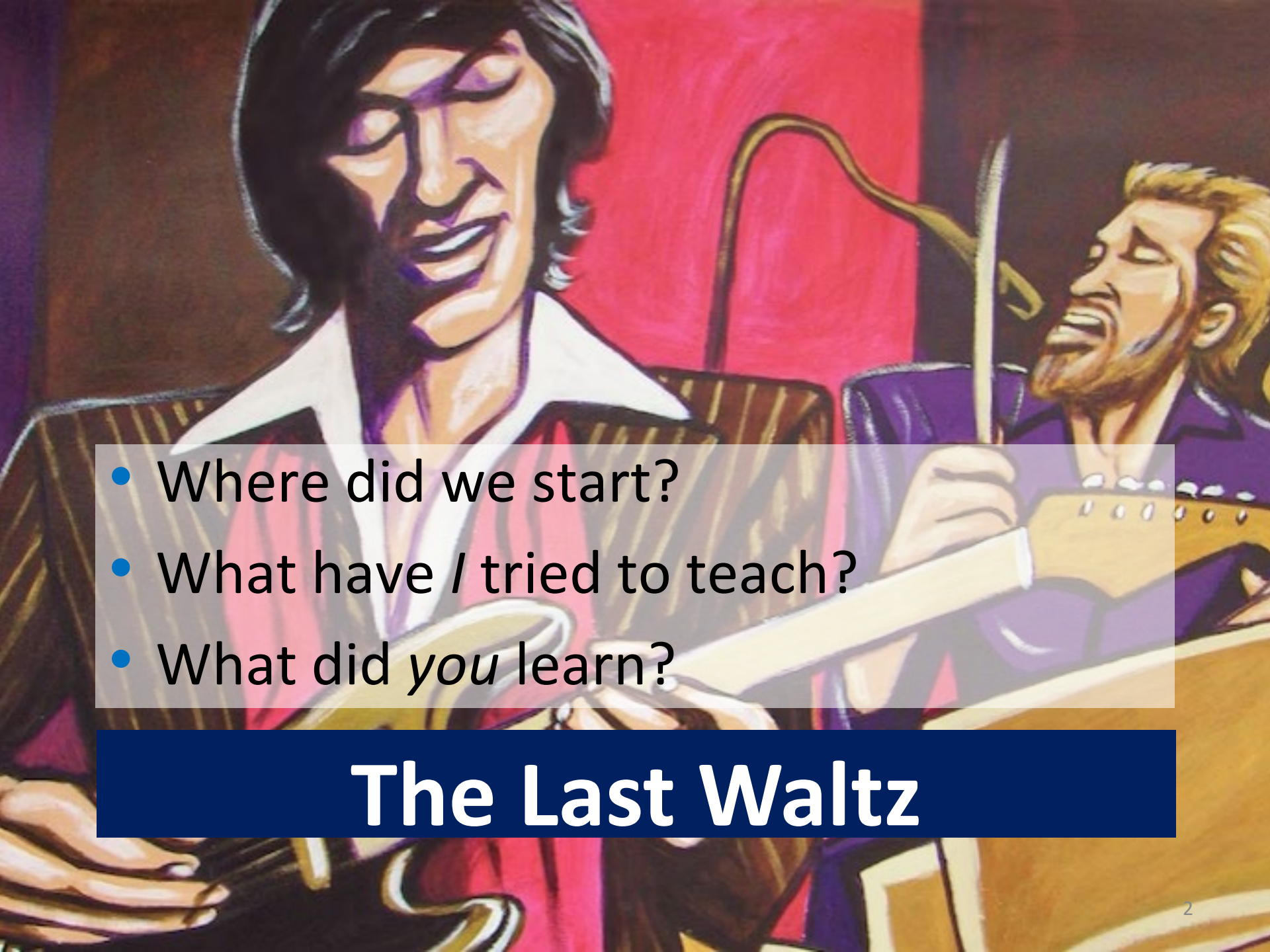
The Last Waltz

Michael Friendly

Psych 6136

http://friendly.github.io/6136



- Where did we start?
- What have *I* tried to teach?
- What did *you* learn?

The Last Waltz

01: Overview

- Categorical data involves some new ideas
  - Discrete variables: unordered or ordered
  - Counts, frequencies as outcomes
- New / different data structures & functions
  - tables – 1-way, 2-way, 3-way, … `table()`, `xtabs()`
  - similar in matrices or arrays `matrix()`, `array()`
  - datasets:
    - frequency form
    - case form
- Graphical methods: often use area ~ Freq
  - Consider: graphical comparisons, effect order
- Models: Most are ≅ natural extensions of lm()

Categorical data: Structures

Categorical (frequency) data appears in various forms

- Tables: often the result of `table()` or `xtabs()`
  - 1-way
  - 2-way – 2 × 2, r × c
  - 3-way

Gender compared to handedness

	Handed		
	Left	Right	
Female	7	46	53
Male	5	63	68
	12	109	121

margins

- Matrices: `matrix()`, with row & col names
- Arrays: `array()`, with `dimnames()`

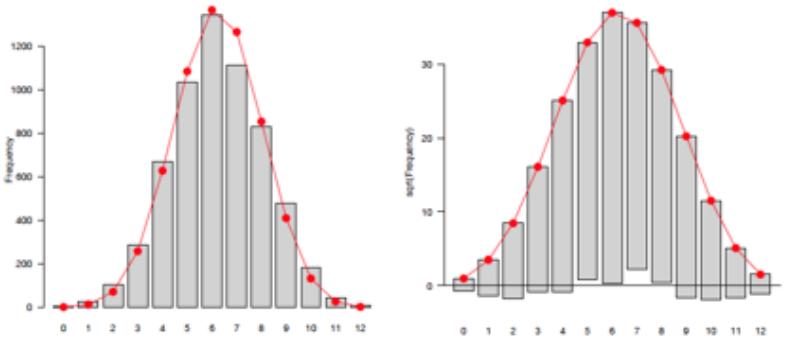
- Data frames
  - Case form (individual observations)
  - Frequency form

```
  Hair  Eye Freq
1 Black Brown   68
2 Brown Brown  119
3   Red Brown   26
4 Blond Brown    7
5 Black  Blue   20
6 Brown  Blue   84
7   Red  Blue   17
8 Blond  Blue   94
```

1-way tables: graphs

For a particular distribution in mind:

- Plot the data together with the fitted frequencies
- Better still: **hanging rootogram**: freq on sqrt scale; hang bars from fitted values



Effect ordering: Frequency tables

- Effect ordering and high-lighting for tables

Table: Hair color - Eye color data: Effect ordered

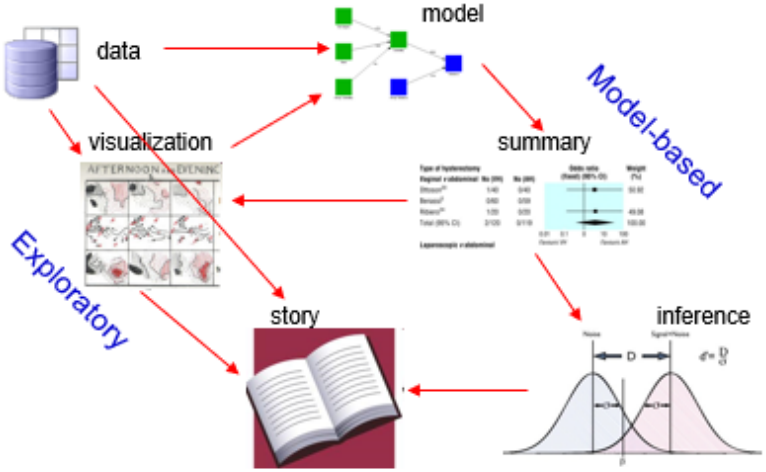
Eye color	Hair color: Black	Brown	Red	Blond
Brown	68	119	26	7
Hazel	15	54	14	10
Green	5	29	14	16
Blue	20	84	17	94

Model:	*Independence*: [Hair][Eye] χ^2 (9)= 138.29						
Color coding:	<-4	<-2	<-1	0	>1	>2	>4
n in each cell:	*n* < expected				*n* > expected		

The pattern is clearer when the eye colors are permuted: light hair goes with light eyes & vice-versa

Data, pictures, models & stories

Now, tell the story!



02: Discrete distributions

- Discrete distributions are the building blocks for categorical data analysis
  - Typically consist of basic counts of occurrences, with varying frequencies
  - Most common: binomial, Poisson, negative binomial
  - Others: geometric, log-series
- Fit with goodfit(); plot with rootogram()
  - Diagnostic plots: Ord_plot(), distplot()
- Models with predictors
  - Binomial $\rightarrow$ logistic regression
  - Poisson $\rightarrow$ poisson regression; logliner models
  - These are special cases of generalized linear models

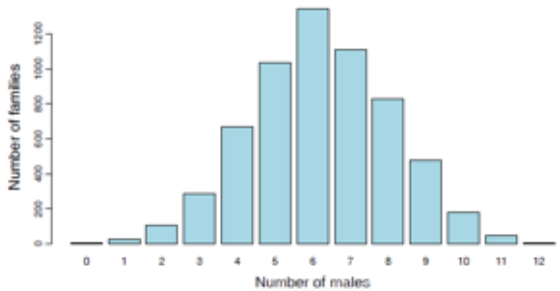
Examples: binomial

Human sex ratio (Geissler, 1889): Is there evidence that Pr(male) = 0.5?

Saxony families

Saxony families with 12 children having $k = 0, 1, \ldots 12$ sons.

k	0	1	2	3	4	5	6	7	8	9	10	11	12
n_k	3	24	104	286	670	1033	1343	1112	829	478	181	45	7



Common discrete distributions

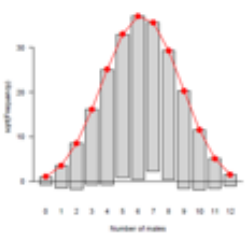
Distribution	Counts, k	Values of X	Pr(X=k)	Mean, E(X)	Var, V(X)
Bernoulli(p)	Success in 1 trial	k={0, 1}	$p^k(1-p)^{1-k}$	p	$p(1-p)$
Binomial(n,p)	# successes in n trials	0, 1, ..., n	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
Geometric(p)	# of trials to 1st success	0, 1, 2, ...	$p(1-p)^k$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Neg. binomial(k,p)	# of trials to kth success	0, 1, 2, ...	$\binom{n+k-1}{k}p^n(1-p)^k$	$\frac{k(1-p)}{p}$	$\frac{k(1-p)}{p^2}$
Poisson(λ)	# of events in interval	0, 1, 2, ...	$\frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ
Log series(p)	# of types observed	0, 1, 2, ...	$\frac{p^k}{n\log(1-p)}$		

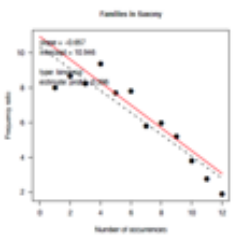
Graphing discrete distributions

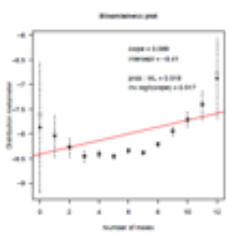
Rootograms

Ord plots

Robust distribution plots



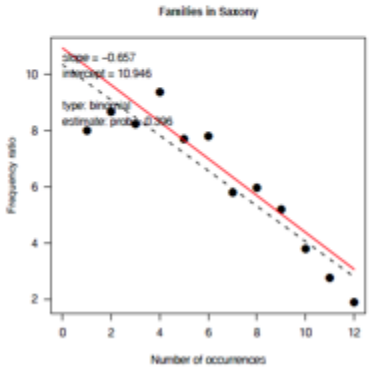


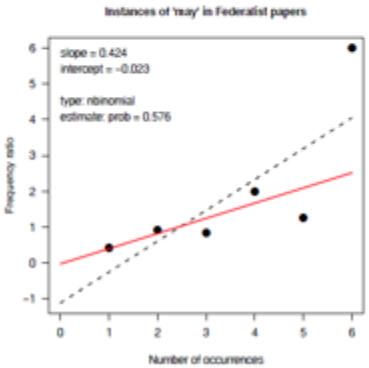


Ord plots: Examples

Ord plots for the Saxony and Federalist data

```
> Ord_plot(Saxony, main = "Families in Saxony", gp=gpar(cex=1), pch=16)
> Ord_plot(Federalist, main = "Instances of 'may' in Federalist papers", gp=gpar(cex=1), pch=16)
```





03: Two-way tables

- Two-way tables summarize frequencies of two categorical factors
  - 2×2: a special case, with odds ratio as a measure
  - $r \times c$: factors can be unordered or ordered
  - $r \times c \times k$: stratified tables, $r \times c$ with groups or circumstances
- Tests & measures of association
  - Pearson χ^2, LR G^2: general association
  - More powerful CMH tests for ordered factors
- Visualization
  - 2×2: fourfold plots
  - r × c: sieve diagrams, tile plots, …
  - More graphical methods to come …

Measures of association

- 2 × 2 tables
  - Odds ratio

$$\theta = \frac{\text{odds}(B_1 \mid A_1)}{\text{odds}(B_1 \mid A_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}}$$

  - Phi coefficient
    - Analog of correlation
    - ϕ^2 = % of variance

$$\phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{1+}n_{2+}n_{+1}n_{+2}} = \pm\sqrt{\chi^2/n}$$

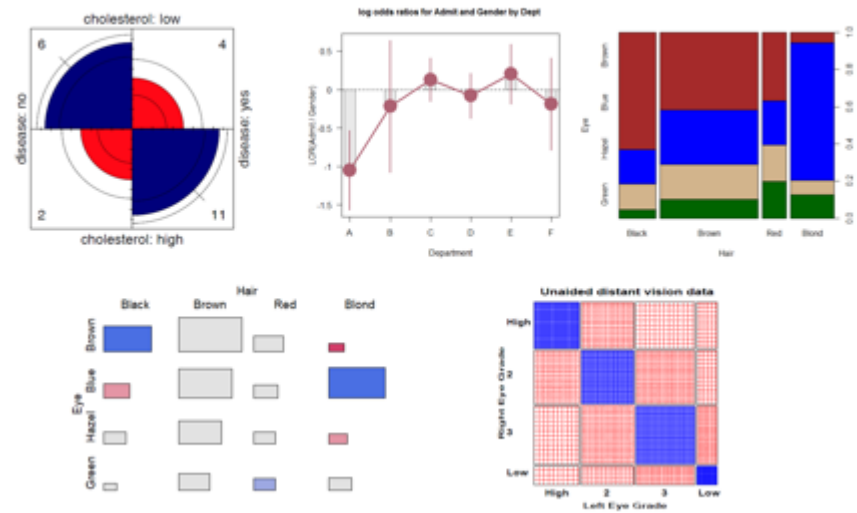
- $r \times c$ tables
  - Cramer's V – generalization of phi

$$\text{Cramer V} = \sqrt{\frac{\chi^2}{n\min(r-1, c-1)}}$$

  - Pearson contingency coef

$$\text{Pearson C} = \sqrt{\frac{\chi^2}{\chi^2/n}}$$

Visualizing association



CMH tests for ordinal factors

Three types of CMH tests:

Non-zero correlation

- Use when *both* row and column variables are ordinal.
- CMH $\chi^2 = (N-1)r^2$, assigning scores (1, 2, 3, ...)
- most powerful for *linear* association

Row/Col Mean Scores Differ

- Use when only *one* variable is ordinal
- Analogous to the Kruskal-Wallis non-parametric test (ANOVA on rank scores)

General Association

- Use when *both* row and column variables are nominal.
- Similar to overall Pearson χ^2 and Likelihood Ratio G^2.

Observer agreement

- **Inter-observer agreement** often used as to assess reliability of a subjective classification or assessment procedure
  - → square table, Rater 1 x Rater 2
  - Levels: diagnostic categories (normal, mildly impaired, severely impaired)
- **Agreement vs. Association:** Ratings can be strongly associated without strong agreement
- **Marginal homogeneity:** Different frequencies of category use by raters affects measures of agreement
- **Measures of Agreement:**
  - Intraclass correlation: ANOVA framework— multiple raters!
  - Cohen's κ: compares the observed agreement, $P_o = \sum p_{ii}$, to agreement expected by chance if the two observer's ratings were independent, $P_c = \sum p_{i+}\, p_{+i}$.

$$\kappa = \frac{P_o - P_c}{1 - P_c}$$

04: Loglinar models, mosaic displays

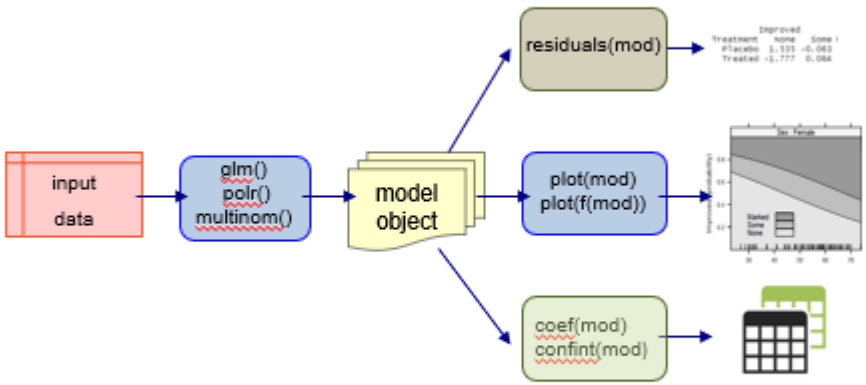
- Mosaic plots use sequential splits to show marginal and conditional frequencies in an *n*-way table
  - Shading: sign and **magnitude** of residuals $\rightarrow$ contributions to χ^2
  - Shows the pattern of association not accounted for
  - Permuting rows/cols often helps
- Loglinear models
  - Express associations with ANOVA-like interaction terms: A*B, A*C
    - Joint independence: [AB][C] ≡ A * B + C
    - Conditional independence: [AC][BC] ≡ $A \perp B \mid C$
  - Fitting models ≅ "cleaning the mosaic"
  - Response models: include all associations among predictors
- Sequential / partial plots & models
  - Sequential: Decompose all associations: V_1; $V_2|V_1$; $V_3|\{V_1, V_2\}$, …
  - Partial: Decompose conditional associations: $[V_1, V_2]|\ V_3 = \{a, b, \ldots\}$

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(Freq), with Poisson dist^n of errors
- **Logit models**: Loglinear, simplified for a binary response

Model-based methods: Fitting & graphing



data | model function | object | method | output

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:

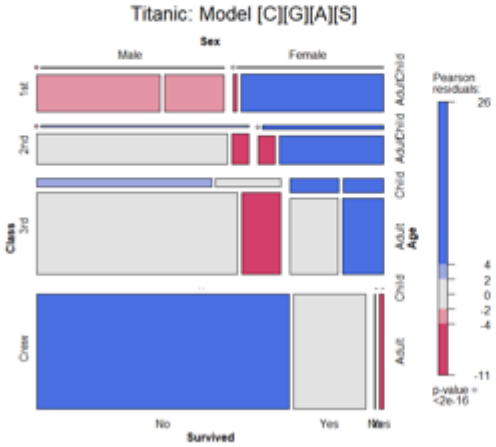
  [A][B] ≡ A ⊥ B ≡ A independent of B
- Special names for various submodels

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

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!!

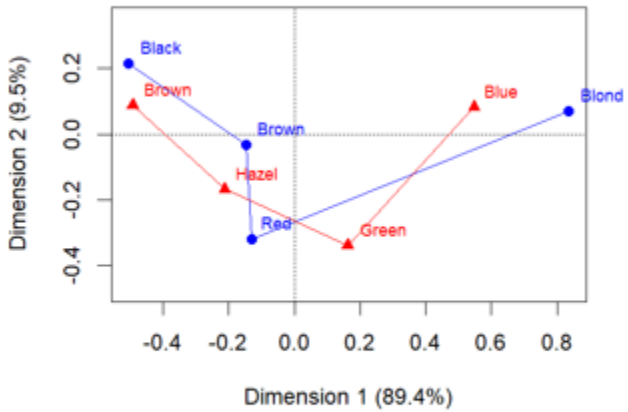
Note the scale of residuals:
+26 -- -11

05: Correspondence analysis

- CA is an exploratory method designed to account for association (Pearson χ^2) in a small number of dimensions
  - Row and column scores provide an optimal scaling of the category levels
  - Plots of these can suggest an explanation for association
- CA uses the singular value decomposition to approximate the matrix of residuals from independence
- Standard and principal coordinates have different geometric properties, but are essentially re-scalings of each other
- Multi-way tables can be handled by:
  - Stacking approach— collapse some dimensions interactively to a 2-way table
  - Each way of stacking $\rightarrow$ a loglinear model
  - MCA analyzes the full $n - way$ table using an indicator matrix or the Burt matrix

Given a new 2-way table, my first thought is nearly always: `plot(ca(mytable))`

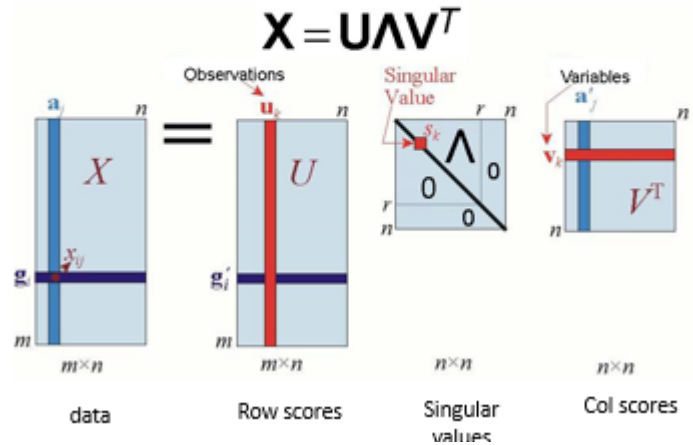
```
plot(haireye.ca, lines=TRUE)
```



- Rough interpretation: row/col points "near" each other are positively associated (independence residuals d_{ij} >> 0)
- Dim 1: 89.4% of χ^2 (dark → light)
- Dim 2: 9.5% of χ^2 (Red/Green vs. others)

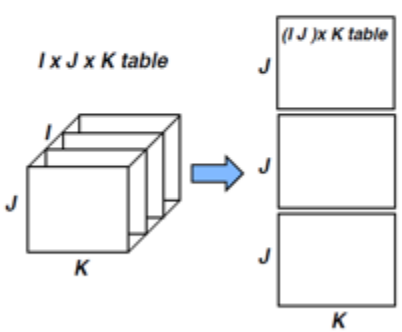
Singular value decomposition

The singular value decomposition (SVD) is a basic technique for factoring a matrix and for matrix approximation
For an m x n matrix **X** of rank $r \le \min(m, n)$ the SVD of **X** is:



Multi-way tables: Stacking

A 3-way table of size $I \times J \times K$ can be sliced and stacked as a two-way table in several ways



- The variables combined are treated "interactively"
- Each way of stacking corresponds to a loglinear model
  - $(I \times J) \times K \rightarrow$ [AB][C]
  - $I \times (J \times K) \rightarrow$ [A][BC]
  - $J \times (I \times K) \rightarrow$ [B][AC]
- Only the associations in separate [ ] terms are analyzed and displayed
- The stacked table is analyzed with ordinary CA of the two-way stacked table

Multiple correspondence analysis

- Extends CA to n-way tables
- Useful when simpler stacking approach doesn't work well, e.g., 10 categorical attitude items
- Analyzes all pairwise bivariate associations. Analogous to:
  - Correlation matrix (numbers)
  - Scatterplot matrix (graphs)
  - All pairwise χ^2 tests (numbers)
  - Mosaic matrix (graphs)
- Provides an optimal scaling of the category scores for each variable
- Can plot all factors in a single plot
- An extension, joint correspondence analysis, gives a better account of inertia for each dimension

06: Logistic regression

- loglm() provides only overall tests of model fit
- Model-based methods, glm(), provide hypothesis tests, CIs & tests for individual terms
- Logistic regression: A glm() for a binary response
  - linear model for the log odds Pr(Y=1)
  - All similar to classical ANOVA, regression models
- Plotting
  - Conditional, full-model plots show data and fits
  - Effect plots show predicted effects averaged over others
- Model diagnostics
  - Influence plots are often informative

Modeling approaches: Overview

Association models

- Loglinear models
  (contingency table form)
  [Admit][Gender Dept]
  [Admit Dept][Gender Dept]
  [AdmitDept][AdmitGender][GenderDept]
- Poisson GLMs
  (Frequency data frame)
  Freq ~ Admit + Gender * Dept
  Freq ~ Admit*Dept + Gender*Dept
  Freq ~ Admit*(Dept + Gender) +
  Gender*Dept
- Ordinal variables
  Freq ~ right + left + Diag(right, left)
  Freq ~ right + left + Symm(right, left)

Response models

- Binary response
- Categorical predictors: logit models
  logit(Admit) ~ 1
  logit(Admit) ~ Dept
  logit(Admit) ~ Dept + Gender
- Continuous/mixed predictors
- Logistic regression models
  Pr(Admit) ~ Dept + Gender + Age + **GRE**
- Polytomous response
- Ordinal: proportional odds model
  Improve ~ Age + Sex + Treatment
- General multinomial model
  WomenWork ~ Kids + HusbandIncome

Linear regression vs Logistic regression

OLS regression:
• Assume y|x ~ N(0, σ^2)

Logistic regression:
• Assume Pr(y=1|x) ~ binomial(p)

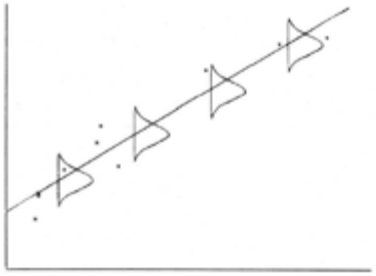


Fig. 2.1. Graphical representation of a simple linear normal regression.

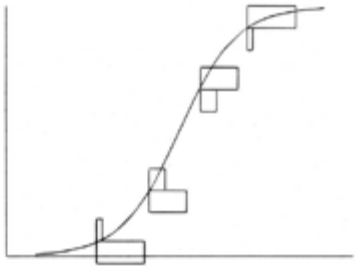


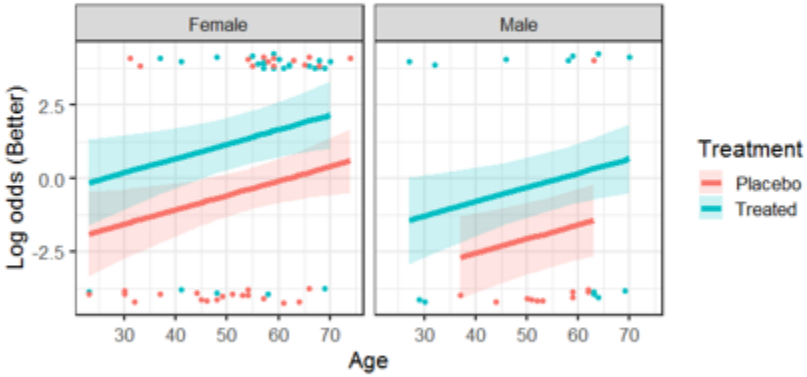
Fig. 2.2. Graphical representation of a simple linear logistic regression.

y linear with x
constant residual variance

y ~ logit (x)
non-constant residual variance ~ p (1-p)

Full-model plot

Plotting on the logit scale shows the additive effects of age, treatment and sex
NB: easier to compare the treatment groups within the same panel

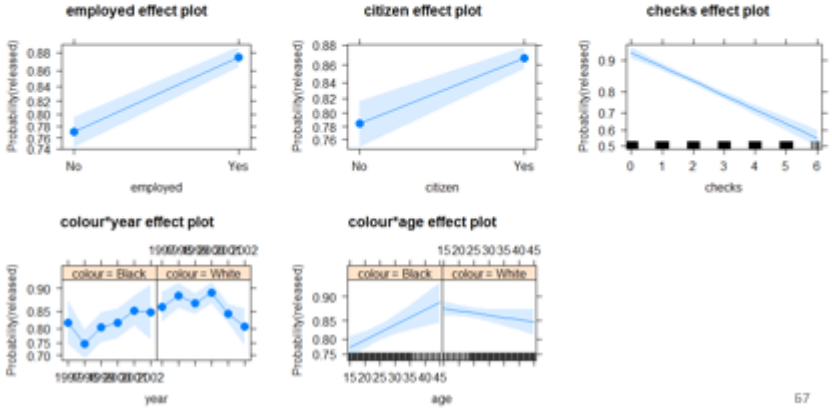


These plots show model uncertainty (confidence bands)
Jittered points show the data

Effect plots: allEffects

All high-order terms can be viewed together using plot(allEffects(mod))

```
arrests.effects <- allEffects(arrests.mod,
xlevels=list(age=seq(15,45,5)))
plot(arrests.effects, ylab="Probability(released)", …)
```

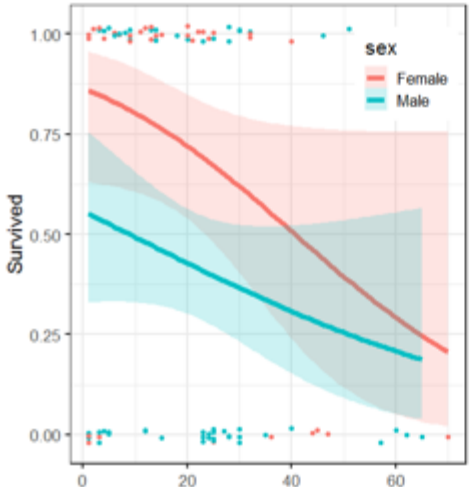


07: Logistic regression: Extensions

- Polytomous responses
  - m response categories $\rightarrow$ (m-1) comparisons (logits)
  - Different models for ordered vs. unordered categories
- Proportional odds model
  - Simplest approach for ordered categories
  - Assumes same slopes for all logits
    - Fit with MASS::polr()
    - Test PO assumption with VGAM::vglm()
- Nested dichotomies
  - Applies to ordered or unordered categories
  - Fit $m - 1$ separate independent models $\rightarrow$ Additive G^2 values
- Multinomial logistic regression
  - Fit $m - 1$ logits as a single model
  - Results usually comparable to nested dichotomies, but diff interpretation
  - R: nnet::multinom()

Exploratory plots

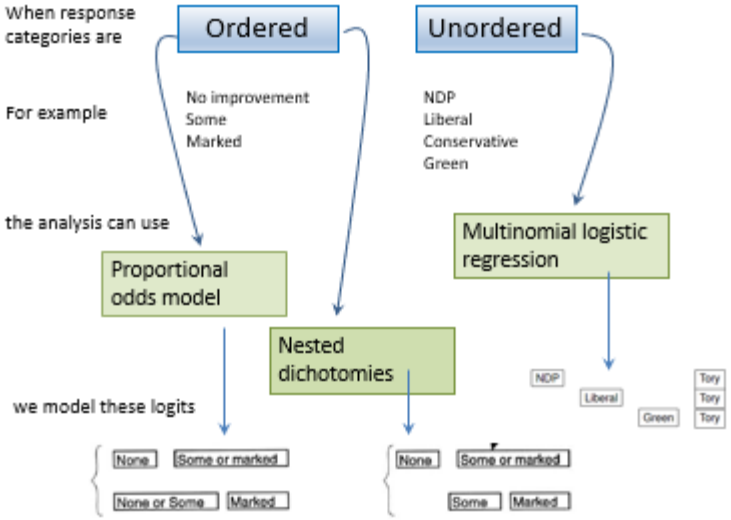
Before fitting models, it is useful to explore the data with conditional ggplots



Survival decreases with age for both men and women

Women more likely to survive, particularly the young

Conf. bands show the data is thin at older ages



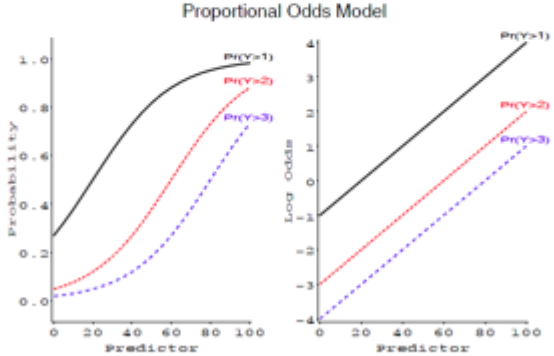
- Consider a logistic regression model for each logit:

$$\text{logit}(\theta_{ij1}) = \alpha_1 + \boldsymbol{x}_{ij}^t \beta_1 \qquad \text{None vs. Some/Marked}$$

$$\text{logit}(\theta_{ij2}) = \alpha_2 + \boldsymbol{x}_{ij}^t \beta_2 \qquad \text{None/Some vs. Marked}$$

- Proportional odds assumption: regression functions are parallel on the logit scale i.e., $\beta_1 = \beta_2$.

Proportional Odds Model



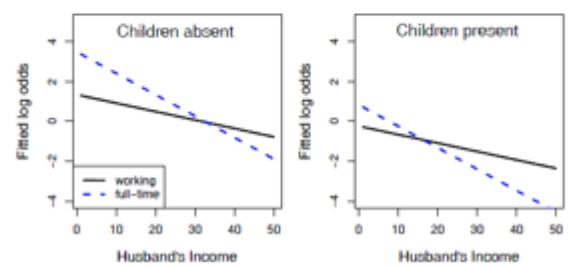
Nested dichotomies: Interpretation

Write out the predictions for the two logits, and compare coefficients:

$$\log\left(\frac{\Pr(\text{working})}{\Pr(\text{not working})}\right) = 1.336 - 0.042\,\text{H\$} - 1.576\,\text{kids}$$

$$\log\left(\frac{\Pr(\text{fulltime})}{\Pr(\text{parttime})}\right) = 3.478 - 0.107\,\text{H\$} - 2.652\,\text{kids}$$

Better yet, plot the predicted log odds for these equations:



08: Extending loglinear models

- Loglinear models, as originally formulated, were quite general, but treated all table variables as unordered factors
  - The GLM perspective is more general, allowing quantitative predictors and handling ordinal factors
  - The logit model give a simplified approach when one variable is a response
- Models for ordered factors give more powerful & focused tests
  - L × L, R, C and R+C models assign scores to the factors
  - RC(1) and RC(2) models estimate the scores from the data
- Models for square tables allow testing structured questions
  - Quasi-independence: ignoring diagonals
  - symmetry & quasi-symmetry
  - theory-specific "topological" models
- These methods can be readily combined to analyze complex tables

Logit models

For a *binary* response, each loglinear model is equivalent to a logit model (logistic regression, with categorical predictors)

- e.g., Admit $\perp$ Gender | Dept (conditional independence $\equiv$ [AD][DG])

$$\log m_{ijk} = \mu + \lambda_i^A + \lambda_j^D + \lambda_k^G + \lambda_{ij}^{AD} + \lambda_{jk}^{DG}$$

So, for admitted ($i = 1$) and rejected ($i = 2$), we have:

$$\log m_{1jk} = \mu + \lambda_1^A + \lambda_j^D + \lambda_k^G + \lambda_{1j}^{AD} + \lambda_{jk}^{DG} \qquad (1)$$

$$\log m_{2jk} = \mu + \lambda_2^A + \lambda_j^D + \lambda_k^G + \lambda_{2j}^{AD} + \lambda_{jk}^{DG} \qquad (2)$$

Thus, subtracting (1)-(2), terms not involving Admit will cancel:

$$\begin{aligned} L_{jk} &= \log m_{1jk} - \log m_{2jk} = \log(m_{1jk}/m_{2jk}) = \text{log odds of admission} \\ &= (\lambda_1^A - \lambda_2^A) + (\lambda_{1j}^{AD} - \lambda_{2j}^{AD}) \\ &= \alpha + \beta_j^{\text{Dept}} \qquad \text{(renaming terms)} \end{aligned}$$

where, α: overall log odds of admission; β_j^{Dept}: effect on admissions of department

Models for ordered categories

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

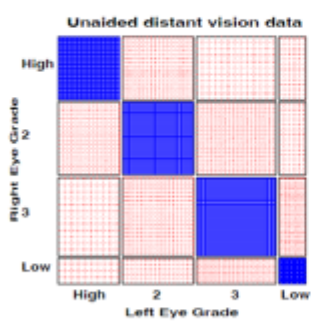
- In many cases, the RC association may be described more simply by assigning numeric scores to the row & column categories.
- For simplicity, we consider only integer scores, 1, 2, ... here
- These models are easily extended to stratified tables

R:C model	μ_{ij}^{RC}	df	Formula
Uniform association	$i \times j \times \gamma$	1	`i:j`
Row effects	$a_i \times j$	$(I-1)$	`R:j`
Col effects	$i \times b_j$	$(J-1)$	`i:C`
Row+Col eff	$ja_i + ib_j$	$I+J-3$	`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`

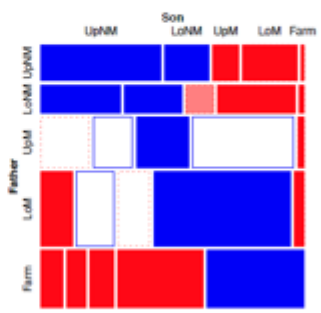
Square tables

Square tables arise when the row and column variables have the same categories, often ordered

Special loglinear models allow us to tease apart different reasons for association



Visual acuity data

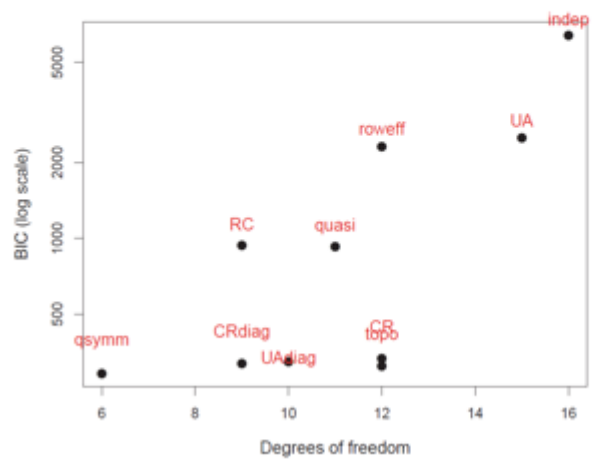


Hauser social mobility data

Model comparison plots

When there are more than a few models, a model comparison plot can show the trade-off between goodness-of-fit and parsimony

- This sorts the models by both fit & complexity



Plot BIC vs. df

Can also use AIC, or G^2 / df in this plot

Plot on log scale to emphasize diffs among better models

And, the winner is: Quasi-symmetry!

09: GLMs for Count Data

- GLMs provide a unified framework for linear models
  - Different families, all estimated in the same way
  - $\rightarrow$link function and associated variance function
- For count data, starting from $\log(\boldsymbol{\mu}) = \mathbf{X}\,\boldsymbol{\beta}$, $\boldsymbol{\mu}|\mathbf{X} \sim$ Poisson:
  - Overdispersion $\rightarrow$ quasi-poisson, negative binomial
  - Standard tools for assessing model fit
- Excess zero counts introduce new ideas & methods
  - ZIP model: structural model for the 0s
  - Hurdle model: random model for 0s, 2nd model for Y>0
- In all this, we rely on data & model plots for understanding

Canonical links and variance functions

- For every distribution family, there is a default, canonical link function
- Each one also specifies the expected relation between the mean and variance

Table 11.2: Common distributions in the exponential family used with generalized linear models and their canonical link and variance functions

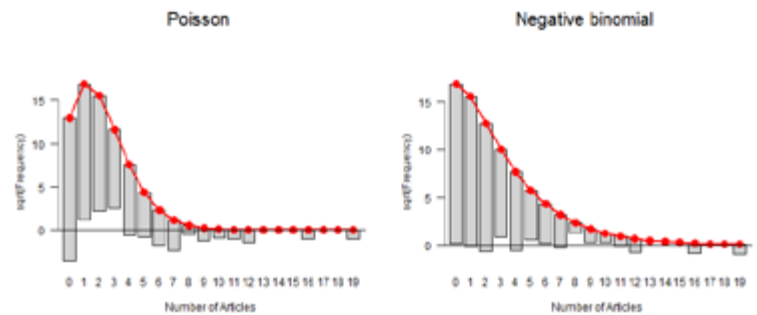
Family	Notation	Canonical link	Range of y	Variance function, $\mathcal{V}(\mu \mid \eta)$
Gaussian	$N(\mu, \sigma^2)$	identity: μ	$(-\infty, +\infty)$	ϕ
Poisson	$\text{Pois}(\mu)$	$\log_e(\mu)$	$0, 1, \ldots, \infty$	μ
Negative-Binomial	$\text{NBin}(\mu, \theta)$	$\log_e(\mu)$	$0, 1, \ldots, \infty$	$\mu + \mu^2/\theta$
Binomial	$\text{Bin}(n, \mu)/n$	$\text{logit}(\mu)$	$\{0, 1, \ldots, n\}/n$	$\mu(1-\mu)/n$
Gamma	$G(\mu, \nu)$	μ^{-1}	$(0, +\infty)$	$\phi\mu^2$
Inverse-Gaussian	$IG(\mu, \nu)$	μ^2	$(0, +\infty)$	$\phi\mu^3$

Choose a basic family:

- Get a default, canonical link, g(μ)
- Also get a variance function for free!

First, look at rootograms:

```
plot(goodfit(PhdPubs$articles), xlab = "Number of Articles",
     main = "Poisson")
plot(goodfit(PhdPubs$articles, type = "nbinomial"),
     xlab = "Number of Articles", main = "Negative binomial")
```



One reason the Poisson doesn't fit: excess 0s (some never published?)

Q: What might some other reasons be?
Think back to assumptions: independent obs; constant probs; unmodelled vars

Quasi-poisson models

- The quasi-poisson model allows the dispersion, ϕ, to be a free parameter, estimates with other coefficients
- The conditional variance is allowed to be a multiple of the mean

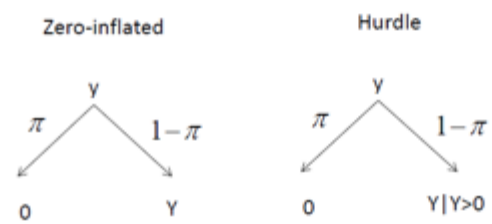
$$\text{Var}(y_i \mid \eta_i) = \phi\, \mu_i$$

- This model is fit with glm() using family=quasipoisson
  - The estimated coefficients $\widehat{\beta}$ are unchanged
  - The standard errors are multiplied by $\phi^{1/2}$
  - Peace, order & good government is restored!

Models for excess zeros

Two types of models, with different mechanisms for zero counts

- ***zero-inflated models***: The responses with $y_i = 0$ arise from a mixture of structural, always 0 values, with $\Pr(y_i = 0) = \pi_i$ and the rest, which are random 0s, with $\Pr(y_i = 0) = 1 - \pi_i$
- ***hurdle models***: One process determines whether $y_i = 0$ with $\Pr(y_i = 0) = \pi_i$. A second process determines the distribution of values of positive counts, $\Pr(y_i \mid y_i > 0)$



10: Models for log odds & LORs

- Logit models for a binary response generalize readily to a polytomous response
  - $\rightarrow$Models for log odds, familiar interpretation
  - Handles 3+ way table, ordinal variables
  - Simple plots for interpretation
- Generalized odds ratios handle bivariate responses
  - Simple linear models for LOR
  - Easy to model log odds for each response and the LOR simultaneously
  - Easy to visualize results

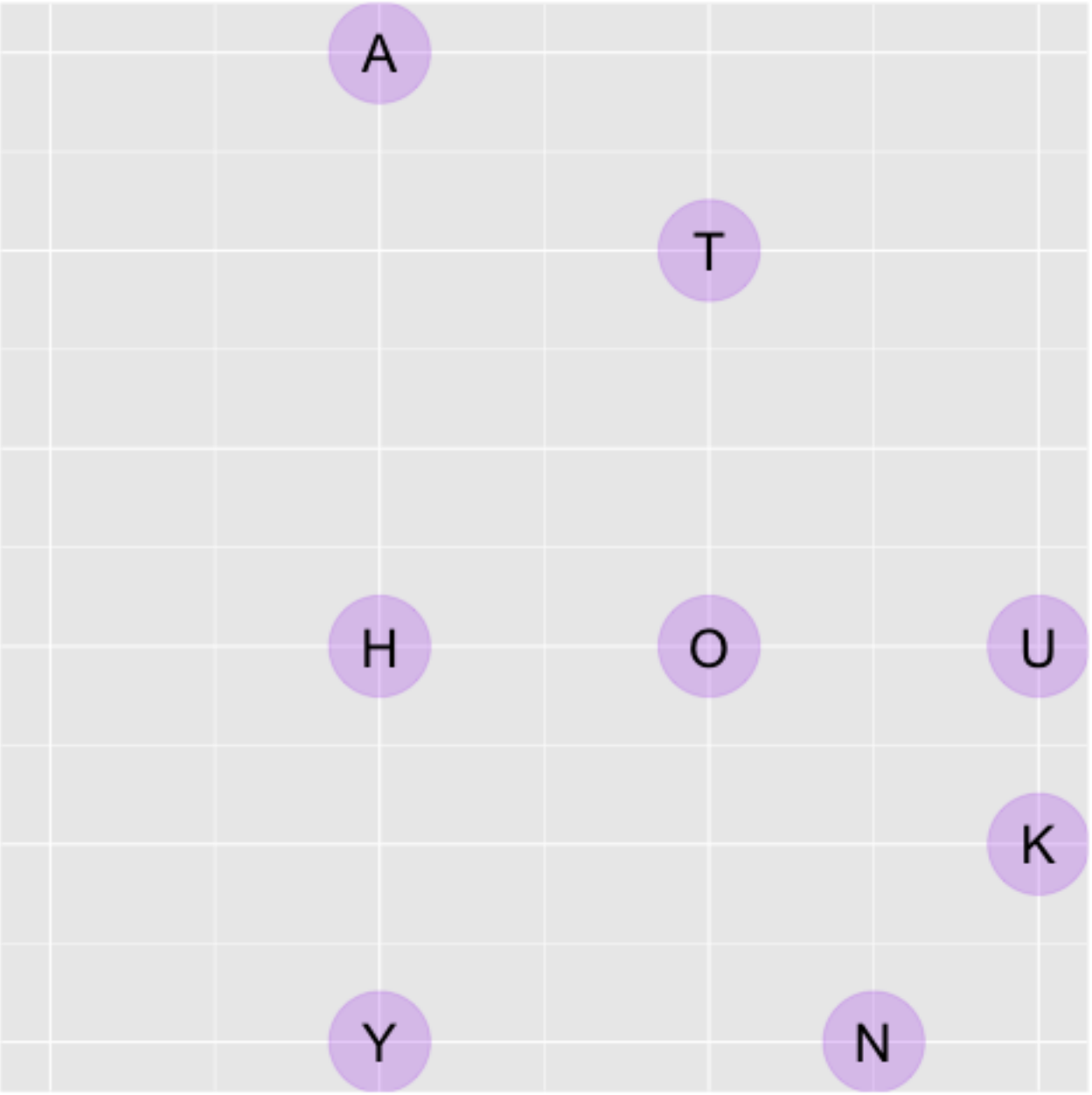
Your turn: Feedback?

What did you like/dislike about 6136?

- Topics: what were the:
  - most interesting?
  - most boring?
  - Most challenging?
- What did you learn most from?
- What gave you the most difficulty?
- How does this relate to your own work?

Tips for next time …

- What should I try to differently the next time?
  - More of X?
  - Less of Y?
  - Aspects of how the course is structured?
  - Evaluation?


A
T
H
O
U
K
Y
N