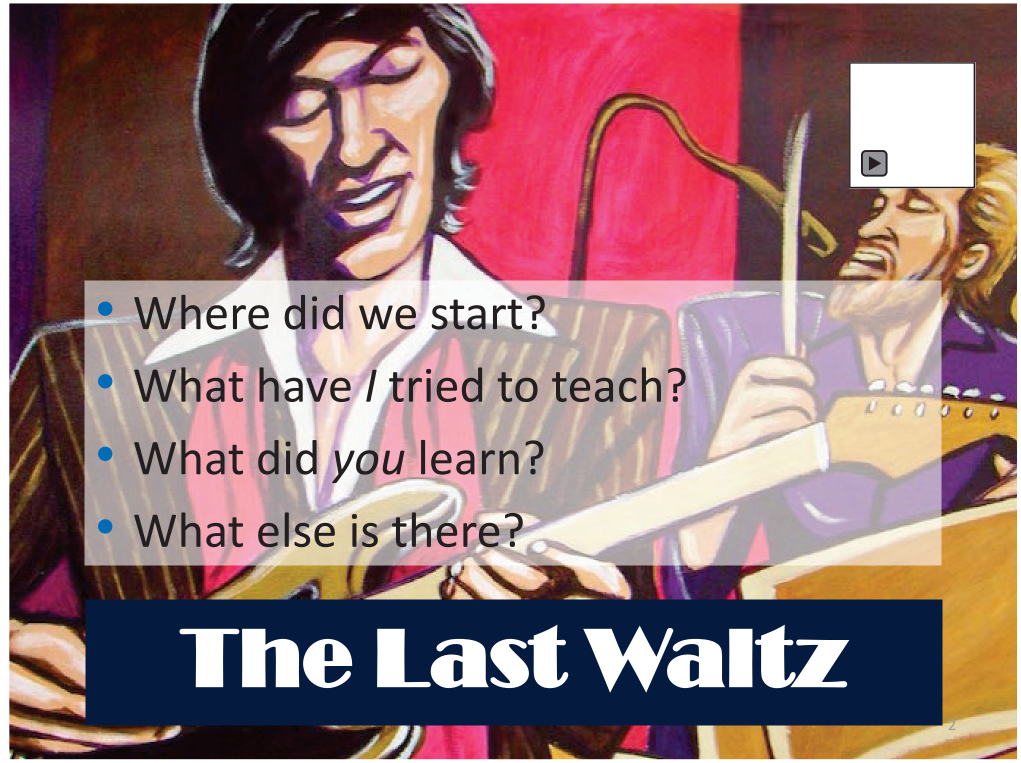
The Last Waltz

Michael Friendly

Psych 6136

https://friendly.github.io/6136



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

The Last Waltz

My goals

- Start with descriptive, hypothesis testing methods, then progress to model-based methods
- Visual tools for thinking & understanding
  - Sieve plots, mosaic plots, spineplots, …
  - Correspondence analysis: best 2D summary
  - Effect plots, Data + Model plots
- Build from simple, loglinear models to more complex ones

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
- Matrices: `matrix()`, with row & col names
- Arrays: `array()`, with `dimnames()`
- Data frames
  - Case form (individual observations)
  - Frequency form

Gender compared to handedness

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

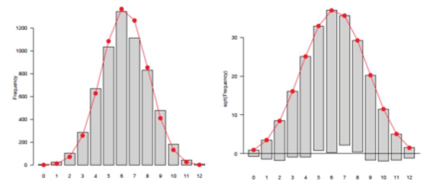
margins

	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

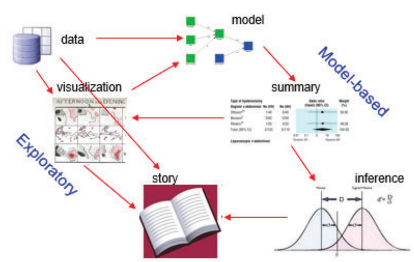
Model:	*Independence*: [Hair][Eye] χ^2 (9)= 138.29						
Color coding:	<-4	<-2	<-1	0	>1	>2	>4
n in each cell:	$n <$ expected				$n >$ expected		

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

43

Data, pictures, models & stories

Now, tell the story!



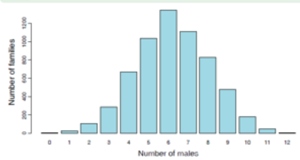
Examples: binomial

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, ...	$\frac{p^k}{n\log(1-p)}$		

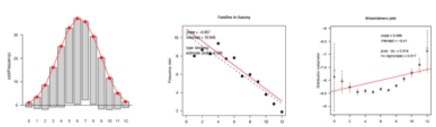
11

Graphing discrete distributions

Rootograms

Ord plots

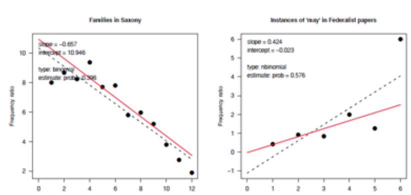
Robust distribution plots



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)
```



40

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 → logistic regression
  - Poisson → poisson regression; logliner models
  - These are special cases of generalized linear models

6

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 …

8

Measures of association

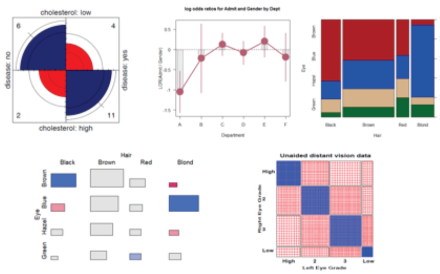
- 2×2 tables
  - Odds ratio $\theta = \dfrac{\text{odds}(B_1 \mid A_1)}{\text{odds}(B_1 \mid A_2)} = \dfrac{n_{11}/n_{12}}{n_{21}/n_{22}}$
  - Phi coefficient $\phi = \dfrac{n_{11}n_{22} - n_{12}n_{21}}{n_{1+}n_{2+}n_{+1}n_{+2}} = \pm\sqrt{\chi^2/n}$
    - Analog of correlation
    - ϕ^2 = % of variance
- $r \times c$ 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: Loglinear models, mosaic displays

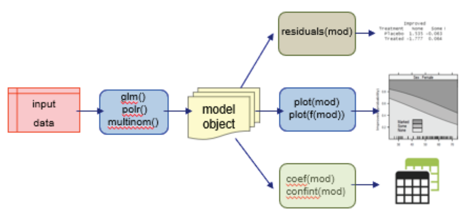
- Mosaic plots use sequential splits to show marginal and conditional frequencies in an *n*-way table
  - Shading: sign and **magnitude** of residuals → contributions to χ^2
  - Shows the pattern of association not accounted for
  - Permuting rows/cols often helps
- Loglinear models
  - Express associations with ANOVA-like interaction terms: A*B, A*C
    - Joint independence: [AB][C] ≡ A * B + C
    - Conditional independence: [AC][BC] ≡ A ⊥ B | 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 distn 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

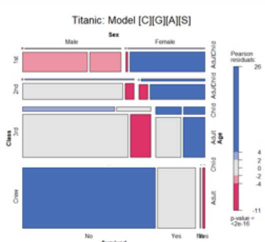
Table: Log-linear Models for Three-Way Tables

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 Class, Gender, Age and Survived



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

Let's clean this mosaic!!

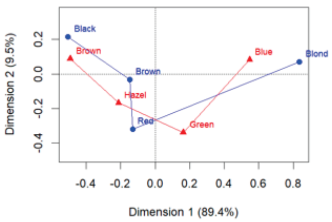
Note the scale of residuals: +26 -- -11

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 → 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)
```



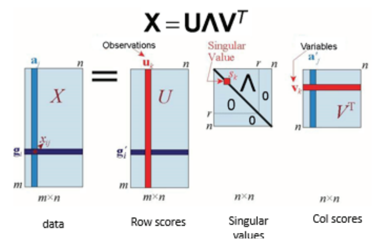
- Rough interpretation: row/col points "near" each other are positively associated (independence residuals $d_{ij} >> 0$)
- Dim 1: 89.4% of χ^2 (dark → light)
- Dim 2: 9.5% of χ^2 (Red/Green vs. others)

6

Singular value decomposition

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

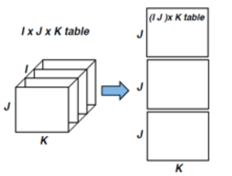
$$\mathbf{X} = \mathbf{U} \Lambda \mathbf{V}^T$$



15

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

38

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

49

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

14

Modeling approaches: Overview

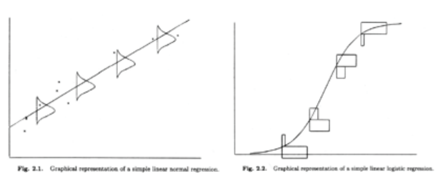
Association models	Response models
• Loglinear models (contingency table form) [Admit][Gender Dept] [Admit Dept][Gender Dept] [AdmitDept][AdmitGender][GenderDept]	• Binary response • Categorical predictors: logit models logit(Admit) ~ 1 logit(Admit) ~ Dept logit(Admit) ~ Dept + Gender
• Poisson GLMs (Frequency data frame) Freq ~ Admit + Gender * Dept Freq ~ Admit*Dept + Gender*Dept Freq ~ Admit*(Dept + Gender) + Gender*Dept	• Continuous/mixed predictors • Logistic regression models Pr(Admit) ~ Dept + Gender + Age + **GRE**
• Ordinal variables Freq ~ right + left + Diag(right, left) Freq ~ right + left + Symm(right, left)	• Polytomous response • Ordinal: proportional odds model Improve ~ Age + Sex + Treatment • General multinomial model WomenWork ~ Kids + HusbandIncome

4

Linear regression vs Logistic regression

OLS regression:
• Assume y|x ~ N(0, σ²)

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



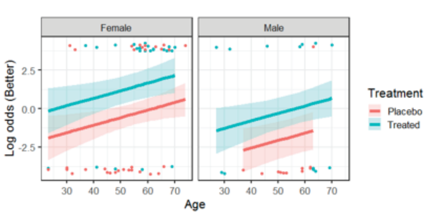
y linear with x
constant residual variance

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

25

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



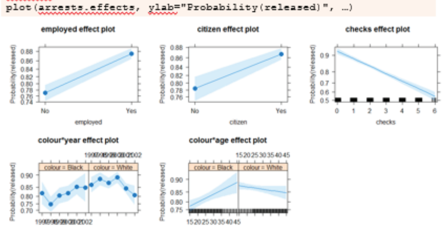
These plots show model uncertainty (confidence bands)
Jittered points show the data

49

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)", …)
```

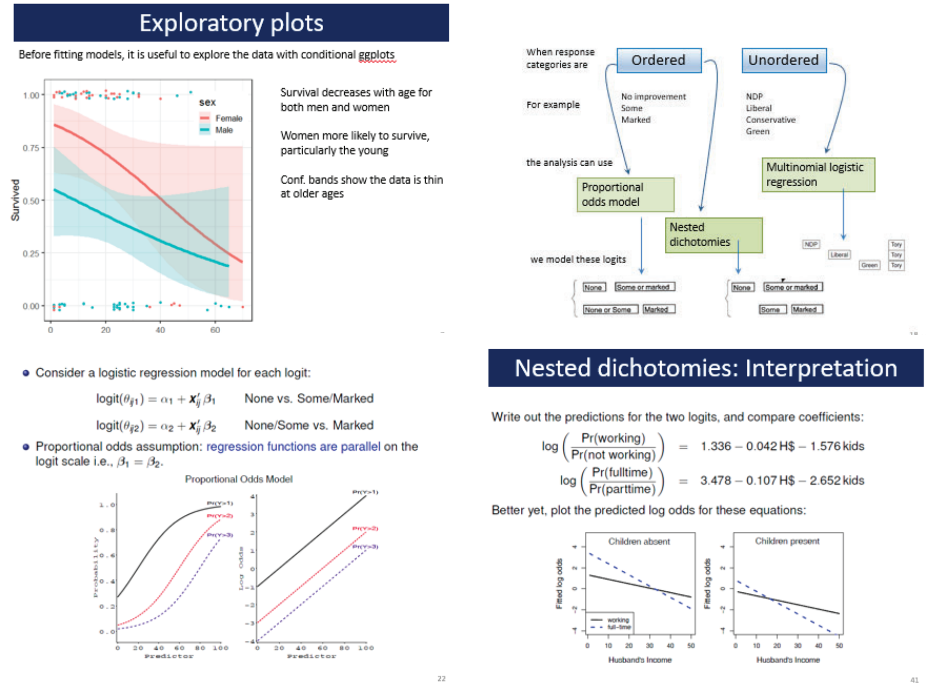


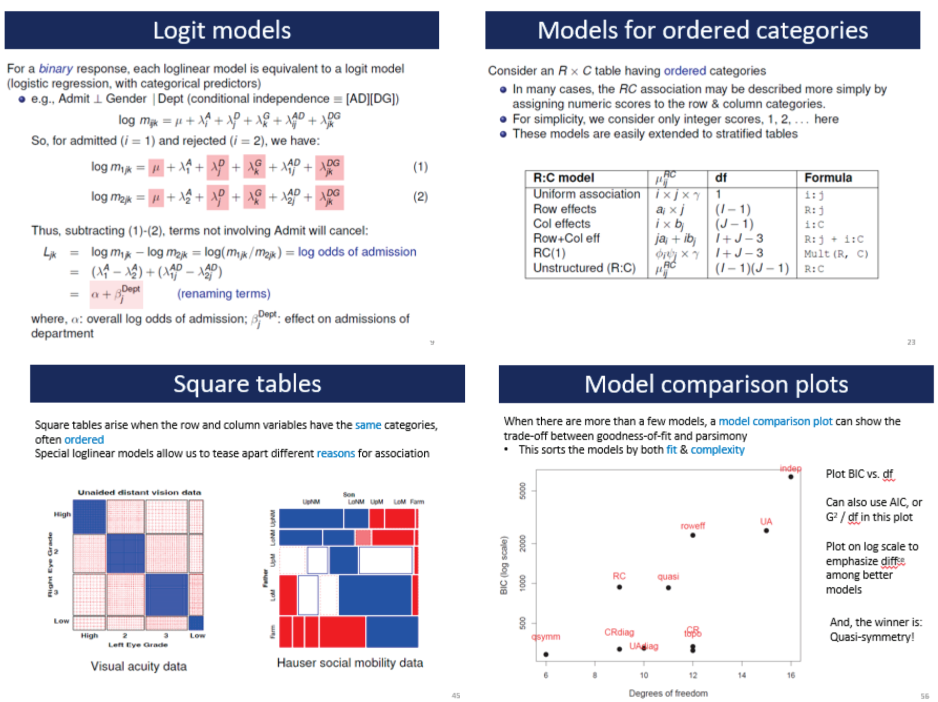
67

07: Logistic regression: Extensions

- Polytomous responses
  - m response categories → (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 → 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()

16





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

09: GLMs for Count Data

- GLMs provide a unified framework for linear models
  - Different families, all estimated in the same way
  - →link function and associated variance function
- For count data, starting from log(μ) = **X** β, μ|**X** ~ Poisson:
  - Overdispersion → 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

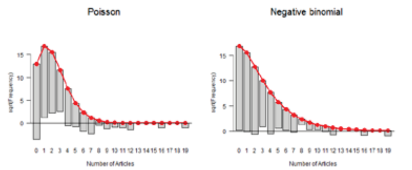
Family	Notation	Canonical link	Range of y	Variance function, $\mathcal{V}(\mu \mid \eta)$
Gaussian	$N(\mu, \sigma^2)$	identity: μ	$(-\infty, +\infty)$	ϕ
Poisson	$\mathrm{Pois}(\mu)$	$\log_e(\mu)$	$0, 1, \ldots, \infty$	μ
Negative-Binomial	$\mathrm{NBin}(\mu, \theta)$	$\log_e(\mu)$	$0, 1, \ldots, \infty$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

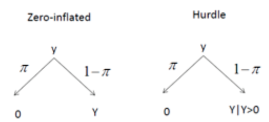
$$\mathrm{Var}(y_i \mid \eta_i) = \phi\, \mu_i$$

- This model is fit with glm() using family=quasipoisson
  - The estimated coefficients $\hat{\beta}$ 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
  - →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

Main ideas

- Familiar case— Binary responses:
  - Every loglinear model for a binary response has an equivalent form in terms of log odds ["logit" models]
  - Log odds models have simple interpretations
  - Data + model plots give simple descriptions of data and models
- Extend to two-way ($I \times J$) and three-way + ($I \times J \times K_1 \ldots$) tables:
  - Log odds as contrasts in log($\boldsymbol{n}$)
  - Variety of simple models for log odds (ANOVA-like)
  - Easily incorporate ordinal variables
  - Data + model plots give simple descriptions of data and models
- Generalized log odds ratios capture associations between two focal variables
  - Simple linear models for LOR
  - Direct visualization (Data + model plots) $\Longrightarrow$ more sensitive comparisons

Based on my CARME (2015) presentation, https://www.datavis.ca/papers/CARME2015-2x2.pdf

Models for log odds

A variety of simple models can be specified in terms of log odds:

Table: Models for adjacent log odds in an $I \times J$ table with B as the response

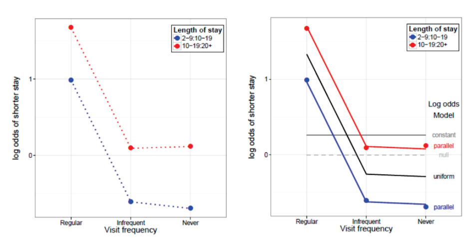
Model	log odds parameters	degrees of freedom
null log odds	$\psi_{ij}^{AB} = 0$	$I(J-1)$
constant log odds	$\psi_{ij}^{AB} = \psi$	$I(J-1) - 1$
uniform B log odds	$\psi_{ij}^{AB} = \psi_i^A$	$I(J-2)$
parallel log odds	$\psi_{ij}^{AB} = \psi_i^A + \psi_j^B$	$(I-1)(J-2)$
saturated	ψ_{ij}^{AB} unspecified	

- The log odds, ψ_{ij}^{AB} can be viewed as entries in an $I \times (J-1)$ table
- These models are analogous to ANOVA tests of the A, B and $A * B$ effects in this table.

These are similar to some glm() models we looked at before, but here in a simpler way by thinking in terms of log odds

Visualizing log odds and models

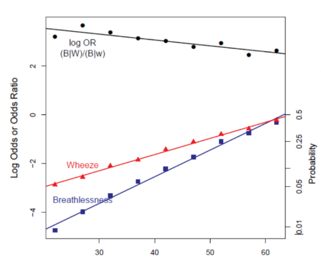
Plots of observed and fitted log odds: easy interpretation of data and models



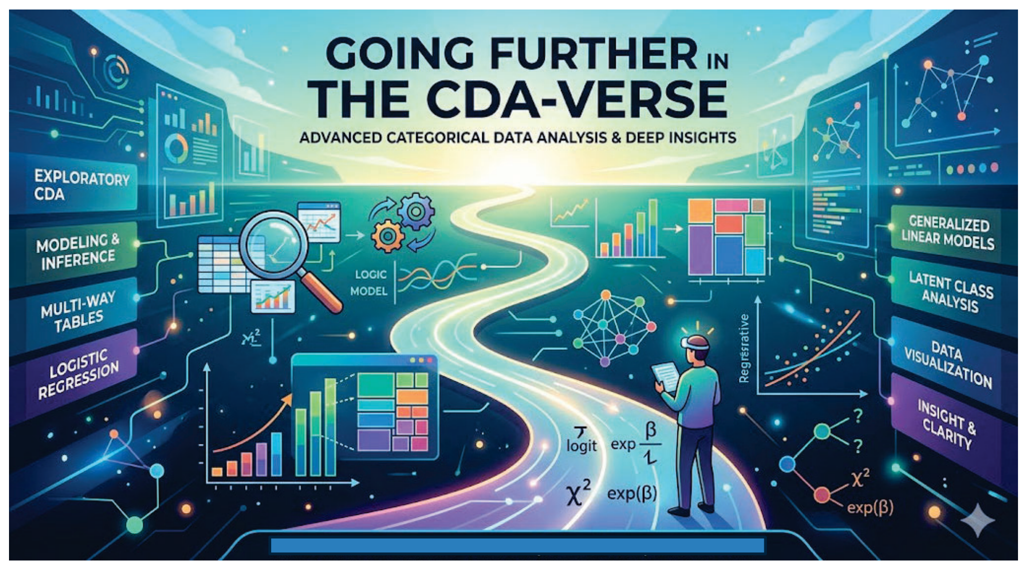
Data plot: Observed log odds

Data + Model plot (fitted log odds)

Linear model for log odds and log odds ratios



Log odds & LORs have similar scales, so it is not terrible to plot them together



Going further: What did I leave out?

- This course, like DDAR, takes an entirely frequentist approach
  - Treats model parameters as fixed, constants
  - *p*-values, CIs interpreted as “over many repeated samples …” --- 95% would contain the true value
  - Never really know if your sample gives correct inferences
- Bayesian approach
  - Treat parameters as random variables, incorporating prior beliefs and updating them with observed data to produce a posterior distribution. Much more intuitive.
  - A Bayesian 95% credible interval **directly** states there is a 95% probability the parameter lies within that range

Fisher vs. Bayes

- Small samples, sparse data
  - Frequentist methods may fail to converge, or require fixups (add 0.5 to avoid 0 cells) for χ^2 tests to be valid
  - Bayesian approaches use priors to "smooth" empty cells, $\Longrightarrow$ more robust estimates in high-dimensional tables
- Logistic regression
  - Frequentist: relies on large-sample asymptotics for inference
  - Bayesian: provides full posterior dist^n for odds ratios, βs
- Computation
  - Frequentist: very fast, particularly for large datasets
  - Bayesian:
    - Requires Bayesian sampling, e.g., MCMC; did my chains converge??
    - Originally (in DDAR): software quite limited, awkward to use
    - Now, much easier to use.

Bayesian advances

- Regularization methods: stabilize inference with small or zero counts
- New variable selection methods
  - Effect clustering: merge levels of categorical predictor that have the same effect $\Longrightarrow$ more stable in high-D problems
- Applications:
  - Clinical trials: Bayesian adaptive designs can use historical data to speed up decision on efficacy of drug or treatment
  - Epidemiology: Bayesian hierarchical models used in disease mapping (estimating risk across areas), pooling strength

JAMA

Published online March 23, 2026

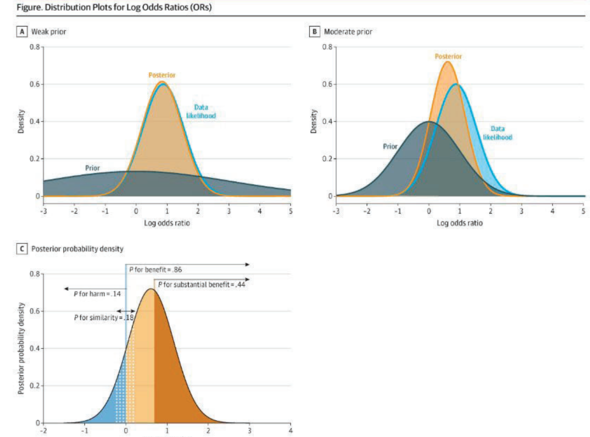
BREAKING NEWS: Bayesian methods adopted in JAMA recommendations to FDA
https://t.co/7mMecV2EI1

PERSPECTIVE

Embracing Bayesian Methods in Clinical Trials
FDA's Long-Awaited Draft Guidance

J. Jack Lee, PhD; Frank E. Harrell Jr, PhD; Lisa M. LaVange, PhD; David J. Spiegelhalter, PhD

Figure. Distribution Plots for Log Odds Ratios (ORs)



Hierarchical, longitudinal data

- All methods in this course — χ^2 tests, loglinear models, logistic regression, ...— assume observations are **independent**
- Fails when:
  - **same person** is measured at multiple time points (longitudinal / repeated measures), or
  - people are **nested within groups** such as schools, clinics, or geographic regions (hierarchical).
- Ignoring dependence:
  - Underestimated std. errors: CIs too narrow, *p*-values too small (👻 phantom precision)
  - Biased parameter estimates

Longitudinal CDA designs

Data structure: n subjects, each measured at T occasions.

- Categorical response: — e.g., disease present/absent, symptom severity (ordered), count of events in a period.
- Examples:
  - Panel surveys: political preference over election cycles
  - Clinical: patients worse, stable, improved at 4, 8, 12 weeks
  - Epidemiology: hospital visits per patient over 3 years

Hierarchical (clustered / multilevel) designs

Data structure: Individuals at level 1 are nested within groups at level 2 (maybe: level 3, ...).

- The grouping creates dependence: two students in same school, or two patients at same clinic, share unmeasured contextual influences.
- Examples:
  - Students (level 1) nested in classrooms (level 2) nested in schools (level 3); binary outcome = passed a standardized test.
  - Patients nested in hospitals; ordinal outcome = satisfaction rating.
- Why standard methods fail:
  - Observations within the same cluster are positively correlated (intraclass correlation, ICC)
  - Std. logistic / Poisson regression treats all as independent, inflating the effective sample size, giving over-confident inference
  - Group-level predictors tested against wrong error term.

Modeling approaches: GEE

- Marginal models via Generalized Estimating Equations (GEE)
  - Fit a population-average GLM (logistic, Poisson, ordinal, ...) as usual, but explicitly model the within-S correlation structure (exchangeable, AR(1), unstructured, ...).
  - GEE adjusts standard errors so inference is valid even if the correlation structure is mis-specified (robust/“sandwich” SEs).
  - Interpretation: coefficients describe the **average effect in the population**, not effects for a specific individual.
- R: `geepack::geeglm()`, `gee::gee()`

General Linear Mixed Models (GLMM)

The natural extension of mixed models to non-normal outcomes. For a binary response with two-level nesting:

$$\text{logit}\left[P\left(Y_{ij}=1\right)\right]=\mathbf{x}_{ij}^{T}\boldsymbol{\beta}+u_j$$

where $u_j \sim N(0, \sigma_u^2)$ is a random intercept for group j.

- $\boldsymbol{\beta}$: fixed effects (individual- and group-level predictors).
- σ_u^2: variance of group-level random effects
- Can add random slopes if effect of predictor varies across groups
- R: `lme4::glmer()`, `glmmTMB::glmmTMB()`, `brms::brm()` (Bayesian)

Transition (Markov) models

- Model response at time t as a function of covariates *and* the lagged response $Y_{i,t-1}$ (or several lags).
- Directly captures how prior state predicts current state
- Natural for panel /clinical data where change over time is the scientific focus.
- Equivalent to a conditional logistic regression on each transition.

Connections to course topics

Course topic	Longitudinal/hierarchical extension
Logistic regression (binary)	GLMM with logit link; GEE with binary family
Ordinal logistic (proportional odds)	Mixed proportional-odds model (`ordinal::clmm()`)
Poisson / negative binomial (counts)	Poisson GLMM; NB GLMM (`glmmTMB`)
Multinomial logistic	Mixed multinomial (`mclogit::mblogit()`)
Loglinear models	Mixed log-linear for repeated cross-classifications

R software summary

Package	Primary use
`lme4`	GLMMs: `glmer()` for binary/count; fast, widely used
`glmmTMB`	GLMMs with more families (NB, zero-inflated, ordinal), random slopes
`geepack`	GEE: `geeglm()`; marginal models, robust SEs
`ordinal`	Mixed ordinal models: `clmm()`
`brms`	Bayesian GLMMs via Stan; full posterior for all variance components

WHAT! There's even more??

Five other areas in a totally comprehensive CDA course:

1. Latent Class Analysis (LCA)
2. Survival Analysis / Event History Models
3. Item Response Theory (IRT)
4. Missing Data & Multiple Imputation
5. Regularized Logistic Regression

Latent Class Analysis & Survival Analysis

Latent Class Analysis (LCA)

- Factor analysis for categorical indicators: finds latent subgroups explaining associations among observed variables
- Widely used for diagnostic categories, behavioral typologies, health risk profiles
- R: `poLCA`, `mclust`
- Agresti Ch. 13; Collins & Lanza (2010), *Latent Class and Latent Transition Analysis*

Survival Analysis / Event History Models

- Outcome is *time to event* (relapse, death, dropout, recovery) — possibly censored
- Workhorses: discrete-time logistic regression; Cox proportional hazards
- Very common in clinical and health psychology
- R: `survival`, `flexsurv`; Agresti Ch. 12; Hosmer, Lemeshow & May (2008)

Item Response Theory & Missing Data

Item Response Theory (IRT)

- Models P(correct or endorsed response) as a function of a latent trait (ability, attitude)
- Rasch, 2PL, 3PL models are logistic regression with a latent predictor
- Natural extension of GLMs — directly relevant to psychometrics
- R: `ltm`, `mirt`, `TAM`; Embretson & Reise (2000)

Missing Data / Multiple Imputation

- Cannot mean-impute a categorical variable; complete-case analysis distorts association estimates
- MICE (multiple imputation via chained equations) handles categorical predictors and outcomes directly
- R: `mice`; van Buuren (2018), *Flexible Imputation of Missing Data* (free online)

Regularized Logistic Regression

The problem: when ordinary logistic regression fails

- Many predictors → instability, perfect separation, near-collinearity
- LASSO (L_1) shrinks coefficients to zero → automatic variable selection
- Ridge (L_2) shrinks toward zero → stable estimation, all predictors retained
- Elastic net combines both penalties

Tools and further reading

- Frequentist complement to Bayesian regularization (shrinkage priors)
- R: `glmnet`
- Hastie, Tibshirani & Friedman (2009), *The Elements of Statistical Learning*, Ch. 4, 18

CDA Dad Jokes

- **Q**: What did the χ^2 statistic say to her *p*-value?
  - A: You're the only one who can tell me if our relationship is real
- **Q**: Why was the mosaic plot so calm?
  - A: He had already made peace with all his residuals.
- **Q**: Why did logistic regression break up with his linear regression?
  - A: It's not you honey, it's me! My love is binary — I can only give you 0 or 1.
- A loglinear model walks into a bar.
  - Bartender says: "We don't serve your type here.“
  - Model says: “Fine — I'll just fit a main effects model and assume no one here is associated."

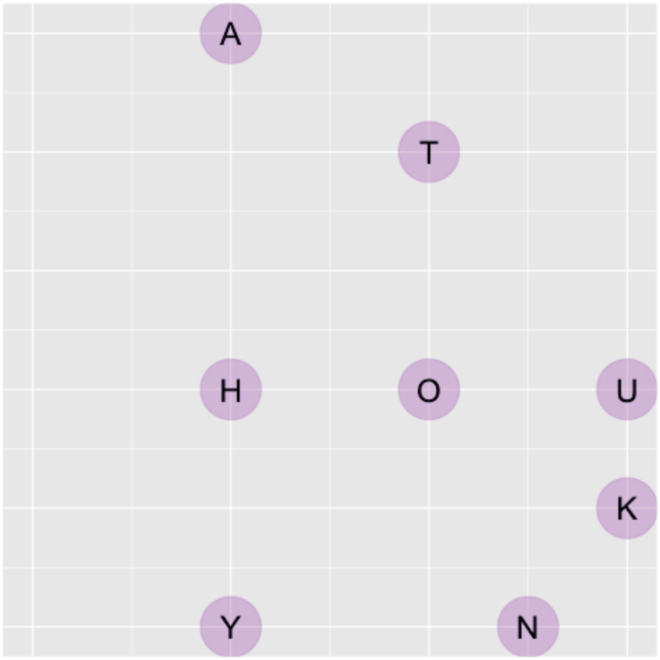
Your turn: Feedback?

What did you like/dislike about 6136?

- Topics: what were the:
  - most interesting?
  - most boring?
  - most challenging/difficult?
- How useful were the web materials for the course?
  - Quizes? R scripts? Resources / extra readings?
- What topics/activities 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