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

Graphical Methods for Categorical Data

- 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?

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

<table>
<thead>
<tr>
<th>Hair color</th>
<th>Eye color</th>
<th>Count</th>
<th>Count</th>
<th>Count</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>Brown</td>
<td>68</td>
<td>119</td>
<td>26</td>
<td>7</td>
</tr>
<tr>
<td>Hazel</td>
<td>Brown</td>
<td>15</td>
<td>54</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>Green</td>
<td></td>
<td>5</td>
<td>29</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>Blue</td>
<td></td>
<td>20</td>
<td>84</td>
<td>17</td>
<td>94</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Hair color</th>
<th>Eye color</th>
<th>Count</th>
<th>Count</th>
<th>Count</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond</td>
<td>Black</td>
<td>94</td>
<td>20</td>
<td>17</td>
<td>84</td>
</tr>
<tr>
<td>Hazel</td>
<td>Brown</td>
<td>7</td>
<td>68</td>
<td>26</td>
<td>119</td>
</tr>
<tr>
<td>Green</td>
<td></td>
<td>10</td>
<td>54</td>
<td>14</td>
<td>29</td>
</tr>
<tr>
<td>Blue</td>
<td></td>
<td>16</td>
<td>5</td>
<td>14</td>
<td>29</td>
</tr>
</tbody>
</table>

Model: Independence ($\text{Hair} \times \text{Eye}$)

\[2 \chi^2(9) = 138.29\]

Color coding:

- $< -4$
- $< -2$
- $< -1$
- $0$
- $1$
- $2$
- $4$

$n$ in each cell:

- $n < \text{expected}$
- $n > \text{expected}$

Comparisons—make visual comparisons easy

Small multiples (Tufte, 1983)—combine stratified graphs into coherent displays

Baselines—compare data to model against a line, preferably horizontal

Visual grouping—connect with lines, make key comparisons contiguous

Model:

<table>
<thead>
<tr>
<th>Eye color</th>
<th>Hair color</th>
<th>Count</th>
<th>Count</th>
<th>Count</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hazel</td>
<td>Brown</td>
<td>16</td>
<td>5</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>Green</td>
<td>Brown</td>
<td>10</td>
<td>15</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>Brown</td>
<td></td>
<td>7</td>
<td>68</td>
<td>26</td>
<td>119</td>
</tr>
<tr>
<td>Blue</td>
<td></td>
<td>94</td>
<td>20</td>
<td>17</td>
<td>94</td>
</tr>
</tbody>
</table>

Principles of Graphical Displays

- Effect ordering and high-lighting for tables (Friendly, 2000)
Categorical Data Analysis with Graphics

Exploratory methods

- Minimal assumptions (like non-parametric methods)
- Show the data, not just summaries
- Help detect patterns, trends, anomalies, suggest hypotheses

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

- Make these methods available and accessible in SAS

Power calculations

- POWERLOG: Power calculations for logistic regression
- POWER2x2: Power calculations for a 2x2 table
- POWERX: Power calculations for two-way frequency table

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

- BARS, EQUATE, GDISPLAY: Graphics utility macros

Visual metaphors

- Quantitative data: magnitude, position along an axis
- Frequency data: count, area

Sex: Male
Admit?: Yes
Sex: Female
Admit?: No

11981493
5571278

Fourfold display for 2x2 table

Model: (DeptGender)(Admit)

Mosaic plot

Goals

- Diagnostic plots - influence, violation of assumptions
- Effect plots - estimated probabilities of response or log odds
- Residual plots - departures from model, omitted terms...
- Model-based methods

- Help detect patterns, trends, anomalies, suggest hypotheses
- Show the data, not just summaries

Poissonness plot

Hanging rootograms

Observer agreement chart

Plotting PROC CORRESP results

Fourfold displays for 2x2 tables (macro)

Mosaic displays (SAS/IML)

Sieve diagrams (SAS/IML)

SAS/IML modules for mosaic displays

Mosaic matrices (macro)

Construct a grouped frequency table, with recoding

Trilinear plots for 3-way tables

Visual metaphors

- MACROS and SAS/IML programs

VCD and SSG - Make these methods available and accessible in SAS

Practical power = Statistical power

Probability of Use

Today's goal: take-home knowledge

Tomorrow's goal: dynamic, interactive graphics for categorical data
Discrete distributions

Questions:
1. What process gave rise to the distribution?
2. Form of distribution: uniform, binomial, Poisson, negative binomial, geometric, etc.?
3. Estimate parameters
4. Visualize goodness of fit

For example:
- Federalist Papers: might expect a Poisson(\lambda) distribution.
- Families in Saxony: might expect a Bin(n; p) distribution with n = \# of papers, p = 0.5 as well.

Example:
- In text analysis, might expect a binomial distribution with \# of papers: n = \# of papers.

Example cases:
- Federalist Papers: might expect a Poisson(\lambda) distribution.
- Families in Saxony: might expect a Bin(n; p) distribution with n = \# of papers, perhaps p = 0.5 as well.

Example:
- In text analysis, might expect a binomial distribution with \# of papers: n = \# of papers.

Example cases:
- Federalist Papers: might expect a Poisson(\lambda) distribution.
- Families in Saxony: might expect a Bin(n; p) distribution with n = \# of papers, perhaps p = 0.5 as well.

Example:
- In text analysis, might expect a binomial distribution with \# of papers: n = \# of papers.

Example cases:
- Federalist Papers: might expect a Poisson(\lambda) distribution.
- Families in Saxony: might expect a Bin(n; p) distribution with n = \# of papers, perhaps p = 0.5 as well.

Example:
- In text analysis, might expect a binomial distribution with \# of papers: n = \# of papers.

Example cases:
- Federalist Papers: might expect a Poisson(\lambda) distribution.
- Families in Saxony: might expect a Bin(n; p) distribution with n = \# of papers, perhaps p = 0.5 as well.

Example:
- In text analysis, might expect a binomial distribution with \# of papers: n = \# of papers.

Example cases:
- Federalist Papers: might expect a Poisson(\lambda) distribution.
- Families in Saxony: might expect a Bin(n; p) distribution with n = \# of papers, perhaps p = 0.5 as well.

Example:
- In text analysis, might expect a binomial distribution with \# of papers: n = \# of papers.

Example cases:
- Federalist Papers: might expect a Poisson(\lambda) distribution.
- Families in Saxony: might expect a Bin(n; p) distribution with n = \# of papers, perhaps p = 0.5 as well.

Example:
- In text analysis, might expect a binomial distribution with \# of papers: n = \# of papers.

Example cases:
- Federalist Papers: might expect a Poisson(\lambda) distribution.
- Families in Saxony: might expect a Bin(n; p) distribution with n = \# of papers, perhaps p = 0.5 as well.

Example:
- In text analysis, might expect a binomial distribution with \# of papers: n = \# of papers.

Example cases:
- Federalist Papers: might expect a Poisson(\lambda) distribution.
- Families in Saxony: might expect a Bin(n; p) distribution with n = \# of papers, perhaps p = 0.5 as well.

Example:
- In text analysis, might expect a binomial distribution with \# of papers: n = \# of papers.

Example cases:
- Federalist Papers: might expect a Poisson(\lambda) distribution.
- Families in Saxony: might expect a Bin(n; p) distribution with n = \# of papers, perhaps p = 0.5 as well.

Example:
- In text analysis, might expect a binomial distribution with \# of papers: n = \# of papers.
Categorical Data Analysis with Graphics

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 = 25.24312131
Prob > chi-square = 0.0001250511

Degrees of freedom = 5

The poisson model does not fit! Why?

The GOODFIT macro gives a table of observed and fitted frequencies, Pearson X2:

Instances of 'may' in Federalist papers

COUNT BLOCKS PHAT EXP CHI DEV

<table>
<thead>
<tr>
<th>COUNT</th>
<th>BLOCKS</th>
<th>PHAT</th>
<th>EXP</th>
<th>CHI</th>
<th>DEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>156</td>
<td>0.51867</td>
<td>135.891</td>
<td>1.72499</td>
<td>6.56171</td>
</tr>
<tr>
<td>1</td>
<td>63</td>
<td>0.34050</td>
<td>89.211</td>
<td>-2.77509</td>
<td>-6.62056</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>0.11177</td>
<td>29.283</td>
<td>-0.05231</td>
<td>-0.75056</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>0.02446</td>
<td>6.408</td>
<td>0.62890</td>
<td>1.88423</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.00401</td>
<td>1.052</td>
<td>2.87493</td>
<td>3.26912</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.00053</td>
<td>0.138</td>
<td>2.31948</td>
<td>1.98992</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0.00006</td>
<td>0.015</td>
<td>8.01267</td>
<td>2.89568</td>
</tr>
</tbody>
</table>

Degrees of freedom = 5

The poisson model does not fit! Why?

Hang & root them!

Hanging rootograms

Tukey (1972, 1977):
shift histogram bars to the fitted curve
judge deviations vs. horizontal line.

plot p freq
smaller frequencies are emphasized.

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

The goodfit macro gives a table of observed and fitted frequencies, Pearson X2.

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 & rootograms

Sqrt(frequency)
-2.0 0.0 2.0 4.0 6.0
0 20 40 60 80 100 120 140
160
Number of Occurrences
0 1 2 3 4 5 6

Hang & root them — Hanging rootograms

Filling discrete distributions

Hang & root them — Hanging rootograms

What's wrong with histograms?
ORDPLOT macro

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

Diagnoses distribution as NegBin
Estimates \( \hat{p} = 0.576 \):

\[
\text{slope} = 0.424, \quad \text{intercept} = -0.023
\]

type: Negative binomial
parm: \( p = 0.576 \)

Instances of 'may' in Federalist papers

Frequency Ratio, \( \frac{k \cdot n(k)}{n(k-1)} \)

<table>
<thead>
<tr>
<th>Occurrences of 'may'</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 2 3 4 5 6</td>
</tr>
</tbody>
</table>

SUGI 2820

Michael Friendly

Categorical Data Analysis with Graphics

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 \( kp \) against \( k \) is linear

signs of intercept and slope give rough estimates of

\( \hat{p} \) and \( \hat{\lambda} \)

For Poisson, plot count metameter \( \log_e \left( \frac{k! \cdot n(k)}{N} \right) \) vs. \( k \)

Linear relation

Poisson, slope gives \( ^{\hat{\lambda}} \)

CI for points, diagnostic (influence) plot

POISPLOT macro

Robust distribution plots: Poisson

Ord plots lack robustness

one discrepant frequency, \( n_k \) affects points for both \( k \) and \( k + 1 \)

Robust plots for Poisson distribution (Hoaglin and Tukey, 1985)

For Poisson, plot count metameter \( \log_e \left( \frac{k! \cdot n(k)}{N} \right) \) vs. \( k \)

Linear relation

Poisson, slope gives \( ^{\hat{\lambda}} \)

CI for points, diagnostic (influence) plot

POISPLOT macro

Butterfly species collected in Malaya

Frequency Ratio, \( \frac{k \cdot n(k)}{n(k-1)} \)

<table>
<thead>
<tr>
<th>Number collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 10 20 30</td>
</tr>
</tbody>
</table>

Logarithmic series

\( \text{slope} = -0.657, \quad \text{intercept} = 10.946 \)

\( \text{type}: \text{Binomial} \)

parm: \( p = 0.396 \)

Ord plot: Families in Saxony

Frequency Ratio, \( \frac{k \cdot n(k)}{n(k-1)} \)

<table>
<thead>
<tr>
<th>Number of males</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
</tbody>
</table>

Binomial

SUGI 2821

Michael Friendly

Categorical Data Analysis with Graphics

Ord plots

Other distributions:

\( \text{slope} = 1.061, \quad \text{intercept} = -0.709 \)

\( \text{type}: \text{Logarithmic series} \)

parm: \( \theta = 1.061 \)

\( \text{Butterfly species collected in Malaya} \)

\( \text{Frequency Ratio, } \frac{k \cdot n(k)}{n(k-1)} \)

<table>
<thead>
<tr>
<th>Number collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 10 20 30</td>
</tr>
</tbody>
</table>

Logarithmic series

\( \text{slope} = -0.657, \quad \text{intercept} = 10.946 \)

\( \text{type}: \text{Binomial} \)

parm: \( p = 0.396 \)

Ord plot: Families in Saxony

Frequency Ratio, \( \frac{k \cdot n(k)}{n(k-1)} \)

<table>
<thead>
<tr>
<th>Number of males</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
</tbody>
</table>

Binomial
Categorical Data Analysis with Graphics

POISPLOT: output

Curvilinear relation is not Poisson

slope = 0.228
intercept = -1.530

lambda: mean = 0.656
exp(slope) = 1.256

Instances of 'may' in Federalist papers

Poisson metameter, ln(k! / n(k) / N)

-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10

Number of Occurrences

0 1 2 3 4 5 6

Influence plot for change in

Parameter change

Scaled Leverage

2 3 4 5 6 7 8 9 10

Correspondence analysis and MCA — (CORRESP macro)

Contingency tables

two-way tables — Visualize odds ratio (FFOLD macro)
k tables — Homogeneity of association
r c tables — Trilinear plots (TRIPLOT macro)
r c tables — Visualize association (SIEVE program)
Square R x I tables — Visualize agreement (GAUGE program)
R x C tables — Visualize association (MOSAIC macro)
r x c tables — Visualize association (SIEVE program)
r x c tables — Trilinear plots (TRIPLOT macro)
r x 2 x I tables — Hypothesis of association
2 x 2 x I tables — Visualize odds ratio (ROID macro)
r x 2 x 2 tables — Visualize agreement

Generalized robust distribution plots

DISTPLOT: macro: example

Title "Instances of 'may' in Federalist papers"

Data: madison, count=blocks, freq=count

Chi-square test of association, count=blocks, freq=count

Distributed macro:

???

Other distributions: Analogous plots, for suitable count metameter, (n/k) vs. k.

Generalized robust distribution plots

DISTPLOT: example

Title "Instances of 'may' in Federalist papers"

Data: madison, count=blocks, freq=count

Chi-square test of association, count=blocks, freq=count

Distributed macro:

???
Categorical Data Analysis with Graphics

Standard analysis:

PROC FREQ

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

Output:

<table>
<thead>
<tr>
<th>Statistics</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>1</td>
<td>92.2053</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>1</td>
<td>93.4494</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Continuity Adj. Chi-Square</td>
<td>1</td>
<td>91.6096</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>92.1849</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Phi Coefficient</td>
<td>0.1427</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How to visualize and interpret?

Methods for $2 \times 2$ tables

Bickel et al. (1975): data on admissions to graduate departments at U.C. Berkeley

Aggregate data for the six largest departments:

Table 3: Admissions to Berkeley graduate programs

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admit?</td>
<td>1198</td>
<td>557</td>
<td>1755</td>
</tr>
<tr>
<td>Yes</td>
<td>1493</td>
<td>1278</td>
<td>2771</td>
</tr>
</tbody>
</table>

Evidence for gender bias?

G$^2$ (1) = 93.27, 2(1) = 92.2; p < .0001

Odds ratio, $Odds(\text{Admit | Male}) / Odds(\text{Admit | Female}) = 1.84$

Males 84% more likely to be admitted.

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

Data in Table 3 had been pooled over departments

Shading: highlight departments for which $H_a: i \neq 0$.

Each $2 \times 2$ table standardized to equate marginal frequencies

Confidence rings: Visual test of $H_0: \theta = 0$

Confidence rings do not overlap: $\theta < 0$

Confidence rings: Visual test of $H_0: \theta = 1$

Confidence rings: Visual test of $H_0: \theta > 1$

Confidence rings: Visual test of $H_0: \theta < 1$

Confidence rings: Visual test of $H_0: \theta = 0$

Confidence rings: Visual test of $H_0: \theta = 1$

Confidence rings: Visual test of $H_0: \theta > 1$

Confidence rings: Visual test of $H_0: \theta < 1$

Confidence rings: Visual test of $H_0: \theta = 0$

Confidence rings: Visual test of $H_0: \theta = 1$

Confidence rings: Visual test of $H_0: \theta > 1$

Confidence rings: Visual test of $H_0: \theta < 1$

Confidence rings: Visual test of $H_0: \theta = 0$

Confidence rings: Visual test of $H_0: \theta = 1$

Confidence rings: Visual test of $H_0: \theta > 1$

Confidence rings: Visual test of $H_0: \theta < 1$

Confidence rings: Visual test of $H_0: \theta = 0$

Confidence rings: Visual test of $H_0: \theta = 1$

Confidence rings: Visual test of $H_0: \theta > 1$

Confidence rings: Visual test of $H_0: \theta < 1$

Confidence rings: Visual test of $H_0: \theta = 0$

Confidence rings: Visual test of $H_0: \theta = 1$

Confidence rings: Visual test of $H_0: \theta > 1$

Confidence rings: Visual test of $H_0: \theta < 1$

Confidence rings: Visual test of $H_0: \theta = 0$

Confidence rings: Visual test of $H_0: \theta = 1$

Confidence rings: Visual test of $H_0: \theta > 1$

Confidence rings: Visual test of $H_0: \theta < 1$

Confidence rings: Visual test of $H_0: \theta = 0$

Confidence rings: Visual test of $H_0: \theta = 1$

Confidence rings: Visual test of $H_0: \theta > 1$

Confidence rings: Visual test of $H_0: \theta < 1$

Confidence rings: Visual test of $H_0: \theta = 0$

Confidence rings: Visual test of $H_0: \theta = 1$

Confidence rings: Visual test of $H_0: \theta > 1$

Confidence rings: Visual test of $H_0: \theta < 1$

Confidence rings: Visual test of $H_0: \theta = 0$

Confidence rings: Visual test of $H_0: \theta = 1$

Confidence rings: Visual test of $H_0: \theta > 1$

Confidence rings: Visual test of $H_0: \theta < 1$
Categorical Data Analysis with Graphics

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).

```
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);
```

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.

Sieve diagrams

When row/col variables are independent, each cell can be represented as a rectangle, with area = height \times width.

When row/col variables are independent, height/width = marginal frequencies, n_i + n_j

Area = expected frequency, n_i + n_j

Shading = observed frequency, n_ij, color: sign (n_ij - n_i + n_j)

Independence: Shown when density of shading is uniform.
Mosaic displays and Log-linear Models

Sieve diagrams: Example

```sas
proc iml;
%include iml(sieve);
*-- frequency table;
tab = f1520 266 124 66, 234 1512 432 78, 117 362 1772 205, 36 82 179 492;
*-- variable and level names;
vnames = f'Right Eye Grade' 'Left Eye Grade';
lnames = f'High' '2' '3' 'Low','High' '2' '3' 'Low';
title = f'unaided distant vision data';
*-- Global options;
font='hwpsl011';
run sieve(tab, vnames, lnames, title );
quit;
```

Did departments differ in the total number of applicants?

Did men and women apply differentially to departments?

Model: \( (\text{Dept})(\text{Gender}) \)

G\(^2\) (5) = 1220.6.

Note: Departments ordered A–F by overall rate of admission.

---

**Mosaic displays**

**Shading**: Sign and magnitude of Pearson \( G^2 \) residuals,

\[ d_{ij} = (n_{ij} - m_{ij}^p)^2 / m_{ij}^p = G^2 (\text{L.R.}) \]

Sign:

- negative in red;
- positive in blue

Magnitude: intensity of shading:

- \( j d_{ij} j > 0 \);
- \( 2, 4, \ldots \); : : : 

Independence: Rows align, or cells are empty!

E.g., aggregate Berkeley data, independence model:

<table>
<thead>
<tr>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admitted</td>
<td>Rejected</td>
</tr>
</tbody>
</table>

Model: \( (\text{Gender})(\text{Admit}) \)

Standardized residuals:

\[ < -4 \]  \[ -4 : -2 \]  \[ -2 : -0 \]  \[ 0 : 2 \]  \[ 2 : 4 \]  \[ > 4 \]

---

**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

<table>
<thead>
<tr>
<th>Model</th>
<th>Model Symbol</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independence</td>
<td>( [A][B][C] )</td>
<td>Mutual independence</td>
</tr>
<tr>
<td>Mutual independence</td>
<td>( {A} {B} {C} )</td>
<td>( A ? B ? C )</td>
</tr>
<tr>
<td>Conditional independence</td>
<td>( {AC} {BC} )</td>
<td>( A ? B \ j C )</td>
</tr>
<tr>
<td>All two-way associations</td>
<td>( {AB} {AC} {BC} )</td>
<td>(none)</td>
</tr>
<tr>
<td>Saturated model</td>
<td>( {ABC} )</td>
<td>(none)</td>
</tr>
</tbody>
</table>

E.g., the model for conditional independence (\( A \? C \ j B \)):

\[ \log m_{ijk} = \alpha_{AC} + \beta_{B} + \gamma_{C} + \delta_{AB} + \epsilon_{BC} \]

---

**SUGI 2841 Michael Friendly**

**SUGI 2842 Michael Friendly**

**SUGI 2840 Michael Friendly**

**SUGI 2843 Michael Friendly**
Categorical Data Analysis with Graphics

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

Visual fitting:
Pattern of lack-of-fit (residuals)

"better" model—smaller residuals

"cleaning the mosaic"

"better" model—empty cells

best done interactively!

Regions in mosaic

Table macro

Conditional independence

Model: (DeptGender)(DeptAdmit)

E.g., Add [Dept Admit] association

Conditional independence:

Fit: Poisson model (\(\phi = 2.74\))

Conditional independence: \(\phi = 2.74\)

E.g., Add [Dept Admit] association

SAS/IML modules:
mosaics.sas program

Examples:
Many in VCD and on web site

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 (Jordana) — http://www.math.yorku.ca/SCS/Online/jvcd/

Autostereo display applet from the graphical modules from VCD.

Java module displaying mosaic tables (two-way).

Other implementations:

Macro interface:
mosaic macro, table macro, mosmat macro

Easiest to use:
Direct input from a SAS dataset

No knowledge of SAS/IML required

Re-order table variables; collapse, reorder table levels with table macro

Convenient interface to partial mosaics (\(BY=\))

Create frequency table from raw data

Conditional independence to partial mosaics (\(BY=\))

Prefer these techniques using SAS/IML (\(BY=\))

Create frequency table from raw data

Differentiate table variables, collapse, reorder levels with table macro

Mosaic macro:

Macro interface.

Mosaic macro, table macro, mosaic macro

Software for Mosaic Displays

Mosaic displays for multiway tables
### MOSAIC MACRO EXAMPLE: BERKELEY DATA

**Title:**

Title: Berkeley Admissions data

**Data Source:**

- **Data set:** Berkeley
- **Variables:** Admit, Gender, Dept
- **Frequency table:**

<table>
<thead>
<tr>
<th>Gender</th>
<th>Dept</th>
<th>Accept</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>A</td>
<td>512</td>
<td>313</td>
</tr>
<tr>
<td>Female</td>
<td>A</td>
<td>89</td>
<td>19</td>
</tr>
<tr>
<td>Male</td>
<td>B</td>
<td>353</td>
<td>207</td>
</tr>
<tr>
<td>Female</td>
<td>B</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>Male</td>
<td>C</td>
<td>120</td>
<td>205</td>
</tr>
<tr>
<td>Female</td>
<td>C</td>
<td>202</td>
<td>391</td>
</tr>
<tr>
<td>Male</td>
<td>D</td>
<td>138</td>
<td>279</td>
</tr>
<tr>
<td>Female</td>
<td>D</td>
<td>131</td>
<td>244</td>
</tr>
<tr>
<td>Male</td>
<td>E</td>
<td>53</td>
<td>138</td>
</tr>
<tr>
<td>Female</td>
<td>E</td>
<td>94</td>
<td>299</td>
</tr>
<tr>
<td>Male</td>
<td>F</td>
<td>22</td>
<td>351</td>
</tr>
<tr>
<td>Female</td>
<td>F</td>
<td>24</td>
<td>317</td>
</tr>
</tbody>
</table>

**SAS/IML Example:**

```sas
proc iml;
*-- Berkeley Admissions data;
dim = f 2 2 6 g;
vnames = f "Admit" "Gender" "Dept" g;
/* var names */
lnames = f "Admitted" "Rejected" " " " " " " "Male" "Female" " " " " " "A" "B" "C" "D" "E" "F" g;
/* level names */
table = f 512 313, 89 19, 353 207, 17 8, 120 205, 202 391, 138 279, 131 244, 53 138, 94 299, 22 351, 24 317 g;
reset storage=mosaic.mosaic;
load module=_all_;" title = 'Model: (Dept)(Gender)'; plots=2:3;
fittype='joint';run mosaic(dim, table, vnames, lnames, plots, title);quit;
```
Categorical Data Analysis with Graphics

Logit models

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

\[ \log \frac{m_{ijk}}{1 - m_{ijk}} = \theta_0 + \theta_1 A_i + \theta_2 D_j + \theta_3 G_k + \theta_4 AD_{ij} + \theta_5 DG_{jk} \]

\[ \log \frac{m_{ij}}{m_{ij}} = \theta_0 + \theta_1 D_i + \theta_j G_{ij} \]

where,

- \( \log \frac{m_{ij}}{m_{ij}} \): log odds of admission for males as vs. females,
- \( D_i \): effect on admissions of department,
- \( G_{ij} \): effect of gender in Dept. A.


Plots for logit models

Logit models

Fitting procedures

PROC CATMOD
PROC LOGISTIC
PROC GENMOD / dist=poisson
SAS/INSIGHT (Fit Y X) Options

Visualization procedures

HALFNORM macro - half-normal plot of residuals for generalized linear models
INFLUENCE macro - influence plots for generalized linear models
CATPLOT macro - plot predicted, observed log odds from CATMOD

SUGI 2853
Michael Friendly

Categorical Data Analysis with Graphics

mosmat
macro: Mosaic matrices

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

Male    Female
Admit   Reject
A       B       C       D       E       F
Male    Female
A       B       C       D       E       F
Male    Female
A       B       C       D       E       F

Department
A B C D E F

Logit(Admit) = Dept DeptA*Gender
Gender Female
Male

Log Odds (Admitted)
-3 -2 -1 0 1 2

Department
A B C D E F

SUGI 2852
Michael Friendly

Categorical Data Analysis with Graphics

Plots for logit models

Logit models

Fitting procedures

PROC CATMOD
PROC LOGISTIC
PROC GENMOD / dist=poisson
SAS/INSIGHT (Fit Y X) Options

Visualization procedures

HALFNORM macro - half-normal plot of residuals for generalized linear models
INFLUENCE macro - influence plots for generalized linear models
CATPLOT macro - plot predicted, observed log odds from CATMOD

SUGI 2855
Michael Friendly

Categorical Data Analysis with Graphics

Plots for logit models

Logit models

Fitting procedures

PROC CATMOD
PROC LOGISTIC
PROC GENMOD / dist=poisson
SAS/INSIGHT (Fit Y X) Options

Visualization procedures

HALFNORM macro - half-normal plot of residuals for generalized linear models
INFLUENCE macro - influence plots for generalized linear models
CATPLOT macro - plot predicted, observed log odds from CATMOD

SUGI 2854
Michael Friendly
Categorical Data Analysis with Graphics
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 | 71.45 | <.0001
depth A/G | 1 | 16.04 | <.0001
Likelihood Ratio | 5 | 2.68 | 0.7489

Analysis of Maximum Likelihood Estimates

Standard Chi-Square
Parameter Estimate | Error | Square | Pr > ChiSq
Intercept | -0.6685 | 0.0392 | 291.22 | <.0001
dept A | 1.1606 | 0.0705 | 271.21 | <.0001
depth B | 1.2113 | 0.0802 | 227.95 | <.0001
depth C | 0.0528 | 0.0687 | 0.59 | 0.4426
depth D | 0.00358 | 0.0727 | 0.00 | 0.9607
depth E | -0.4210 | 0.0871 | 23.34 | <.0001
depth F | 0.0521 | 0.2627 | 16.04 | <.0001

How to interpret?

SUGI 2856
Michael Friendly
Categorical Data Analysis with Graphics
Plots for logit models
Fit:
PROC CATMOD;
plot:
CATPLOT macro
Admit?
Gender
Dept, except for Dept. A
catberk6.sas
1
%include catdata(berkeley);
data berkeley;
set berkeley;
/*-- Dummy variable for Gender in Dept A;*/
depth1AG = (gender='F') * (dept=1);
format dept dept.;
8
proc catmod order=data
data=berkeley;
weight freq;
population dept gender;
direct dept1AG;
response / out=predict;
model admit = dept dept1AG / ml;
run;
...
SUGI 2859
Michael Friendly
Categorical Data Analysis with Graphics
Plots for logit models
PROC CATMOD: observed and predicted logits:
proc print data=predict;
id dept gender;
var _obs_ _pred_ _sepred_;format _numeric_ 6.3 dept dept.;where(_type_='FUNCTION');
depth gender 0.492 0.492 0.072 dept F 1.544 1.544 0.253 dept M 0.534 0.543 0.086 dept B 0.754 0.543 0.086 dept F 0.536 -0.616 0.069 dept C 0.660 -0.616 0.069 dept D -0.704 -0.665 0.075 dept D -0.622 -0.665 0.075 dept E -0.957 -1.090 0.095 dept E -1.157 -1.090 0.095 dept F -2.770 -2.676 0.152 dept F -2.581 -2.676 0.152
SUGI 2858
Michael Friendly
INFLGLIM macro: Example

```
Berkeley data, model
```

```
\[ L_{ij} = + \]
```

```
%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);
```

```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
```
Logistic regression models

Response variable:
- Binary response: success/failure, vote: yes/no
- Ordinal response: none, some, severe depression
- Polytomous response: vote Liberal, Tory, Alliance, NDP

Explanatory variables:
- Quantitative regressors: age, dose
- Transformed regressors: $\log(dose)$, $\log(age)$
- Polynomial regressors: $\log(age)^2$, $\log(age)^3$
- Categorical predictors: treatment, sex
- Interaction regressors: treatment $\times$ age, sex $\times$ age

Finding: PROC LOGISTIC (OR ROBUST macro— M-estimation)

Logistic regression models: Binary response

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

$$ \logit(i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p $$

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

$$ \logit(i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p $$

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

$$ \logit(i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p $$

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

$$ \logit(i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p $$

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

$$ \logit(i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p $$

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

```sas
options pageno=10000 lreen=0;
%include catdata(arthrit);
%logodds(data=arthrit,
x=age, y=Better,
smooth=0.5,
plot=logit);
```

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
- Fitted and smoothed curves
- Standard confidence intervals
- Smoothing parameter
- Plot on logit scale

Data in case form:

Consider first as binary response: None vs. (Some or Marked) = Better

Response: Improvement (none, some, marked)

Predictors: Sex, Treatment (treated, placebo), Age

Log Odds Better = 1

0 1 2 3
-3 -2 -1 0 1 2

AGE

20 30 40 50 60 70 80

Log Odds Better = 1

Log Odds Better = 1

Log Odds Better = 1
Parameter estimates (reference cell coding):

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>95% Wald Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex Female vs Male</td>
<td>4.427</td>
<td>(1.380, 14.204)</td>
</tr>
<tr>
<td>treat Treated vs Placebo</td>
<td>5.811</td>
<td>(2.031, 16.632)</td>
</tr>
<tr>
<td>age</td>
<td>1.050</td>
<td>(1.008, 1.093)</td>
</tr>
</tbody>
</table>

Females are 4.43 times more likely to be better than Males.

Treated patients are 5.81 times more likely to be better than Placebo.

The odds ratio is 1.05, indicating a 5% increase in the odds of improvement each year. Over 10 years, the odds of improvement are multiplied by $e^{10 \cdot 0.0487} = 1.63$, a 63% increase.

**PROC LOGISTIC**

```
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**

<table>
<thead>
<tr>
<th>Effect</th>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>1</td>
<td>6.2576</td>
<td>0.0124</td>
</tr>
<tr>
<td>treat</td>
<td>1</td>
<td>10.7596</td>
<td>0.0010</td>
</tr>
<tr>
<td>age</td>
<td>1</td>
<td>5.5655</td>
<td>0.0183</td>
</tr>
</tbody>
</table>

**PROC LOGISTIC**

```
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:

- Probability scales at right (PSCALE macro)
- Show 67% error bars (BARS macro)
- Custom legend and panel labels (LABEL macro)
- Placebo vs. Treated

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;
```

```glogist1c.sas``
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

Output data set
Fitted values in the grid
Standard errors automatically calculated
Plot fitted values in the grid
(Not yet a macro)

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)

= X and link function g ()

Estimate b of and covariance matrix of b.

V of b.

Standard errors are square roots of diag( X V(X) X T )

Plot , or values transformed back to scale of response, g1 ( )

Note: This provides a general means to visualize interactions in all linear and generalized linear models.

SUGI 2880
Michael Friendly

Categorical Data Analysis with Graphics

Effect plots: Example
Cowles and Davis (1987)— Volunteering for a psychology experiment
Predictors: Sex, Neuroticism, Extraversion!

Strong interaction, Neuroticism \times Extraversion

Extraversion

Neuroticism

Extraversion

0 6 12 18 24 30
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9

Probability of Volunteering

SUGI 2883
Michael Friendly
Influence measures and diagnostic plots

- **Leverage**: Potential impact of an individual case
- **Residuals**: How observations are poor fitted
- **Influence**: Actual impact of an individual case
  - leverage
  - residual

**C, CBAR** – analogs of Cook's D in OLS
  - standardized change in regression coefficients when i-th case is deleted.

**DIFCHISQ, DIFDEV** – $\chi^2$-s when i-th case is deleted.

**Effect plots: Example**

cowles3.sas

```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;
```

**Problems:**

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

**PROC LOGISTIC** provides printed output with the influence and iplots options.

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

**IPLots**

- Deviance Residual
- Hat Matrix Diagonal

<table>
<thead>
<tr>
<th>Number</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
<th>Value 5</th>
<th>Value 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.812</td>
<td>*</td>
<td>*</td>
<td>0.089</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.360</td>
<td>*</td>
<td>*</td>
<td>0.031</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.685</td>
<td>*</td>
<td>*</td>
<td>0.087</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.425</td>
<td>*</td>
<td>*</td>
<td>0.034</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.700</td>
<td>*</td>
<td>*</td>
<td>0.086</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.488</td>
<td>*</td>
<td>*</td>
<td>0.038</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.703</td>
<td>*</td>
<td>*</td>
<td>0.084</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.499</td>
<td>*</td>
<td>*</td>
<td>0.039</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1.396</td>
<td>*</td>
<td>*</td>
<td>0.066</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.511</td>
<td>*</td>
<td>*</td>
<td>0.040</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1.142</td>
<td>*</td>
<td>*</td>
<td>0.064</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.523</td>
<td>*</td>
<td>*</td>
<td>0.041</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1.234</td>
<td>*</td>
<td>*</td>
<td>0.065</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.599</td>
<td>*</td>
<td>*</td>
<td>0.051</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1.121</td>
<td>*</td>
<td>*</td>
<td>0.065</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.599</td>
<td>*</td>
<td>*</td>
<td>0.051</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>
```

**Effect plots: Example**

```sas
/* Custom legend:
proc plot 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;
```
%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

Specialized version of INFLGLIM macro for logistic regression
Plots a measure of change in $\chi^2$ (DIFCHISQ or DIFDEV) vs. predicted probability
Bubble symbols show actual influence (C or CBAR)
Shows standard cutoffs for "large" values
Labels outlying cases with large leverage, residual or influence.

Arthritis Treatment data
Bubble size: Influence on Coefficients (C)
Change in Pearson Chi Square
Estimated Probability
Leverage (Hat Value)
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. Described and illustrated in VCD.

Many new methods developed over the last 10–15 years. Described and illustrated in VCD.

Graphical methods for categorical data

VCD provides a general platform for SAS/IML programmers.

Graphical methods for categorical data

VCD provides a general platform for SAS/IML programmers.