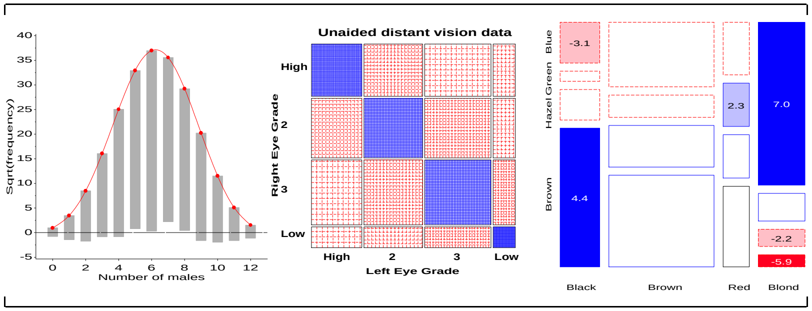
Categorical Data Analysis with Graphics



Michael Friendly

York University, Toronto, `<friendly@yorku.ca>`

SUGI 28

April, 2003

Outline

- Overview: Categorical Data *and* Graphics
- Methods for discrete distributions
  - Hanging rootograms
  - Robust distribution plots
- Methods for two-way frequency tables
  - Fourfold displays
  - Sieve diagrams
- Mosaic displays and loglinear models for *n*-way tables
  - Mosaic displays
  - Mosaic matrices
- Logistic and logit regression
  - Logit plots, effect plots
  - Diagnostic plots

Color version of these slides:
`http://www.math.yorku.ca/SCS/sugi/sugi28.pdf`

Categorical Data Analysis

Methods of analysis for categorical data fall into two main categories:

- **Non-parametric, randomization-based methods**
  - make minimal assumptions
  - useful for hypothesis-testing
  - SAS: `PROC FREQ`
    - Pearson Chi-square
    - Fisher's exact test (for small expected frequencies)
    - Mantel-Haenszel tests (ordered categories: test for *linear* association)
- **Model-based methods**
  - Must assume random sample (possibly stratified)
  - Useful for estimation purposes
  - Greater flexibility; fitting specialized models (e.g., symmetry)
  - More suitable for multi-way tables
  - SAS: `PROC LOGISTIC`, `CATMOD`, `GENMOD` , `INSIGHT` (Fit YX)
    - estimate standard errors, covariances for model parameters
    - confidence intervals for parameters, predicted $\Pr\{\text{response}\}$

Graphical Methods for Categorical Data

If I can't picture it, I can't understand it. Albert Einstein

Getting information from a table is like extracting sunlight from a cucumber.

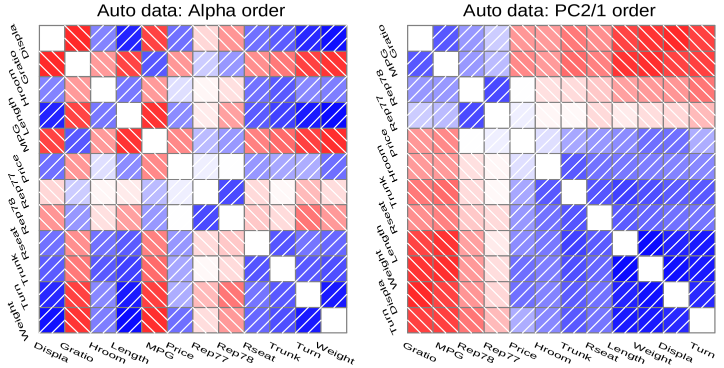
Farquhar & Farquhar, 1891

- **Tables vs. Graphs**
  - Tables are best suited for *look-up*— read off exact numbers
  - Graphs are better for showing *patterns, trends, anomalies*, making *comparisons*
  - Visual presentation as *communication*: what do you want to say?

Principles of Graphical Displays

- **Effect ordering** (Friendly and Kwan, 2002)— In tables and graphs, sort unordered factors according to the effects you want to see/show.

Auto data: Alpha order   Auto data: PC2/1 order



"Corrgrams: Exploratory displays for correlation matrices" (Friendly, 2002)

- Effect ordering and high-lighting for tables (Friendly, 2000)

Table 1: Hair color - Eye color data: Effect ordered

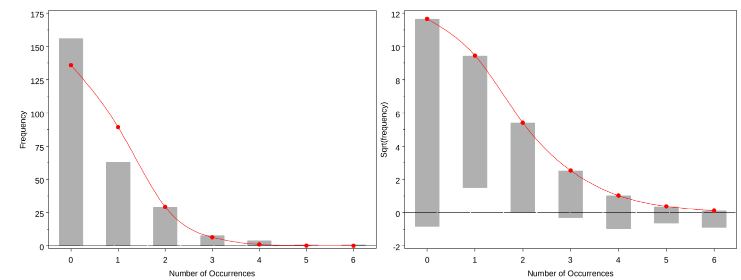
Eye color	Hair color: Black	Brown	Red	Blond
Brown	68	119	26	7
Hazel	15	54	14	10
Green	5	29	14	16
Blue	20	84	17	94

Table 2: Hair color - Eye color data: Alpha ordered

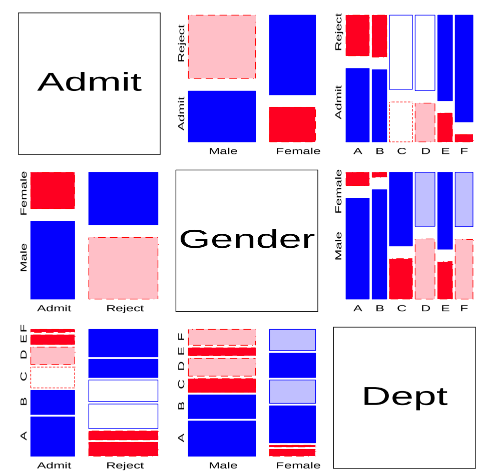
Eye color	Hair color: Blond	Black	Brown	Red
Blue	94	20	17	84
Brown	7	68	26	119
Green	10	15	14	54
Hazel	16	5	14	29

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		

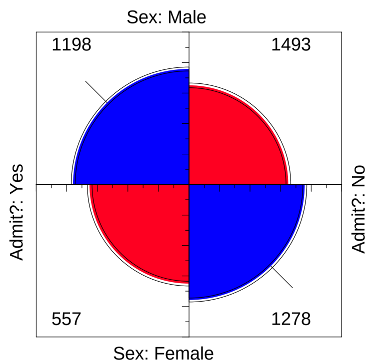
- **Comparisons**— Make visual comparisons easy
  - Visual grouping— connect with lines, make key comparisons contiguous
  - Baselines— compare *data* to *model* against a line, preferably horizontal



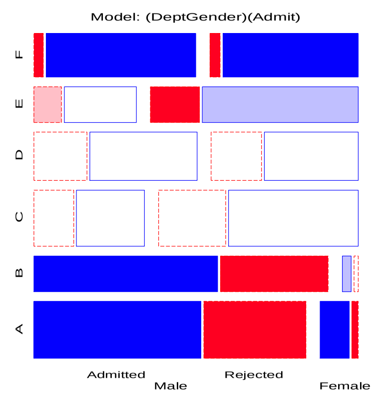
- **Comparisons**— Make visual comparisons easy
  - Small multiples (Tufte, 1983)— combine stratified graphs into coherent displays



- **Visual metaphors**
  - Quantitative data: **magnitude $\sim$ position along an axis**
  - Frequency data: **count $\sim$ area**



Fourfold display for 2×2 table



Mosaic plot for 3-way table

- **Exploratory methods**
  - Minimal assumptions (like non-parametric methods)
  - Show the *data*, not just *summaries*
  - Help detect *patterns, trends, anomalies*, suggest hypotheses
- **Plots for model-based methods**
  - Residual plots - departures from model, omitted terms, ...
  - Effect plots - estimated probabilities of response or log odds
  - Diagnostic plots - influence, violation of assumptions
- **Goals**
  - *VCD* and *SSSG* - Make these methods *available* and *accessible* in SAS
  - **Practical power = Statistical power $\times$ Probability of Use**
  - Today's goal: take-home knowledge
  - Tomorrow's goal: dynamic, interactive graphics for categorical data

VCD Macros & SAS/IML programs

- Macros, datasets available at `www.math.yorku.ca/SCS/vcd/`

Discrete distributions

`DISTPLOT`	Plots for discrete distributions
`GOODFIT`	Goodness-of-fit for discrete distributions
`ORDPLOT`	Ord plot for discrete distributions
`POISPLOT`	Poissonness plot
`ROOTGRAM`	Hanging rootograms

Two-way and n-way tables

`AGREE`	Observer agreement chart
`CORRESP`	Plot `PROC CORRESP` results
`FFOLD`	Fourfold displays for $2 \times 2 \times k$ tables (macro)
`FOURFOLD`	Fourfold displays for $2 \times 2 \times k$ tables (SAS/IML)
`SIEVE`	Sieve diagrams (SAS/IML)
`MOSAIC`	Mosaic displays (macro)
`MOSAICS`	SAS/IML modules for mosaic displays
`MOSMAT`	Mosaic matrices (macro)
`TABLE`	Construct a grouped frequency table, with recoding
`TRIPLOT`	Trilinear plots for $n \times 3$ tables

Model-based methods

`ADDVAR`	Added variable plots for logistic regression
`CATPLOT`	Plot results from `PROC CATMOD`
`HALFNORM`	Half-normal plots for generalized linear models
`INFLGLIM`	Influence plots for generalized linear models
`INFLOGIS`	Influence plots for logistic regression
`LOGODDS`	Plot empirical logits and probabilities for binary data
`POWERLOG`	Power calculations for logistic regression
`POWERRxC`	Power calculations for two-way frequency table
`POWER2x2`	Power calculations for a 2×2 table
`ROBUST`	Robust fitting for linear models

Utility macros

`DUMMY`	Create dummy variables
`LAGS`	Calculate lagged frequencies for sequential analysis
`PANELS`	Arrange multiple plots in a panelled display
`SORT`	Sort a dataset by the value of a statistic or formatted value
`Utility`	Graphics utility macros: `BARS`, `EQUATE`, `GDISPLA`, `GENSYM`, `GSKIP`, `LABEL`, `POINTS`, `PSCALE`

VCD Archive (`vcdprog.zip`) available to purchasers at:
`support.sas.com/publishing/bbu/56571_sample.html`

Discrete distributions

- **Counts of occurrences:** accidents, words in text, blood cells with some characteristic.
- **Data:** Basic outcome value, k , $k = 0, 1, \ldots$, and number of observations, n_k, with that value.
- **Example:** distributions of key "marker" words: *from*, *may*, *whilst*, . . . in *Federalist Papers* by James Madison, e.g., blocks of 200 words with *may*:

Occurrences (k)	0	1	2	3	4	5	6
Blocks (n_k)	156	63	29	8	4	1	1

- **Example:** 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

Discrete distributions

- **Questions:**
  - What process gave rise to the distribution?
  - Form of distribution: uniform, binomial, Poisson, negative binomial, geometric, etc.?
  - Estimate parameters
  - Visualize goodness of fit
- **For example:**
  - *Federalist Papers:* might expect a Poisson(λ) distribution.
  - *Families in Saxony:* might expect a Bin(n, p) distribution with $n = 12$. Perhaps $p = 0.5$ as well.

Fitting and graphing discrete distributions

VCD methods to fit, visualize, and diagnose discrete distributions:

- **Fitting:** `GOODFIT` macro fits uniform, binomial, Poisson, negative binomial, geometric, logarithmic series distributions (or any specified multinomial)
- **Hanging rootograms:** Sensitively assess departure between Observed, Fitted counts (`ROOTGRAM` macro)
- **Ord plots:** Diagnose form of a discrete distribution (`ORDPLOT` macro)
- **Poissonness plots:** Robust fitting and diagnostic plots for Poisson (`POISPLOT` macro)
- **Robust distribution plots** (`DISTPLOT` macro)

`GOODFIT` macro: Fitting discrete distributions

- `GOODFIT` macro fits uniform, binomial, Poisson, negative binomial, geometric, logarithmic series distributions (or any specified multinomial)
- E.g., Try fitting Poisson model

madfit.sas

```
title "Instances of 'may' in Federalist papers";
data madison;
   input count blocks;
   label count='Number of Occurrences'
         blocks='Blocks of Text';
datalines;
  0    156
  1     63
  2     29
  3      8
  4      4
  5      1
  6      1
;
 %goodfit(data=madison, var=count, freq=blocks,
      dist=poisson);
```

Fitting discrete distributions

The GOODFIT macro gives a table of observed and fitted frequencies, Pearson χ^2 residuals (CHI) and likelihood-ratio deviance residuals (DEV).

```
            Instances of 'may' in Federalist papers

COUNT    BLOCKS      PHAT         EXP        CHI          DEV

  0        156     0.51867    135.891     1.72499     6.56171
  1         63     0.34050     89.211    -2.77509    -6.62056
  2         29     0.11177     29.283    -0.05231    -0.75056
  3          8     0.02446      6.408     0.62890     1.88423
  4          4     0.00401      1.052     2.87493     3.26912
  5          1     0.00053      0.138     2.31948     1.98992
  6          1     0.00006      0.015     8.01267     2.89568
         ======    =======    =======
           262     0.99999    261.998
```


Fitting discrete distributions

In addition, it provides the overall goodness-of-fit tests:

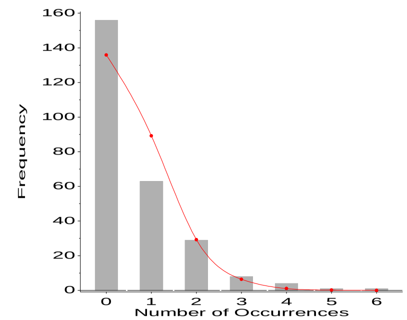
```
Goodness-of-fit test for data set MADISON

Analysis variable:        COUNT Number of Occurrences
Distribution:             POISSON
Estimated Parameters:     lambda = 0.6565

Pearson chi-square     = 88.92304707
Prob > chi-square      = 0

Likelihood ratio G2    = 25.243121314
Prob > chi-square      = 0.0001250511

Degrees of freedom     = 5
```

The poisson model does not fit! Why?

What's wrong with histograms?

- Discrete distributions often graphed as histograms, with a theoretical fitted distribution superimposed.

```
%goodfit(data=madison, var=count, freq=blocks,
   dist=poisson);
```



Problems:

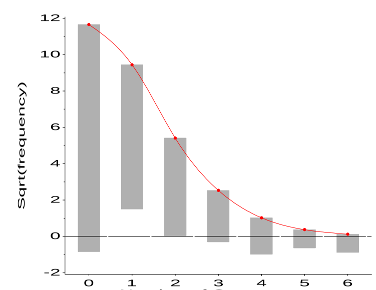
- largest frequencies dominate display
- must assess deviations vs. a curve

Hang & root them $\longrightarrow$ Hanging rootograms

Tukey (1972, 1977):

- shift histogram bars to the fitted curve $\longrightarrow$ judge deviations vs. horizontal line.
- plot $\sqrt{\text{freq}}$ $\longrightarrow$ smaller frequencies are emphasized.

```
%goodfit(data=madison, var=count, freq=blocks,
  dist=poisson, out=fit);
%rootgram(data=fit, var=count, obs=blocks);
```



Ord plots: Diagnose form of discrete distribution

- How to tell which discrete distributions are likely candidates?
- Ord (1967): for each of Poisson, Binomial, Negative Binomial, and Logarithmic Series distributions,
  - plot of kp_k / p_{k-1} against k is linear
  - signs of intercept and slope $\longrightarrow$ determine the form, give rough estimates of parameters

Slope (b)	Intercept (a)	Distribution (parameter)	Parameter estimate
0	$+$	Poisson (λ)	$\lambda = a$
$-$	$+$	Binomial (n, p)	$p = b/(b-1)$
$+$	$+$	Neg. binomial (n,p)	$p = 1 - b$
$+$	$-$	Log. series (θ)	$\theta = b$
			$\theta = -a$

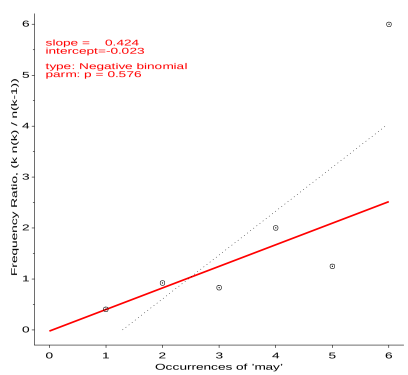
- Fit line by WLS, using $\sqrt{n_k}$ as weights

Ord plots

- ORDPLOT macro

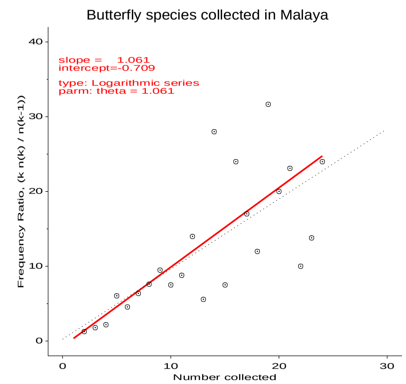
```
%ordplot(data=madison, count=Count, freq=blocks);
```

  - Diagnoses distribution as NegBin
  - Estimates $\hat{p} = 0.576$

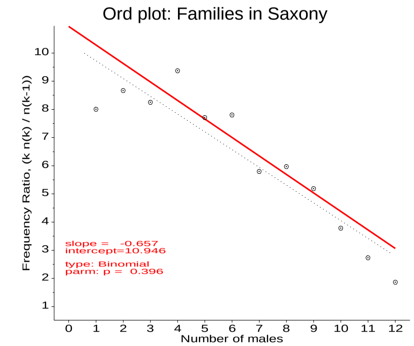


Ord plots

Other distributions:



Logarithmic series



Binomial

Robust distribution plots: Poisson

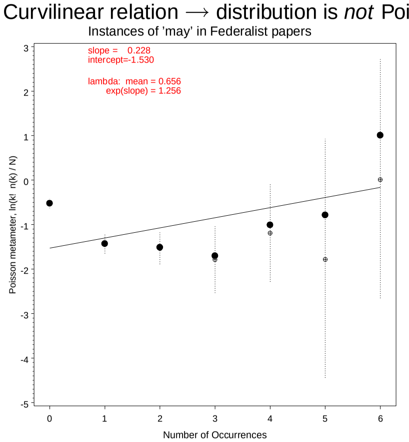
- Ord plots lack robustness
  - one discrepant freqency, n_k affects points for both k and $k+1$
- Robust plots for Poisson distribution (Hoaglin and Tukey, 1985)
  - For Poisson, plot ***count metameter*** = $\phi(n_k) = \log_e(k!\, n_k / N)$ vs. k
  - Linear relation $\Rightarrow$ Poisson, slope gives $\hat{\lambda}$
  - CI for points, diagnostic (influence) plot
  - POISPLOT macro

POISPLOT macro: example

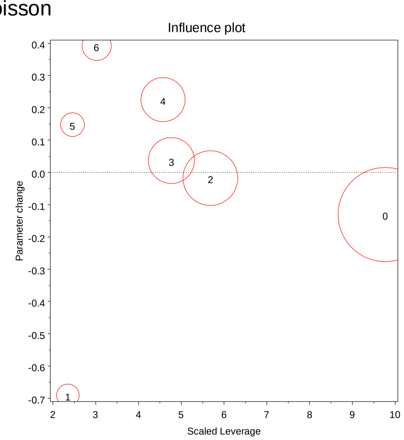
```
title "Instances of 'may' in Federalist papers";
data madison;
   input count blocks;
   label count='Number of Occurrences'
         blocks='Blocks of Text';
datalines;
  0    156
  1     63
  2     29
  3      8
  4      4
  5      1
  6      1
;
%poisplot(data=madison,count=count, freq=blocks);
```


POISPLOT macro: output

Curvilinear relation ⟶ distribution is *not* Poisson



Poissonness plot



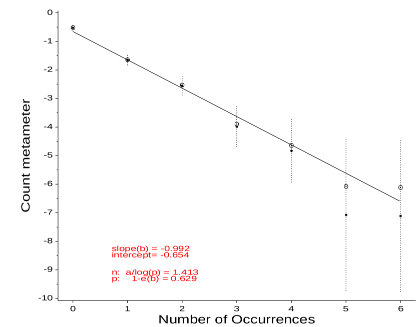
Influence plot for change in λ

Generalized robust distribution plots

Other distributions: Analogous plots, for suitable count metameter, $\phi(n_k)$ vs. k.

- Linear relation ⇒ correct distribution, slope gives parameter estimates
- CI reflect variability of the individual counts, n_k
- DISTPLOT macro

```
%distplot(data=madison, count=count, freq=blocks,
      dist=negbin);
```



Contingency tables

- Two-way tables
  - 2×2 tables — Visualize odds ratio (FFOLD macro)
  - $2 \times 2 \times k$ tables — Homogeneity of association
  - $r \times 3$ tables — Trilinear plots (TRIPLOT macro)
  - $r \times c$ tables — Visualize association (SIEVE program)
  - $r \times c$ tables — Visualize association (MOSAIC macro)
  - Square $r \times r$ tables — Visualize agreement (AGREE program)
- *n*-way tables
  - Fit loglinear models, visualize lack-of-fit — (MOSAIC macro)
  - Test & visualize partial association — (MOSAIC macro)
  - Visualize pairwise association — (MOSMAT macro)
  - Visualize conditional association — (MOSMAT macro)
  - Visualize loglinear structure — (MOSMAT macro)
- Correspondence analysis and MCA — (CORRESP macro)

Methods for 2×2 tables

- Bickel et al. (1975): data on admissions to graduate depatments at U. C. Berkeley in 1973.
- Aggregate data for the six largest departments:

Table 3: Admissions to Berkeley graduate programs

	Admitted	Rejected	Total	% Admitted
Males	1198	1493	2691	44.52
Females	557	1278	1835	30.35
Total	1755	2771	4526	38.78

- Evidence for gender bias?
  - $G^2_{(1)} = 93.7, \chi^2_{(1)} = 92.2,\ p < 0.0001$
  - Odds ratio, $\theta = \frac{\text{Odds(Admit}\,|\,\text{Male)}}{\text{Odds(Admit}\,|\,\text{Female)}} = 1.84$
  - Males 84% more likely to be admitted.

Standard analysis: PROC FREQ

```
proc freq data=berkeley;
  weight freq;
  tables gender*admit / chisq;
```

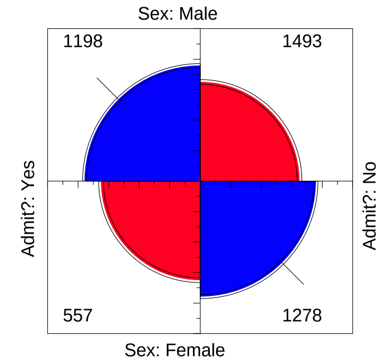
Output:

```
       Statistics for Table of gender by admit

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1     92.2053    <.0001
Likelihood Ratio Chi-Square    1     93.4494    <.0001
Continuity Adj. Chi-Square     1     91.6096    <.0001
Mantel-Haenszel Chi-Square     1     92.1849    <.0001
Phi Coefficient                       0.1427
```

How to visualize and interpret?

Fourfold displays for 2 × 2 tables

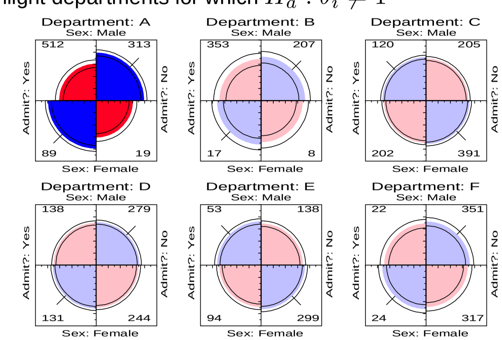
- ***Quarter circles***: radius $\sim \sqrt{n_{ij}} \Rightarrow$ **area** $\sim$ **frequency**
- ***Independence***: Adjoining quadrants $\approx$ align
- ***Odds ratio***: ratio of areas of diagonally opposite cells
- ***Confidence rings***: Visual test of $H_0 : \theta = 1 \leftrightarrow$ adjoining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



- Confidence rings do not overlap: $\theta \neq 1$

Fourfold displays for 2 × 2 × *k* tables

- Data in Table 3 had been pooled over departments
- Stratified analysis: one fourfold display for each department
- Each 2×2 table standardized to equate marginal frequencies
- Shading: highlight departments for which $H_a : \theta_i \neq 1$



- Only one department (A) shows association; $\theta_A = 0.349 \rightarrow$ women $(0.349)^{-1} = 2.86$ times as likely as men to be admitted.

What happened here?

Simpson's paradox:

- Aggregate data are misleading because they falsely assume men and women apply *equally* in each field.
- But:
  - Large differences in admission rates across departments.
  - Men and women apply to these departments differentially.
  - Women applied in large numbers to departments with low admission rates.
- (This ignores possibility of structural bias against women: differential funding of fields to which women are more likely to apply.)
- Other graphical methods can show these effects.

The FOURFOLD program and the FFOLD macro

- The FOURFOLD program is written in SAS/IML.
- The FFOLD macro provides a simpler interface.
- Printed output: (a) significance tests for individual odds ratios, (b) tests of homogeneity of association (here, over departments) and (c) conditional association (controlling for department).

Plot by department:

berk4f.sas

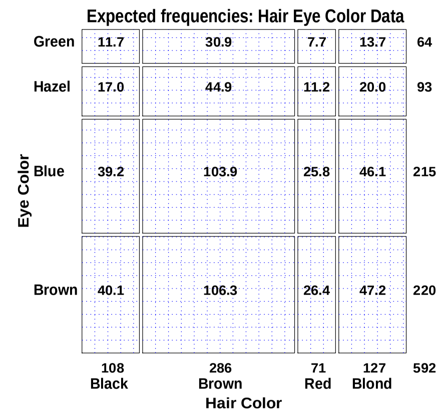
```
%include catdata(berkeley);

%ffold(data=berkeley,
   var=Admit Gender,        /* panel variables   */
   by=Dept,                 /* stratify by dept  */
   down=2, across=3,        /* panel arrangement */
   htext=2);                /* font size         */
```

Aggregate data: first sum over departments, using the TABLE macro:

```
%table(data=berkeley, out=berk2,
   var=Admit Gender,        /* omit dept          */
   weight=count,            /* frequency variable */
   order=data);
%ffold(data=berk2, var=Admit Gender);
```

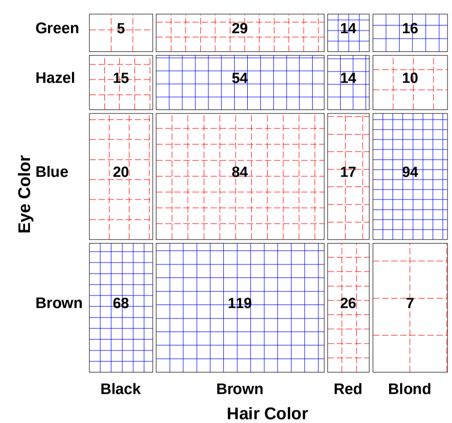

Two-way frequency tables: Sieve diagrams

- **count ∼ area**
  - When row/col variables are independent, $n_{ij} \sim n_{i+} n_{+j}$
  - $\Rightarrow$ each cell can be represented as a rectangle, with area = height × width ∼ frequency, n_{ij}



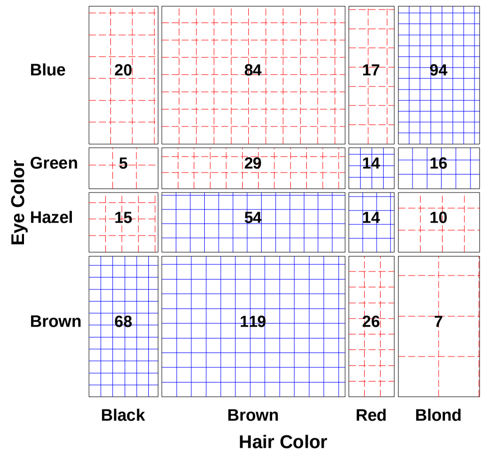
Sieve diagrams

- Height/width ∼ marginal frequencies, n_{i+}, n_{+j}
- Area ∼ expected frequency, $\sim n_{i+} n_{+j}$
- Shading ∼ observed frequency, n_{ij}, color: sign$(n_{ij} - \hat{m}_{ij})$.
- ***Independence***: Shown when density of shading is uniform.



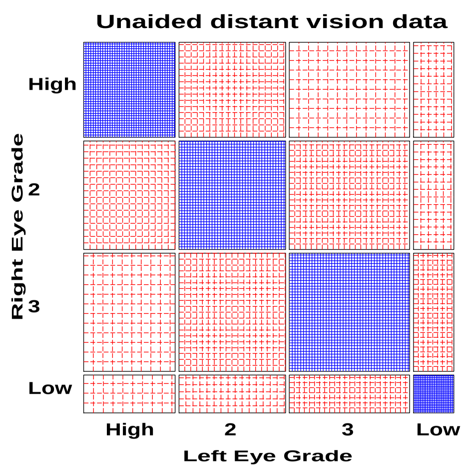
Sieve diagrams

- **Effect ordering**: Reorder rows/cols to make the pattern coherent



Sieve diagrams

- Vision classification data for 7477 women



Sieve diagrams: Example

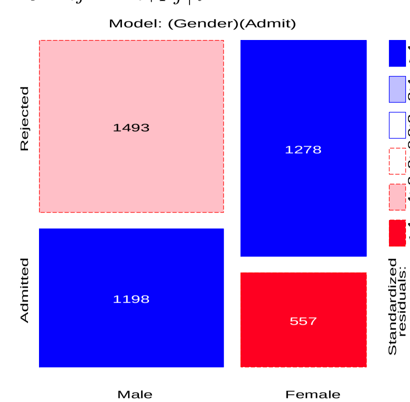
sieve2.sas

```
proc iml;
  %include iml(sieve);
    *-- frequency table;
  tab = {1520   266   124    66,
          234  1512   432    78,
          117   362  1772   205,
           36    82   179   492 };
    *-- variable and level names;
  vnames = {'Right Eye Grade' 'Left Eye Grade'};
  lnames = { 'High' '2' '3' 'Low',
             'High' '2' '3' 'Low'};
  title  = {'Unaided distant vision data'};
    *-- Global options;
  font='hwpsl011';
  run sieve(tab, vnames, lnames, title );
quit;
```

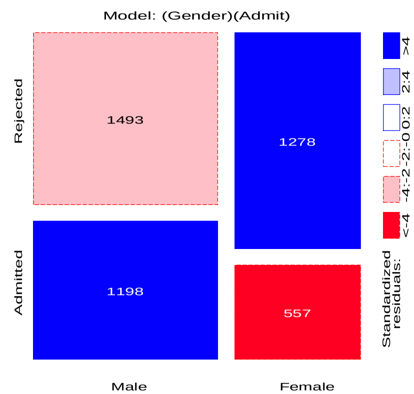

Mosaic displays and Log-linear Models

Hartigan and Kleiner (1981), Friendly (1994, 1999):

- ***Width*** $\sim$ one set of marginals, n_{i+}
- ***Height*** $\sim$ relative proportions of other variable, $p_{j\,|\,i} = n_{ij}/n_{i+}$
- $\Rightarrow$ **area** $\sim$ **frequency**, $n_{ij} = n_{i+}p_{j\,|\,i}$



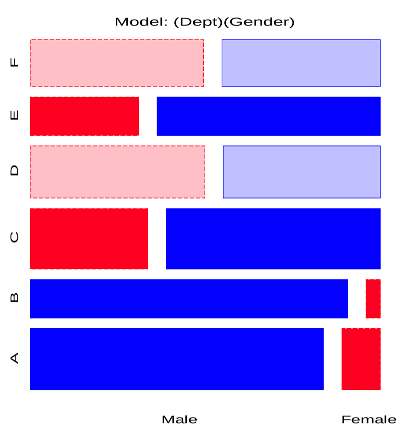
- ***Shading***: Sign and magnitude of Pearson χ^2 residual, $d_{ij} = (n_{ij} - \hat{m}_{ij})/\sqrt{\hat{m}_{ij}}$ (or L.R. G^2)
  - Sign: $-$ negative in red; $+$ positive in blue
  - Magnitude: intensity of shading: $|d_{ij}| > 0, 2, 4, \ldots$
- ***Independence***: Rows $\approx$ align, *or* cells are empty!
- E.g., aggregate Berkeley data, independence model:



Mosaic displays

Departments $\times$ Gender:

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



- Model [Dept] [Gender]: $G^2_{(5)} =$ 1220.6.
- ***Note***: Departments ordered A–F by overall rate of admission.

Mosaic displays for multiway tables

- Generalizes to n-way tables: divide cells recursively
- Can fit *any* log-linear model (e.g., 3-way),

Table 4: Log-linear Models for Three-Way Tables

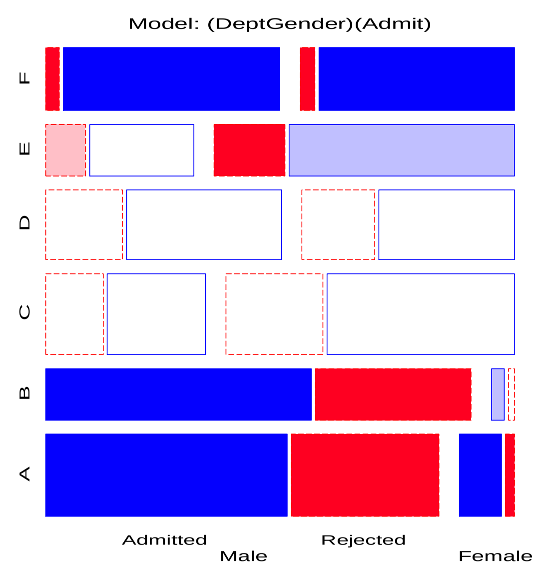
Model	**Model symbol**	**Independence interpretation**
Mutual independence	$[A][B][C]$	$A \perp B \perp C$
Joint independence	$[AB][C]$	$(A\,B) \perp C$
Conditional independence	$[AC][BC]$	$(A \perp B) \mid C$
All two-way associations	$[AB][AC][BC]$	(none)
Saturated model	$[ABC]$	(none)

e.g., the model for conditional independence ($A \perp C \mid B$):

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

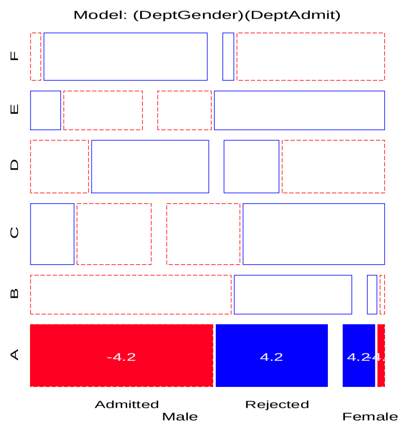
- Each mosaics shows:
  - **DATA** (size of tiles)
  - (some) **marginal** frequencies (spacing $\longrightarrow$ visual grouping)
  - **RESIDUALS** (shading) — what associations have been omitted?

- E.g., Joint independence (null model, Admit as response) [$G^2_{(11)}$ = 877.1]:



Mosaic displays for multiway tables

- Visual fitting:
  - Pattern of lack-of-fit (residuals) ⟶ "better" model— smaller residuals
  - "cleaning the mosaic" ⟶ "better" model— empty cells
  - best done interactively!



- E.g., Add [Dept Admit] association ⟶ Conditional independence:
  - Fits poorly, overall ($G^2_{(6)}$ = 21.74)
  - But, only in Department A!

Software for Mosaic Displays

- **Demonstration web applet**: `http://www.math.yorku.ca/SCS/Online/mosaics/`
  - Runs the *current* version of mosaics via a cgi-script
  - Can run *sample data*, *upload* a data file, *enter* data in a form.
  - Choose model fitting and display options (not all supported).
- **Documentation & software**: `http://www.math.yorku.ca/SCS/mosaics.html`
- **Examples**: Many in *VCD* and on web site
- **SAS/IML modules**: `mosaics.sas` program
  - Enter frequency table directly in SAS/IML, or read from a SAS dataset.
  - Most flexible:
    - Select, collapse, reorder, re-label table levels using SAS/IML statements
    - Specify structural 0s, fit specialized models (e.g., quasi-independence)
    - Interface to models fit using `PROC GENMOD`

Software for Mosaic Displays

- **Macro interface**: `mosaic` macro, `table` macro, `mosmat` macro
- `mosaic` **macro**
  - Easiest to use:
    - Direct input from a SAS dataset
    - No knowledge of SAS/IML required
    - Reorder table variables; collapse, reorder table levels with `table` macro
    - Convenient interface to *partial mosaics* (`BY=`)
- `table` **macro**
  - Create frequency table from raw data
  - Collapse, reorder table categories
  - Re-code table categories using SAS formats, e.g., `1='Male' 2='Female'`
- `mosmat` **macro**
  - Mosaic matrices— analog of scatterplot matrix (Friendly, 1999)

Software for Mosaic Displays

- **Other implementations**:
  - JMP and SAS/INSIGHT both provide rudimentary mosaic displays (two-way only, no interface with model-fitting engines (shame!).
  - The R-Project (`http://www.r-project.org`) now provides the `vcd` package, implementing most of the graphical methods from *VCD*.
  - Truly interactive mosaic displays have been implemented in:
    - Lisp-Stat (ViSta)— `http://forrest.psych.unc.edu/research/`
    - Java (Mondrian)— `http://http:www1.math.uni-augsburg.de/Mondrian/`

mosaic macro example: Berkeley data

berkeley.sas

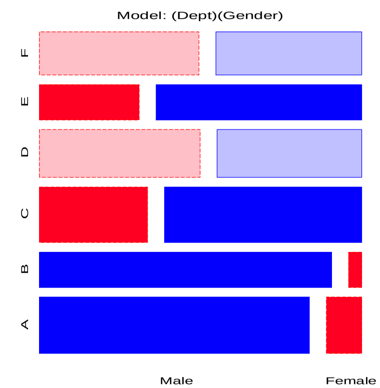
```
title 'Berkeley Admissions data';
proc format;
   value admit 1="Admitted" 0="Rejected"          ;
   value dept  1="A" 2="B" 3="C" 4="D" 5="E" 6="F";
        value $sex  'M'='Male'   'F'='Female';
data berkeley;
   do dept = 1 to 6;
      do gender = 'M', 'F';
         do admit = 1, 0;
            input freq @@;
            output;
   end; end; end;
/* -- Male --   - Female- */
/* Admit  Rej  Admit Rej */
datalines;
     512  313    89   19  /* Dept A */
     353  207    17    8  /*      B */
     120  205   202  391  /*      C */
     138  279   131  244  /*      D */
      53  138    94  299  /*      E */
      22  351    24  317  /*      F */
;
```


mosaic macro example: Berkeley data

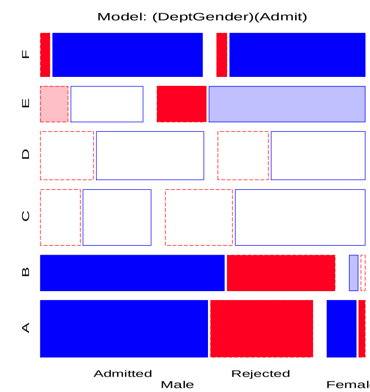
mosaic9m.sas

```
goptions hsize=7in vsize=7in;

%include catdata(berkeley);

*-- apply character formats to numeric table variables;
%table(data=berkeley,
    var=Admit Gender Dept,
    weight=freq,
    char=Y, format=admit admit. gender $sex. dept dept.,
    order=data, out=berkeley);

%mosaic(data=berkeley,
    vorder=Dept Gender Admit, /* reorder variables */
    plots=2:3,                /* which plots?      */
    fittype=joint,            /* fit joint indep.  */
    split=H V V, htext=3);    /* options           */
```


mosaic macro example: Berkeley data



Two-way, Dept. by Gender



Three-way, Dept. by Gender by Admit

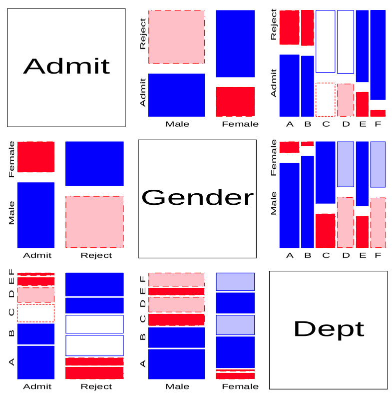
SAS/IML example

mosaic9.sas

```
proc iml;
   *-- Berkeley Admissions data;
   dim = {2 2 6};
   vnames = {"Admit" "Gender" "Dept"};          /* var names   */
   lnames = {                                   /* level names */
     "Admitted" "Rejected" " "  " "  " "  " ",
     "Male"   "Female" " "  " "  " "  " ",
     "A" "B" "C" "D" "E" "F"};
          /* Admit Not Admit Not */
   table = { 512  313,   89   19,
             353  207,   17    8,
             120  205,  202  391,
             138  279,  131  244,
              53  138,   94  299,
              22  351,   24  317};
   reset storage=mosaic.mosaic;
   load module=_all_;               *-- Load mosaic modules;

   title = 'Model: &MODEL';
   plots=2:3;
   fittype='joint';
   run mosaic(dim, table, vnames, lnames, plots, title);
 quit;
```

mosmat macro: Mosaic matrices

```
%include catdata(berkeley);
%mosmat(data=berkeley,
   vorder=Admit Gender Dept, sort=no);
```



Logit models

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

- Admit $\perp$ Gender $|$ Dept (conditional independence)

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

$$\leftrightarrow L_{ij} = \log(m_{ij1}/m_{ij2}) = \alpha + \beta_i^{\text{Dept}}$$

- Admit $\perp$ Gender $|$ Dept, except for Dept. A

$$\log\, m_{ijk} = \mu + \lambda_i^A + \lambda_j^D + \lambda_k^G + \lambda_{ij}^{AD} + \lambda_{jk}^{DG} + \delta_{j=1}\lambda_{ik}^{AG}$$

$$\leftrightarrow L_{ij} = \log(m_{ij1}/m_{ij2}) = \alpha + \beta_i^{\text{Dept}} + \delta_{j=1}\beta^{\text{Gender}}$$

where,

- $L_{ij} = \log(m_{ij1}/m_{ij2})$: log odds of admission for males as vs. females,
- β_i^{Dept}: effect on admissions of department,
- $\delta_{j=1}\beta^{\text{Gender}}$: effect of gender in Dept. A.

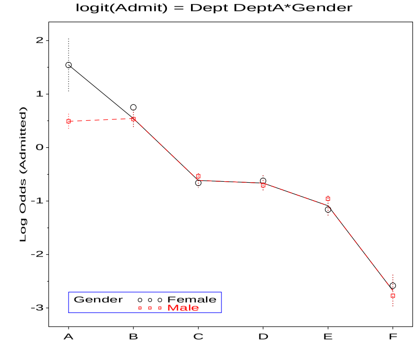
Logit models

- **Fitting procedures**
  - `PROC CATMOD`
  - `PROC LOGISTIC`
  - `PROC GENMOD` / `dist=poisson`
  - SAS/INSIGHT (Fit Y X) Options $\longrightarrow$ Distribution poisson
- **Visualization procedures**
  - `CATPLOT` macro - plot predicted, observed log odds from `CATMOD`
  - `INFLGLIM` macro - influence plots for generalized linear models
  - `HALFNORM` macro - half-normal plot of residuals for generalized linear models

Plots for logit models

- Fit: `PROC CATMOD`; plot: `CATPLOT` macro
- Admit $\perp$ Gender $|$ Dept, except for Dept. A

```
proc catmod order=data data=berkeley;
   ...
   response / out=predict;
   model admit = dept dept1AG / ml;
%catplot(data=predict, xc=dept, class=gender,
    type=FUNCTION, z=1.96, legend=legend1);
```



Plots for logit models

- Fit: PROC CATMOD; plot: CATPLOT macro
- Admit ⊥ Gender | Dept, except for Dept. A

··· catberk6.sas ···

```
%include catdata(berkeley);
data berkeley;
   set berkeley;
   *-- Dummy variable for Gender in Dept A;
   dept1AG = (gender='F') *  (dept=1);
   format dept dept.;

proc catmod order=data
            data=berkeley;
   weight freq;
   population dept gender;
   direct dept1AG;
   response / out=predict;
   model admit = dept dept1AG / ml;
  run;
  ...
```


Plots for logit models

PROC CATMOD output:

```
        Maximum Likelihood Analysis of Variance

     Source              DF   Chi-Square    Pr > ChiSq
     --------------------------------------------------
     Intercept            1       291.22        <.0001
     dept                 5       571.45        <.0001
     dept1AG              1        16.04        <.0001

     Likelihood Ratio     5         2.68        0.7489

        Analysis of Maximum Likelihood Estimates

                               Standard     Chi-
Parameter      Estimate        Error      Square   Pr > ChiSq
--------------------------------------------------------------
Intercept       -0.6685       0.0392     291.22       <.0001
dept       A     1.1606       0.0705     271.21       <.0001
           B     1.2113       0.0802     227.95       <.0001
           C     0.0528       0.0687       0.59       0.4426
           D     0.00358      0.0727       0.00       0.9607
           E    -0.4210       0.0871      23.34       <.0001
dept1AG          1.0521       0.2627      16.04       <.0001
```

How to interpret?

Plots for logit models

PROC CATMOD: observed and predicted logits:

··· catberk6.sas ···

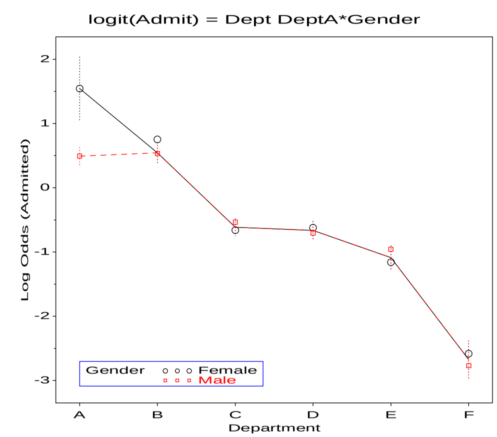
```
proc print data=predict;
  id dept gender;
  var _obs_ _pred_ _sepred_;
  format _numeric_  6.3 dept dept.;
  where(_type_='FUNCTION');
```

dept	gender	_OBS_	_PRED_	_SEPRED_
A	M	0.492	0.492	0.072
A	F	1.544	1.544	0.253
B	M	0.534	0.543	0.086
B	F	0.754	0.543	0.086
C	M	-0.536	-0.616	0.069
C	F	-0.660	-0.616	0.069
D	M	-0.704	-0.665	0.075
D	F	-0.622	-0.665	0.075
E	M	-0.957	-1.090	0.095
E	F	-1.157	-1.090	0.095
F	M	-2.770	-2.676	0.152
F	F	-2.581	-2.676	0.152

Plots for logit models

··· catberk6.sas

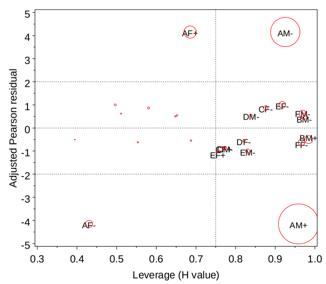
```
title 'logit(Admit) = Dept DeptA*Gender';
%catplot(data=predict, x=dept, class=gender,
   type=FUNCTION,       /* plot the log odds */
   z=1.96);             /* 95% error bars    */
```



Diagnostic plots for Generalized Linear Models

INFLGLIM macro: Influence plots for generalized linear models (Williams, 1987)

- Fit: PROC GENMOD; calculates additional diagnostic measures (Hat value, Cook's D, etc.)
- Plot: measures of residual (GY=$\Delta\chi^2$, χ^2 residual) vs. leverage (GX=hat value), bubble size (area, radius) $\sim$ Cook's D.
- $\longrightarrow$ which cells have undue impact on fitted model?



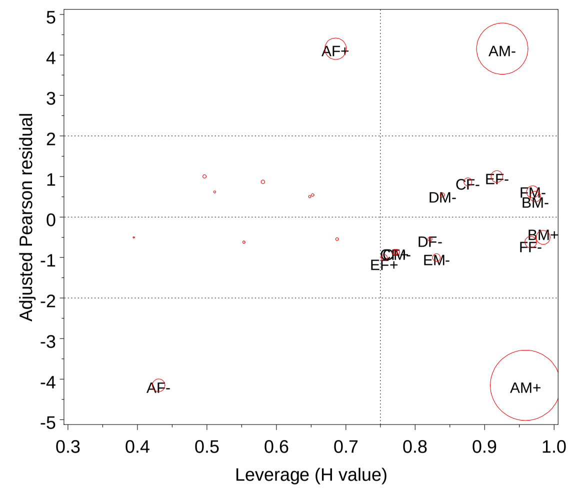
INFLGLIM macro: Example

- Berkeley data, model $[AD][GD] \leftrightarrow L_{ij} = \alpha + \beta_i^{\text{Dept}}$

genberk1.sas

```
%include catdata(berkeley);
*-- make a cell ID variable, joining factors;
data berkeley;
   set berkeley;
   cell = trim(put(dept,dept.)) ||
          gender ||
          trim(put(admit,yn.));

%inflglim(data=berkeley,
   class=dept gender admit,
   resp=freq,
   model=admit|dept gender|dept,
   dist=poisson,
   id=cell,
   gx=hat, gy=streschi);
```


INFLGLIM macro: Example



- All cells which do not fit ($|r_i| > 2$) are for department A.
- Males applying to dept A have large leverage $\Rightarrow$ large influence (Cook's D)

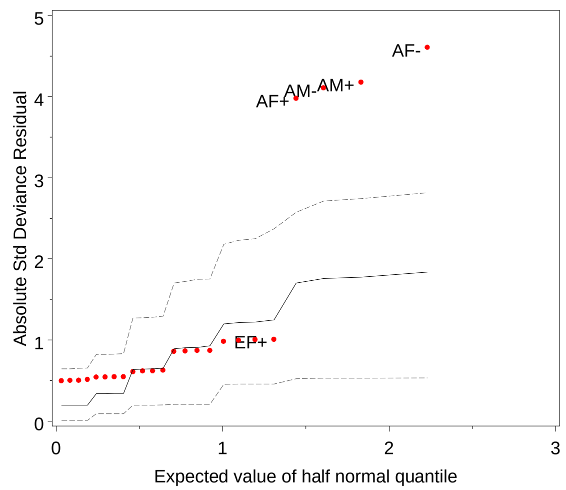
Diagnostic plots for Generalized Linear Models

HALFNORM macro: Half-normal plot of residuals (Atkinson, 1981)

- Plot ordered *absolute* residuals, $|r|_{(i)}$ vs. expected normal values, $|z|_{(i)}$
- Standard normal confidence envelope not suitable for GLMs
- Simulate reference 'line' and envelope with simulated confidence intervals

··· genberk1.sas

```
%halfnorm(data=berkeley,
   class=dept gender admit,
   resp=freq,
   model=dept|gender dept|admit,
   dist=poisson, id=cell);
```



- Points with largest |residual| labeled
- The model fits well, except in department A.

Logistic regression models

- **Response variable**:
  - Binary response: success/failure, vote: yes/no
  - Ordinal response: none, some, severe depression
  - Polytomous response: vote Liberal, Tory, Aliance, NDP
- **Explanatory variables**:
  - Quantitative regressors: age, dose
  - Transformed regressors: $\sqrt{\text{age}}$, log(dose)
  - Polynomial regressors: age^2, age^3, $\cdots$
  - Categorical predictors: treatment, sex
  - Interaction regessors: treatment $\times$ age, sex $\times$ age

Logistic regression models: Binary response

- For a binary response, $Y \in (0, 1)$, let $\boldsymbol{x}$ be a vector of p regressors, and π_i be the probability, $\Pr(Y = 1 \,|\, \boldsymbol{x})$.
- The logistic regression model is a linear model for the *log odds*, or *logit* that $Y = 1$, given the values in $\boldsymbol{x}$,

$$\begin{aligned} \operatorname{logit}(\pi_i) \equiv \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \alpha + \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta} \\ &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \end{aligned}$$

- An equivalent (non-linear) form of the model may be specified for the probability, π_i, itself,

$$\pi_i = \{1 + \exp(-[\alpha + \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}])\}^{-1}$$

- The logistic model is a *linear model* for the log odds, but also a *multiplicative* model for the odds of "success,"

$$\frac{\pi_i}{1-\pi_i} = \exp(\alpha + \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}) = \exp(\alpha)\exp(\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta})$$

so, increasing x_{ij} by 1 increases $\operatorname{logit}(\pi_i)$ by β_j, and multiplies the odds by e^{β_j}.

Logistic regression models: Binary response

- **Fitting**: `PROC LOGISTIC` (or `ROBUST` macro— M-estimation)
  - Data:
    - Frequency form (from `PROC FREQ`)— when all predictors are discrete
    - Case form— when any predictors are quantitative
  - Models:
    - `CLASS` statement (V7+)— no need for dummy variables
      - discrete predictors
      - can specify *order* and *parameterization* (effect, polynomial, reference cell)
    - `MODEL` statement— allows `GLM` syntax, e.g.,
      model Better = Sex — Treat — Age @2;

Logistic regression models: Binary response

- **Visualization**:
  - Goal: *see* and *understand* the data and fitted model
  - LOGODDS macro: Plot observed responses, fitted and smoothed probabilities
  - Model plots:
    - OUTPUT statement $\longrightarrow$
      - fitted $\hat{\pi}_i$, lower/upper $(1-\alpha)$ CI, and/or
      - fitted logit, $(\alpha + \boldsymbol{x}_i^{\mathsf{T}} \hat{\boldsymbol{\beta}}) \pm z_{1-\alpha/2} se(\text{logit})$
    - Plot with standard procedures (PROC GCHART, GPLOT)
    - Utility macros (BARS, LABEL, POINTS, PSCALE, etc.) for custom displays
  - Effect plots— plot hierarchical subset of effects, averaging over those not included.
  - INFLOGIS macro: Influence plots for logistic regression models
  - ADDVAR macro: Added variable plots for new predictors or transformations of old

Example: Arthritis treatment data

- Predictors: Sex, Treatment (treated, placebo), Age
- Response: improvement (none, some, marked)
  - Consider first as binary response: None vs. (Some or Marked)='Better'
- Data in case form:

arthrit.sas

```
data arthrit;
   length treat $7. sex $6. ;
   input id treat $ sex $ age improve @@ ;
   case = _n_;
   better  = (improve > 0);  *-- Make binary response;
datalines ;
57 Treated Male   27 1   9 Placebo Male   37 0
46 Treated Male   29 0  14 Placebo Male   44 0
77 Treated Male   30 0  73 Placebo Male   50 0
  ... (observations omitted)
56 Treated Female 69 1  42 Placebo Female 66 0
43 Treated Female 70 1  15 Placebo Female 66 1
                        71 Placebo Female 68 1
                         1 Placebo Female 74 2
;
```

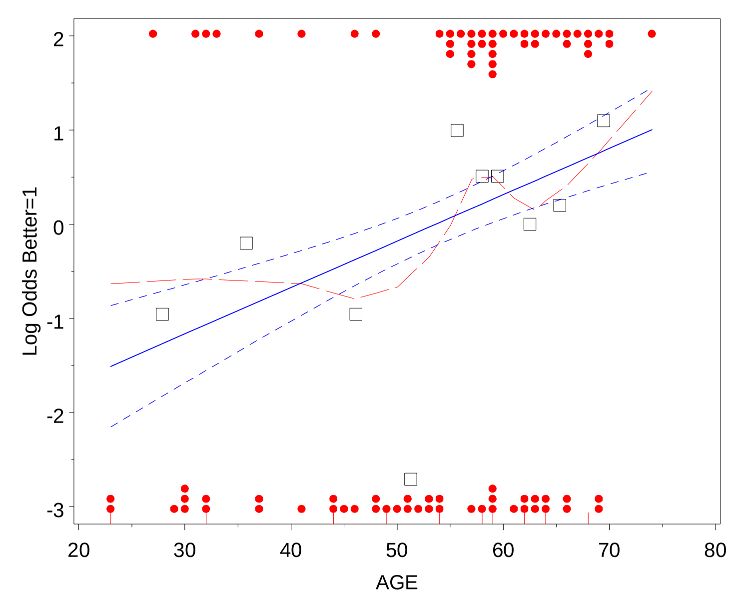

LOGODDS macro: Empirical logit plots

- **Linearity**: Is a linear relation realistic?
- **Smoothing**: Discrete data often requires smoothing to see!

The LOGODDS macro:

- Show the data: Plot (0/1) responses [stacked or jittered]
- Divide X into groups (e.g., deciles), emprical logit, $\log\left(\frac{y_i+1/2}{n_i-y_i+1/2}\right)$, for each
- Linear logistic regression, plus smoothed curve (LOWESS macro)

```
%include catdata(arthrit);
%logodds(data=arthrit,
    x=age, y=Better,  /* vars to plot            */
    smooth=0.5,       /* LOWESS smoothing parameter */
    plot=logit);      /* plot on logit scale      */
```

PROC LOGISTIC: Model fitting and plotting

- Specify ordering of response levels (order= or descending options)
- Specify parameterizations for CLASS variables
- OUTPUT statement to get fitted logits and probabilities

glogist1c.sas ···

```
proc logistic data=arthrit descending;
   class sex (ref=last) treat (ref=first) / param=ref;
   model better = sex  treat  age;
   output out=results
          p=prob l=lower u=upper
          xbeta=logit stdxbeta=selogit / alpha=.33;
```

The output includes:

```
              Type III Analysis of Effects

                                Wald
        Effect       DF    Chi-Square    Pr > ChiSq

        sex           1        6.2576        0.0124
        treat         1       10.7596        0.0010
        age           1        5.5655        0.0183
```



```
            Analysis of Maximum Likelihood Estimates

                                   Standard        Wald
Parameter           DF  Estimate      Error  Chi-Square  Pr > ChiSq

Intercept            1   -4.5033     1.3074     11.8649      0.0006
sex       Female     1    1.4878     0.5948      6.2576      0.0124
treat     Treated    1    1.7598     0.5365     10.7596      0.0010
age                  1    0.0487     0.0207      5.5655      0.0183

                     Odds Ratio Estimates

                                Point          95% Wald
  Effect                     Estimate      Confidence Limits

  sex   Female vs Male          4.427       1.380     14.204
  treat Treated vs Placebo      5.811       2.031     16.632
  age                           1.050       1.008      1.093
```

Parameter estimates (reference cell coding):

- $\beta_1 = 1.49 \Rightarrow$ Females $e^{1.49}$=4.43 times more likely to be better than Males
- $\beta_2 = 1.76 \Rightarrow$ Treated $e^{1.76}$=5.81 times more likely to be better than Placebo
- $\beta_3 = 0.0487 \Rightarrow$ odds ratio=1.05 $\Rightarrow$ odds of improvement increase 5% each year. Over 10 years, odds of improvement multiplied by $e^{10\times 0.0486} = 1.63$, a 63% increase.

PROC LOGISTIC: Model plots

- Plots of fitted values from the dataset specified on the OUTPUT statement
- Plot either predicted probabilities or logits

The first few observations from the results dataset:

```
id sex    treat  age better   prob  lower  upper  logit selogit

57 Male Treated  27    1     0.194  0.103  0.334 -1.427  0.758
 9 Male Placebo  37    0     0.063  0.032  0.120 -2.700  0.725
46 Male Treated  29    0     0.209  0.115  0.350 -1.330  0.728
14 Male Placebo  44    0     0.086  0.047  0.152 -2.358  0.658
77 Male Treated  30    0     0.217  0.122  0.357 -1.281  0.713
73 Male Placebo  50    0     0.112  0.065  0.188 -2.066  0.622
 ...
```


PROC LOGISTIC: Model plots

Basic plots:

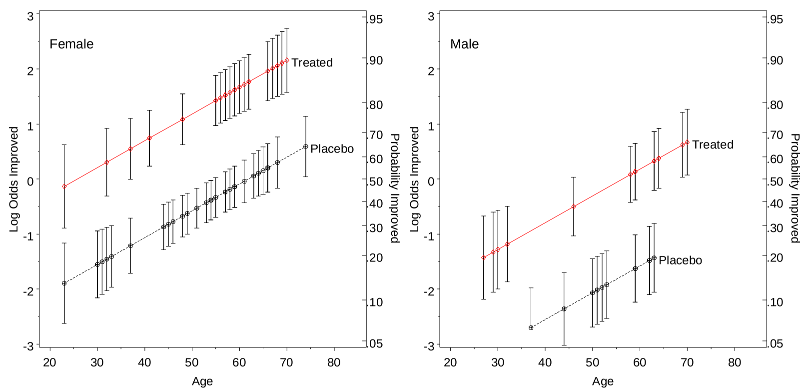
- Plot either logit or probability vs. one predictor (continuous or most levels)
- Separate curves for one factor
- Separate panels for all others (BY statement)

```
proc gplot data=results;
   plot (logit prob) * age = treat;
   by sex;
   symbol1 v=circle i=join l=3 c=black;
   symbol2 v=dot    i=join l=1 c=red;
```

PROC LOGISTIC: Model plots

Enhanced plots:

- Plot on logit scale, with probability scale at right (PSCALE macro)
- Show 67% error bars $\approx \pm 1$ se (BARS macro)
- Custom legend and panel labels (LABEL macro)



PROC LOGISTIC: Model plots

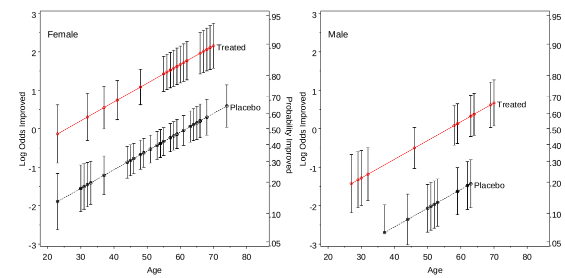
Enhanced plots:

··· glogist1c.sas ···

```
  *-- Error bars, on logit scale;
%bars(data=results, var=logit,
   class=age, cvar=treat, by=age,
   barlen=selogit, out=bars);

  *-- Custom legends and panel labels;
%label(data=results, y=logit, x=age, xoff=1, cvar=treat,
   by=sex, subset=last.treat, out=label1, pos=6, text=treat);
%label(data=results, y=2.5, x=20, size=2,
   by=sex, subset=first.sex, out=label2, pos=6, text=sex);

  *-- Probability scales at right;
%pscale(out=pscale,
   byvar=sex, byval=%str('Female','Male'));

  *-- Join ANNOTATE datasets;
data bars;
   set label1 label2 bars pscale;
proc sort;
   by sex;
```


··· glogist1c.sas

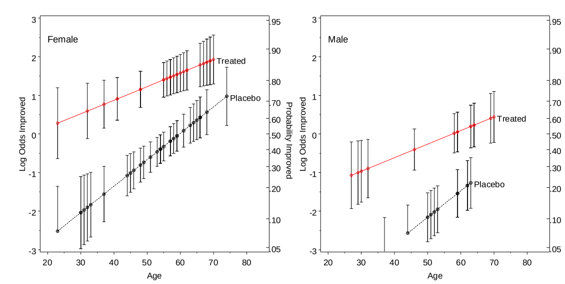
```
title ' '
   h=1.8 a=-90 'Probability Improved'    /* right axis label   */
   h=2.5 a=-90 ' ';                      /* extra space        */
goptions hby=0;                          /* suppress BY values */
proc gplot data=results;
   plot logit * age = treat /
         vaxis=axis1 haxis=axis2 hm=1 vm=1
         nolegend anno=bars frame;
   by sex;
   axis1 label=(a=90 'Log Odds Improved')
         order=(-3 to 3);
   axis2 order=(20 to 80 by 10) offset=(2,6);
   symbol1 v=+ i=join l=3 c=black;
   symbol2 v=- i=join l=1 c=red;
   label age='Age';
run;
```



Models with interactions

- **Plotting fitted values**
  - Only need to change the MODEL statement
  - Output dataset automatically incorporates all model terms
  - Plotting steps remain *exactly* the same

```
proc logistic data=arthrit descending;
  class sex (ref=last) treat (ref=first) / param=ref;
  model  better = sex  treat | age @2;
  output out=results p=prob l=lower u=upper
         xbeta=logit stdxbeta=selogit / alpha=.33;
```



Effect plots for generalized linear models

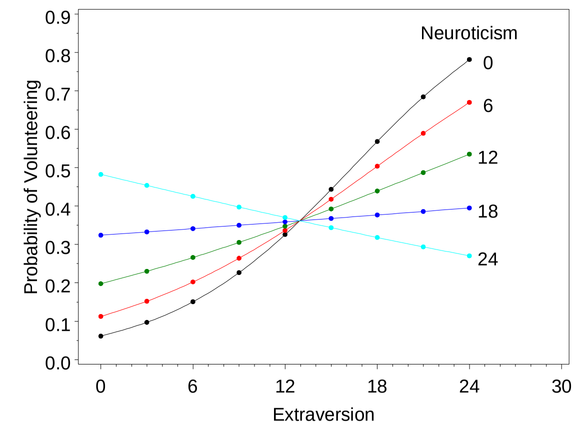
- Fox (1987)— For complex models, often wish to plot a specific main effect or interaction (including lower-order relatives)
  - Fit full model to data with linear predictor (e.g., logit) $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ and link function $g(\boldsymbol{\mu}) = \boldsymbol{\eta} \rightarrow$ estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$ and covariance matrix $\widehat{V(\boldsymbol{b})}$ of $\boldsymbol{b}$.
  - Vary each predictor in the term over its' range
  - Fix other predictors at "typical" values (mean, median, proportion in the data)
  - $\longrightarrow$ "effect model matrix," $\boldsymbol{X}^*$
  - Calculate fitted effect values, $\hat{\boldsymbol{\eta}}^* = \boldsymbol{X}^*\boldsymbol{b}$.
  - Standard errors are square roots of $\mathrm{diag}(\boldsymbol{X}^*\widehat{V(\boldsymbol{b})}\boldsymbol{X}^{*\mathsf{T}})$
  - Plot $\hat{\boldsymbol{\eta}}^*$, or values transformed back to scale of response, $g^{-1}(\hat{\boldsymbol{\eta}}^*)$.
- *Note*: This provides a general means to visualize interactions in *all* linear and generalized linear models.

Effect plots in SAS

- Create a grid of values for predictors in the effect (EXPGRID macro)
- Fix other predictors at "typical" values (mean, median, proportion in the data)
- Concatenate grid with data
- Fit model $\longrightarrow$ output data set $\longrightarrow$ fitted values in the grid
- Standard errors automatically calculated
- Plot fitted values in the grid
- (Not yet a macro)

Effect plots: Example

- Cowles and Davis (1987)— Volunteering for a psychology experiment
  - Predictors: Sex, Neuroticism, Extraversion
  - $\longrightarrow$ strong interaction, Neuroticism $\times$ Extraversion



Effect plots: Example

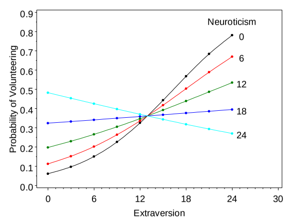
cowles3.sas

```
%include catdata(cowles);
%expgrid(Sex=0.5,              /* Fix Sex at mid value */
   Extraver=0 to 24 by 3,      /* range of predictors  */
   Neurot=0 to 24 by 6,        /*    "          "      */
   _in_=1);                    /* grid select value  */

*-- catenate grid values with data;
data cowles;
   set cowles _grid_;

*-- fit model, output fitted values;
proc logistic data=cowles outest=parm covout;
   model Volunter = Sex Extraver | Neurot  / covb;
   output out=predicted xbeta=logit stdxbeta=selogit
      p=prob u=upper l=lower / alpha=.33;

*-- select grid, replace labels;
data effect;
   set predicted (where =(_in_=1));
   label logit='log odds of Volunteering'
      prob = 'Probability of Volunteering';
```

Effect plots: Example

··· cowles3.sas

```
*-- Custom legend;
%label(data=effect,
   x=Extraver, y=prob,
   subset=Extraver=24,       /* at last.Extraver */
   text=put(Neurot,3.),      /* label text       */
   pos=6, xoff=.2, out=labels);

*-- Plot step;
proc gplot data=effect;
   plot prob * Extraver = Neurot /
      vaxis=axis1 haxis=axis2 vm=1
      anno=labels nolegend;
   symbol v=dot i=spline r=5;
   axis1 label=(a=90 r=0) order=(0 to .9 by .1);
   axis2 order=(0 to 30 by 6) offset=(3,1);
run; quit;
```



Influence measures and diagnostic plots

- **Leverage**: Potential impact of an individual case $\sim$ distance from the centroid in space of predictors
- **Residuals**: Which observations are poorly fitted?
- **Influence**: Actual impact of an individual case $\sim$ leverage $\times$ residual
  - ***C***, ***CBAR*** – analogs of Cook's D in OLS $\sim$ standardized change in regression coefficients when i-th case is deleted.
  - ***DIFCHISQ***, ***DIFDEV*** – $\Delta\chi^2$ when i-th case is deleted.

Influence measures and diagnostic plots

PROC LOGISTIC provides printed output with the influence and iplots options

```
proc logistic data=arthrit ;
   model better = sex treat age / influence iplots;
```

```
                    The LOGISTIC Procedure
                    Regression Diagnostics
          Deviance Residual                   Hat Matrix Diagonal

Case                (1 unit = 0.26)                (1 unit = 0.01)
Number   Value     -8  -4  0 2 4 6 8     Value    0 2 4 6 8  12  16

   1    1.812  | *      |        |    0.089  |          *      |
   2    0.360  |        |*       |    0.031  |    *            |
   3    0.685  |        |  *     |    0.087  |          *      |
   4    0.425  |        | *      |    0.034  |    *            |
   5    0.700  |        |  *     |    0.086  |          *      |
   6    0.488  |        | *      |    0.038  |    *            |
   7    1.703  |  *     |        |    0.084  |         *       |
   8    0.499  |        | *      |    0.039  |    *            |
   9    1.396  |   *    |        |    0.066  |        *        |
  10    0.511  |        | *      |    0.040  |     *           |
  11    1.142  |    *   |        |    0.064  |       *         |
  12    0.523  |        | *      |    0.041  |     *           |
  13    1.234  |        |    *   |    0.065  |       *         |
  14    0.599  |        | *      |    0.051  |      *          |
  15    1.121  |    *   |        |    0.065  |       *         |
  16    0.599  |        | *      |    0.051  |      *          |
```



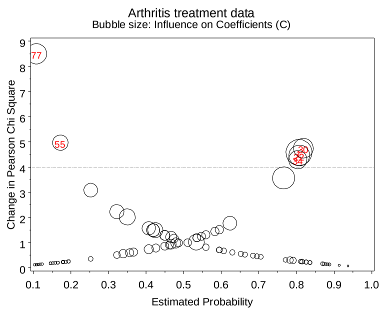
```
  17    1.319  |        |    *   |    0.069  |        *        |
  18    0.640  |        | *      |    0.058  |       *         |
  19    1.319  |        |    *   |    0.069  |        *        |
  20    0.640  |        | *      |    0.058  |       *         |
  21    1.340  |        |    *   |    0.070  |        *        |
  22    1.814  | *      |        |    0.061  |       *         |
  23    1.022  |    *   |        |    0.070  |        *        |
  24    0.529  |        | *      |    0.060  |       *         |
  25    1.449  |        |    *   |    0.078  |         *       |
  26    0.619  |        | *      |    0.053  |      *          |
  27    0.909  |     *  |        |    0.080  |         *       |
 ...
```

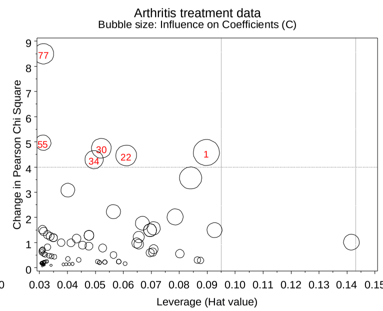
Problems:

- Way too much output
- Doesn't highlight unusual cases well
- Index plots don't consider combinations of measures

INFLOGIS macro

- Specialized version of INFLGLIM macro for logistic regression
- Plots a measure of change in χ^2 (DIFCHISQ or DIFDEV) vs. predicted probability or leverage.
- Bubble symbols show actual influence (C or CBAR)
- Shows standard cutoffs for “large” values
- Labels outlying cases





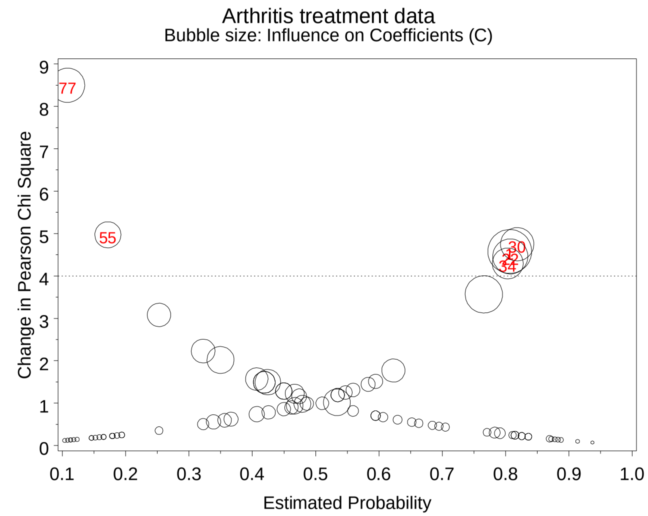
INFLOGIS macro: Example

```
%include data(arthrit);
%inflogis(data=arthrit,
   class=sex treat,      /* CLASS variables */
   y=better,             /* response        */
   x=sex treat age,      /* predictors      */
   id=case,              /* case ID         */
   gy=DIFCHISQ,          /* graph ordinate  */
   gx=PRED HAT);         /* graph abscissas */
```

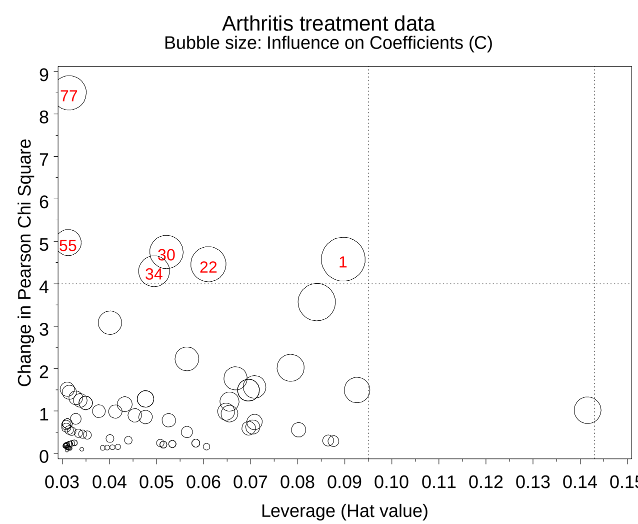
Printed output lists cases with “large” leverage, residual or influence:

case	better	sex	treat	age	pred	hat	difchisq	difdev	c
1	1	Male	Treated	27	.806	.09	4.578	3.695	0.451
22	1	Male	Placebo	63	.807	.06	4.460	3.565	0.290
30	1	Female	Placebo	31	.818	.05	4.749	3.657	0.261
34	1	Female	Placebo	33	.803	.05	4.296	3.464	0.224
55	0	Female	Treated	58	.172	.03	4.970	3.676	0.160
77	0	Female	Treated	69	.108	.03	8.498	4.712	0.276

INFLOGIS macro: Example



INFLOGIS macro: Example



Conclusions

- **Summarization & exposure**
  - Effective data analysis requires *summarization*— hypothesis tests, model fits (& comparisons!), parameter estimates (& precision!)
  - Also requires *exposure*— displays to help the viewer see (& understand!) patterns, trends, and anomalies.
- **Graphical methods for categorical data**
  - Many new methods developed over the last 10–15 years
  - Some novel, others extend familiar methods for quantitative data
  - Described and illustrated in *VCD*
- **Theory into practice**
  - To be useful, statistical methods must be:
    - available— implemented in standard software
    - accessible— easy to use (or at least easier)
  - *VCD* provides $\sim$ 40 general macros and SAS/IML programs

References

Atkinson, A. C. Two graphical displays for outlying and influential observations in regression. *Biometrika*, 68:13–20, 1981.

Bickel, P. J., Hammel, J. W., and O'Connell, J. W. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187:398–403, 1975.

Cowles, M. and Davis, C. The subject matter of psychology: Volunteers. *British Journal of Social Psychology*, 26:97–102, 1987.

Fox, J. Effect displays for generalized linear models. In Clogg, C. C., editor, *Sociological Methodology, 1987*, pp. 347–361. Jossey-Bass, San Francisco, 1987.

Friendly, M. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89:190–200, 1994.

Friendly, M. Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, 8(3):373–395, 1999.

Friendly, M. Multidimensional arrays in SAS/IML. In *Proceedings of the SAS User's Group International Conference*, volume 25, pp. 1420–1427. SAS Institute, 2000.

Friendly, M. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(4):316–324, 2002.

Friendly, M. and Kwan, E. Effect ordering for data displays. *Computational Statistics and Data Analysis*, 37, 2002. In press.

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

Hoaglin, D. C. and Tukey, J. W. Checking the shape of discrete distributions. In Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors, *Exploring Data Tables, Trends and Shapes*, chapter 9. John Wiley and Sons, New York, 1985.

Ord, J. K. Graphical methods for a class of discrete distributions. *Journal of the Royal Statistical Society, Series A*, 130:232–238, 1967.

Tufte, E. R. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983.

Tukey, J. W. Some graphic and semigraphic displays. In Bancroft, T. A., editor, *Statistical Papers in Honor of George W. Snedecor*. Iowa State University Press, Ames, IA, 1972.

Tukey, J. W. *Exploratory Data Analysis*. Addison Wesley, Reading, MA, 1977.

Williams, D. A. Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics*, 36:181–191, 1987.