A Reader's Guide to Visualizing Categorical Data

1 author:

Michael Friendly
York University
73 PUBLICATIONS **1,917** CITATIONS

SEE PROFILE

A Reader's Guide to *Visualizing Categorical Data*

Michael Friendly
York University, friendly@yorku.ca

Abstract

Visualizing Categorical Data presents a comprehensive overview of graphical methods for discrete data— count data, cross-tabulated frequency tables, and discrete response data. These methods are designed to complement traditional numerical summaries and statistical models, expose patterns in the data, and to aid in diagnosing model defects. They are illustrated with real data problems, and implemented in a large collection of SAS macro programs available with the book.

A number of these methods are somewhat novel, and the macro programs may take some effort to use effectively with your own data. In this paper I present an overview of these methods and illustrate the use of the macro programs for graphic analysis to reveal features of the data not apparent in traditional numerical summaries. The goal is to translate "theory into practice," and enable readers to use these techniques productively with their own data.

KEYWORDS: categorical data, graphics, mosaic displays, mosaic matrices, correspondence analysis, diagnostic plots, macros, loglinear models, logistic regression.

1 Introduction

Over the last decade a modest revolution has been brewing in the analysis of categorical data, as graphical methods and techniques of data visualization, so commonly used for quantitative data, have begun to be developed for frequency data and discrete data.

Visualizing Categorical Data (Friendly, 2000) completes the initial steps reported at SUGI 17 (Friendly, 1992). It presents a comprehensive overview of graphical methods for discrete data— count data, cross-tabulated frequency tables, and discrete response data. These methods are designed to complement traditional numerical summaries and statistical models, expose patterns in the data, and to aid in diagnosing model defects. They are illustrated with real data problems, and implemented in a collection of nearly 40 general macros and programs (see Appendix A) available with the book.

A number of these methods are somewhat novel, and the macro programs, while flexible, and easy to use may take some effort to use effectively with your own data. In this paper I present an overview of these methods and illustrate the use of the macro programs for graphic analysis to reveal features of the data not apparent in traditional numerical summaries. The goal is to translate "theory into practice," and enable readers to use these techniques productively with their own data. (Most of the graphs are in color; see the CD version of the Proceedings.)

2 Discrete distributions

Discrete frequency distributions often involve counts of occurrences such as accident fatalities, words in passages of text, births of twins, events of terrorism or suicide, or blood cells with some characteristic. Typically such data consist of a table which records that n_k of the observations pertain to the basic outcome value k , $k = 0, 1, \ldots$.

For such data, we often wish to understand the process which gives rise to these numbers, or to estimate frequencies for outcome values k we did not observe. Both goals can be approached by examining how closely the data follow a particular discrete probability distribution, such as the Poisson, the binomial, or the geometric distribution.

Chapter 2: "Fitting and graphing discrete distributions" describes the properties of some of the most widely used discrete distributions (the binomial, Poisson, negative binomial, log-series and geometric), along with SAS techniques for calculating and visualizing those distributions, as illustrated in Table 1 In some cases, we may not know *which* discrete distribution should be fit to a given dataset; we describe simple graphical methods designed to determine an appropriate distribution type. A number of SAS macros to simplify the fitting and graphing of discrete distributions are illustrated throughout the chapter.

For each of the main distributions, the `GOODFIT` macro estimates parameters, and calculates fitted frequencies and goodness-of-fit tests. The `ROOTGRAM` macro produces a variety of graphs, and the `DISTPLOT` macro provides robust distribution and influence plots.

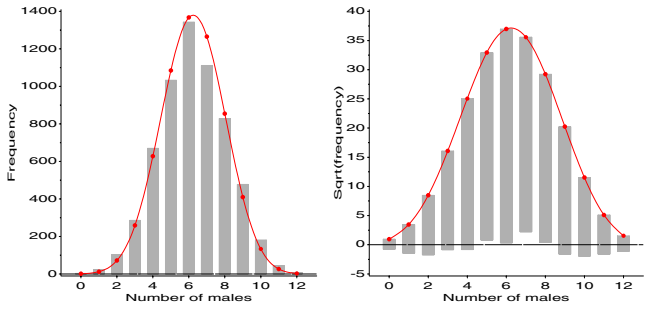

Figure 1: Saxony data, with Binomial fit. Left: histogram; right: hanging rootogram

We concentrate here on visualization methods for binomial distributions. For example, Geissler tabulated a huge dataset on sex distributions in families in Saxony in the 19th century. Included were $N = 6115$ families with $n = 12$ children, which might reasonably be expected to follow a Bin(12,p) distribution. The data are input and fit, using the `GOODFIT` macro as shown below.

Table 1: Some tasks for discrete distributions

Topic	Task	Program	Examples
Binomial distribution	Calculate, plot		§2.2.1
Poisson distribution	Calculate, plot		§2.2.2
”	Robust diagnosis	POISPLOT	2.14
”	Leverage, influence plots	POISPLOT	2.14
Neg. Bin. distribution	Calculate, plot		§2.2.3
Discrete distributions	Fit model	GOODFIT	2.7, 2.8
”	Fit, as loglinear model	Genmod	2.11
”	Estimate model parameters	GOODFIT	2.7, 2.8
”	Plot observed, fitted frequencies	ROOTGRAM	2.9, 2.10
”	Diagnose model form	ORDPLOT	2.12, 2.13
”	Robust diagnosis	DISTPLOT	2.15

```
title 'Number of males in 6115 families in Saxony';
data saxony;
  do males = 0 to 12;
    input families @;
    output;
    end;
  label males='Number of males'
     families='Number of families';
cards;
3 24 104 286 670 1033 1343 1112 829 478 181 45 7
;
%goodfit(data=saxony,var=males,freq=families,dist=binomial);
```

2.1 Hanging rootograms

Discrete frequency distributions are often graphed as histograms, with a theoretical fitted distribution superimposed. Figure 1 (left), for example, shows the data together with the fitted frequencies under a Binomial model. It is hard to compare the observed and fitted frequencies visually, because (a) we must assess deviations against a curvilinear relation, and (b) the largest frequencies dominate the display.

The hanging rootogram (Tukey, 1977) solves these problems by (a) shifting the histogram bars to coincide with the fitted curve, so that deviations may be judged by deviations from a horizontal line, and (b) plotting on a square-root scale, so that smaller frequencies are emphasized. Figure 1 (right) shows more clearly that the observed frequencies differ systematically from those predicted under a Binomial model. The `ROOTGRAM` macro produces a variety of displays like those in Figure 1 For example, the right panel is produced with the statement

```
%rootgram(data=fit, var=males, obs=families, exp=exp);
```

2.2 Robust distribution plots

While χ^2 tests provide an overall measure of goodness-of-fit, they do not reveal the departure is systematic or confined to a single discrepant frequency. Robust distribution plots, following methods described by Hoaglin and Tukey (1985), are provided by the `DISTPLOT` macro.

Figure 2 shows the Binomial distribution plot, produced using the `DISTPLOT` macro, as follows:

```
%distplot(data=saxony, count=males, freq=families,
          dist=binomial);
```

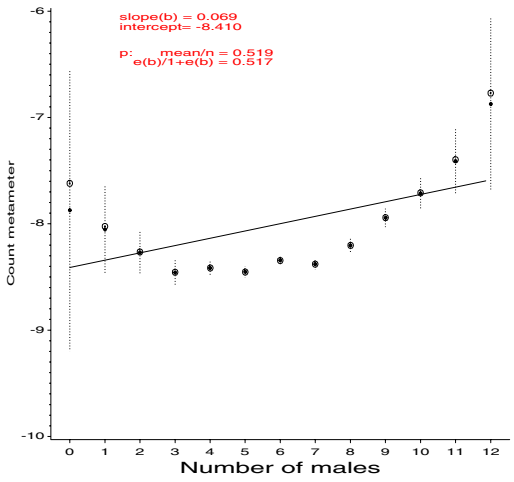

Figure 2: Robust distribution plot for Saxony data

This plot has the property that the circled points are linear in k when the data follow the assumed distribution. However, the ordinate “count metameter” depends only on n_k, and the confidence bars are calculated to take into account the variability of individual counts, n_k, in the observed distribution.

3 Contingency tables

Chapters 3–5 provide a wide variety of facilities for the analysis and visual display of contingency tables, some of which are shown in Table 2

Chapter 3: “Two-way contingency tables” presents methods of analysis designed mainly for two-way tables of frequencies (contingency tables), along with graphical techniques for understanding the patterns of associations between variables. Different specialized displays are focused on visualizing an odds ratio (a fourfold display for 2×2 tables), or the general pattern of association (sieve diagrams), the agreement between row and column categories (agreement charts), and relations in $n \times 3$ tables (trilinear plots).

Chapter 4: “Mosaic displays for n-way tables” introduces the mosaic display, a general method for visualizing the pattern of associations among variables in two-way and larger tables. Extensions of this technique can reveal partial associations, marginal associations, and shed light on the structure of loglinear models them-

Table 2: Some tasks for contingency tables

Topic	Task	Program	Examples
2×2 tables	Visualize odds ratio	FOURFOLD	3.8
$2 \times 2 \times k$ tables	Visualize odds ratios	FOURFOLD	§3.4.3
”	Homogeneity of association	FOURFOLD	3.9
$r \times c$ tables	Display observed, expected frequencies	SIEVE	3.10, 3.11
”	Measures of association	Freq	3.5
”	Use ordinal variables	Freq	§3.2.4
”	Control for other variable(s)	Freq	3.6
”	Homogeneity of association	Freq	3.7
”	Fit independence model	MOSAIC	4.1
”	Visualize association	MOSAIC	4.1
square tables	Visualize agreement	AGREE	3.15
”	Fit quasi-independence	MOSAIC	4.3
$r \times 3$ tables	Trilinear plots	TRIPLOT	3.16–3.18
three-way tables	Fit, visualize models	MOSAIC	§4.3.1
n-way tables	Fit, visualize models	MOSAIC	4.4, 4.5
”	Test, visualize partial association	MOSAIC	4.6
”	Visualize all pairwise association	MOSMAT	4.7, 4.8
”	Visualize conditional associations	MOSMAT	4.7, 4.8
”	Visualize loglinear structure	MOSMAT	§4.5

selves.

Chapter 3: “Correspondence analysis” discusses correspondence analysis, a technique designed to provide visualizations of associations in a two-way contingency table in a small number of dimensions. Multiple correspondence analysis extends this technique to n-way tables. Other graphical methods, including mosaic matrices and biplots provide complementary views of loglinear models for two-way and n-way contingency tables.

3.1 Fourfold displays

For 2×2 (and $2 \times 2 \times k$) tables, the focus is often on the odds ratio as a measure of association. The fourfold display is designed to display such data, indicating the direction and significance of associations. It includes confidence rings, having the property that the rings for adjacent segments overlap when no significant association is shown.

We consider here some data from Gilovich et al. (1985) on the topic of a “hot-hand” in basketball, i.e., whether a player given two free-throws is more likely to have a “hit” following a first hit than following a first miss. The data come from 9 players on the Boston Celtics in the 1980–1982 seasons.

Figure 3 shows the aggregate data. There is a strongly positive, and significant association between first- and second-shot hits, supporting the hot-hand notion. However, the aggregate data are misleading, because collapsing over players assumes that all players get equal chances at free-throws, and they are all equally accurate, overall.

Figure 4 shows the individual data, with the players sorted by their odds ratio of having a second hit, given their first shot was a hit. In all cases, the confidence rings overlap, indicating no association between first- and second-shot accuracy.

These plots are produced by the `FOURFOLD` program and the associated macro, `FFOLD`, which provides a simpler interface to the SAS/IML program. The printed output includes significance tests for individual odds ratios, and tests of homogeneity of association (here, over players) and conditional association (controlling for players).

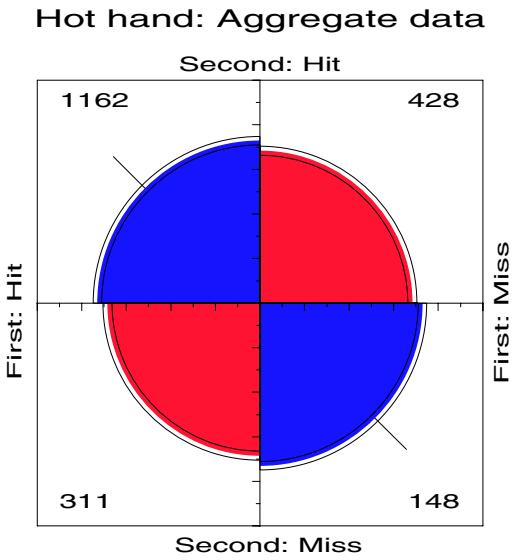

Figure 3: Fourfold display for hot hand data, aggregated over players

For example, the plot of the individual data is produced as follows (omitting steps to sort the player by odds ratio).

```
data hothand;
  input first $ second $ @;
  do player ='LB', 'CM', 'RP', 'NA', 'CF',
             'KM', 'MC', 'RR', 'GH';
    input count @;
    output;
    end;
cards;
Hit   Hit   251 245  164  203  36  93  39  54  77
```

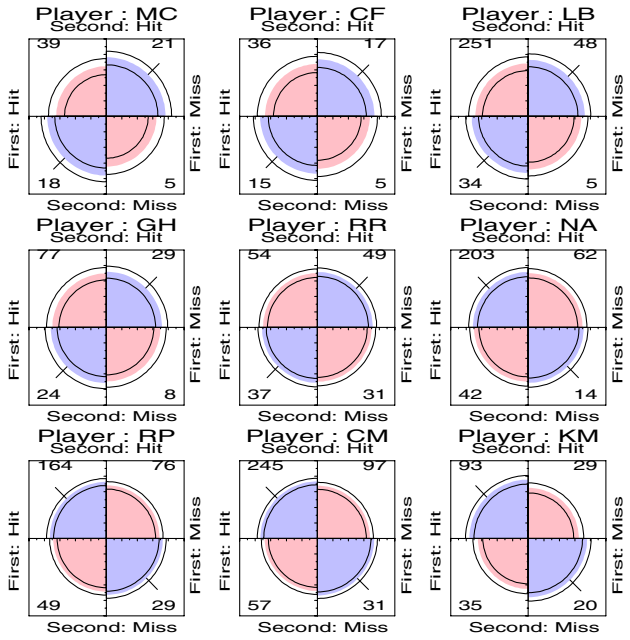

Figure 4: Fourfold display for hot hand data, by player

```
Hit   Miss    34  57   49   42  15  35  18  37  24
Miss  Hit     48  97   76   62  17  29  21  49  29
Miss  Miss     5  31   29   14   5  20   5  31   8
;
%ffold(data=hothand, var=First Second, by=Player,htext=3);
```

To plot the aggregate data, we must first sum the data over players, which is done using the TABLE macro.

```
%table(data=hothand, out=hot2, var=first second,
       order=data, weight=count);
%ffold(data=hot2, var=First Second,
       ptitle=Hot hand: Aggregate data);
```

3.2 Mosaic displays

The mosaic display, proposed by Hartigan and Kleiner (1981) is a graphical method to show the values (cell frequencies) in a contingency table cross-classified by one or more "factors". As extended to show both the data, and residuals from a log-linear model (Friendly, 1994), it has become a primary graphical tool for visualization and analysis of categorical data in the form of contingency tables. In each case, a loglinear model is fit to the data, and the sign and magnitude of residuals in the model are shown by color and shading (blue for positive, red for negative, color intensity ∼ magnitude). The pattern of residuals show the *nature* of associations, and help suggest a more adequate model.

Some examples are shown in Figure 5, for a two-way and in Figure 6 for a three-way table, showing the relations among the categories of hair color, eye color and sex in a sample of individuals. The two way display fits the model of independence; the residuals show, however, that people with dark hair are more likely to have dark eyes, while people with light hair tend to have light eyes. The three-way display fits the model [HairEye] [Sex], asserting that the combinations of hair color and eye color are independent of sex; the residuals show that this is largely true, except for blue-eyed blonds, where the proportion of females is significantly greater.

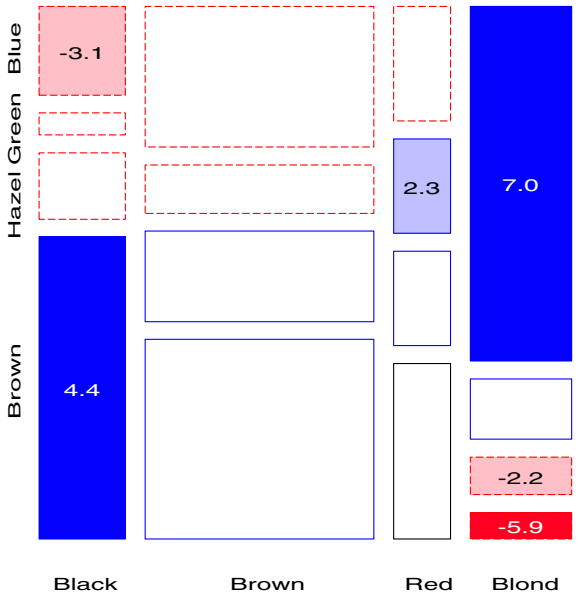

Figure 5: Mosaic display for frequencies of hair color, eye color

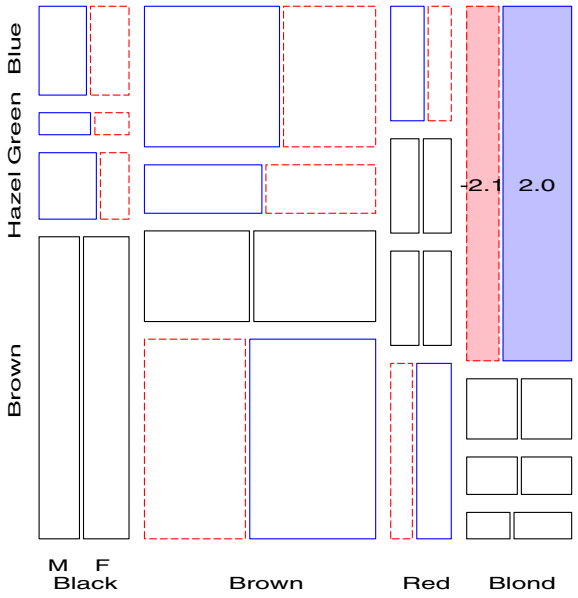

Figure 6: Mosaic display for frequencies of hair color, eye color and sex

Like the FOURFOLD program, MOSAICS is a SAS/IML program with a wide variety of options and features. It may be used directly within SAS/IML, or more easily through an associated macro, MOSAIC. For example, both Figure 5 and Figure 6 are produced using one call to the MOSAIC macro:

```
%include catdata(hairdat);
%mosaic(data=haireye, vorder=Hair Eye Sex, plots=2:3,
  htext=1.75, cellfill=dev);
```

The VORDER parameter specifies the order of variables in the series of mosaics; PLOTS=2:3 produces plots of the two- and three-way tables; CELLFILL=DEV prints the value of the standardized residual in each shaded cell.

Mosaic displays easily generalize to n-way tables, and their great benefit is the ability to find adequate descriptive models for complex tables visually, rather than from tables of parameter estimates.

For example, Figure 7 shows two models fit to data on the relationships pre-marital sex, extra-marital sex, gender, and marital status. The left panel fits the base model [GPE] [M], which says that marital status is independent of all the other variables. the pattern of residuals indicates associations of both pre-marital sex and extra-marital sex with marital status, giving the model [GPE] [PEM], which fits satisfactorally.

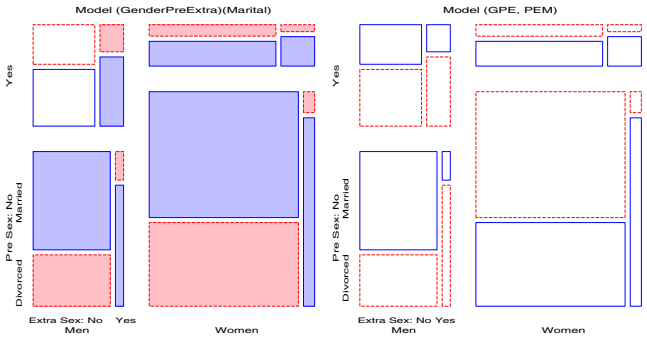

Figure 7: Four-way mosaic for marital status data. Left: model [GPE] [M]; right: model [GPE] [PEM]

3.3 Mosaic matrices

The ***mosaic matrix*** is a discrete analog for multivariate categorical data of the scatterplot matrix (Friendly, 1999). Like the scatterplot matrix, it contains all $p(p-1)$ pairwise plots for a p-variate dataset, but displays the relation of each pair of variables by a mosaic. Extensions of this idea include: (a) a conditional mosaic matrix, which fits a model of conditional independence between each row and column, controlling for one or more of the other variables—a generalization of partial regression plots, (b) mosaic displays of partial association, stratified by one or more variables—a discrete analog of coplots or Trellis displays.

Figure 8 shows the bivariate marginal relations among all pairs of variables in the marital status data, produced with the `MOSMAT` macro, as follows:

```
%include catdata(marital);
%mosmat(data=marital, var=Gender Pre Extra Marital,
   vorder=Marital Extra Pre Gender, devtype=LR ADJ);
```

Viewing Gender, Premarital sex and Extramarital sex as explanatory, and Marital status as the response, the mosaics in row 1 (and in column 1) shows how marital status depends on each predictor marginally. The remaining panels show the relations within the set of explanatory variables.

Thus we see (row 1, column 4) that marital status is independent of gender, by design of the data collection. In the (1, 3) panel, we see that reported premarital sex is more often followed by divorce, while non-report is more prevalent among those still married. The (1, 2) panel shows a similar, but stronger relation between extramarital sex and marriage stability. These effects pertain to the associations of P and E with marital status—the terms [PM] and [EM] in a loglinear model.

Among the background variables, the (2, 3) panel shows a strong relation between premarital sex and subsequent extramarital sex,

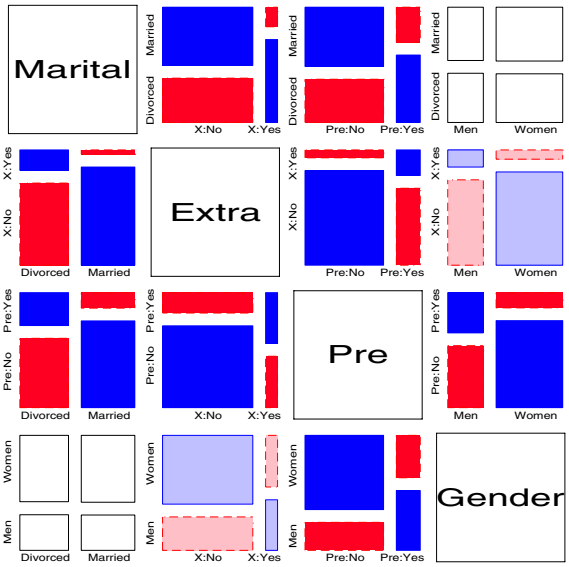

Figure 8: Mosaic matrix for marital status data. Each panel shows the bivariate marginal association.

while the (2, 4) and (3, 4) panels show that men are far more likely to report premarital sex than women in this sample, and also more likely to report extramarital sex.

3.4 Correspondence analysis

Correspondence analysis is an analog of principal components analysis for frequency data, designed to display the association among categorical variables in a small number of dimensions, designed to account for the largest proportion of the Pearson χ^2. Multiple correspondence analysis extends this method to n-way tables, but displays only bivariate associations, analogous to the (marginal) mosaic matrix.

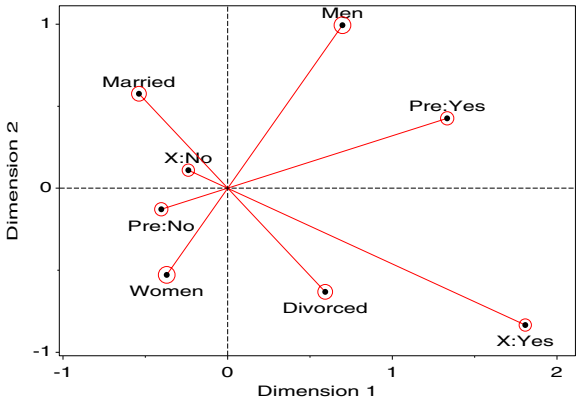

Figure 9: 2D multiple correspondence analysis display for marital status data

Figure 9 shows the 2D MCA solution for the marital status data. This graph was prepared by the `CORRESP` macro as follows:

```
%corresp(data=marital, tables=gender pre extra marital,
   weight=freq, options=mca, interp=vec, inc=1, pos=-,
   symbols=dot);
```

The macro uses the CORRESP procedure for the calculations, but extends it with extensive facilities for easily visualizing the results.

4 Logistic regression & Loglinear models

The final two chapters turn to model-based methods for the analysis of discrete data, where the emphasis is more on confirmatory testing than on data exploration.

Chapter 6: "Logistic regression" introduces the model-building approach of logistic regression, designed to describe the relation between a discrete response, often binary, and a set of explanatory variables. Smoothing techniques are often crucial in visualizations for such discrete data. The fitted model provides both inference and prediction, accompanied by measures of uncertainty. Diagnostic plots help us to detect influential observations which may distort our results.

Chapter 7: "Loglinear and logit models" extends the model building approach to loglinear and logit models. These are most easily interpreted through visualizations, including mosaic displays and plots of associated logit models. As with logistic regression, diagnostic plots and influence plots help to assure that the fitted model is an adequate summary of associations among variables.

4.1 Plotting binary response data

It is sometimes difficult to understand how a binary response can give rise to a smooth, continuous relationship between the predicted response and an explanatory variable, particularly when the predictor is continuous. It is helpful, therefore, to plot the observed sample probabilities (or logits) against X, together with the observations (in a way which avoids overplotting), and the fitted relationships, as we do in Figure 10 for data on treatment outcome for arthritis patients.

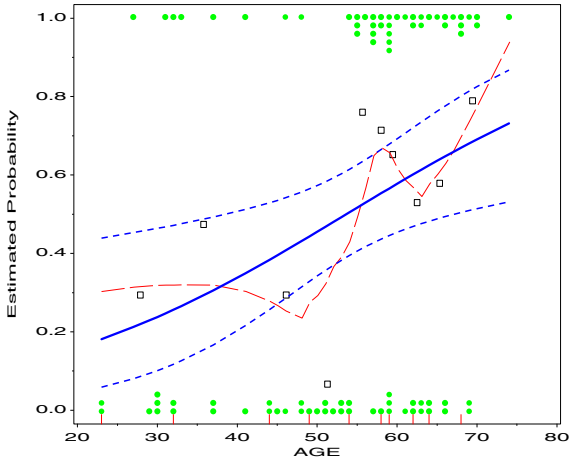

Figure 10: Empirical probability plot for arthritis data.

In this figure the observed responses are shown by stacked points at the top and bottom, and summarized by sample probabilities in 10 intervals; the solid line shows the predicted probability of improvement as a function of age, and the irregular curve is a lowess smooth of the sample probabilities. Such plots are produced by the LOGODDS macro,

```
%logodds(data=arthrit, x=Age, y=Better, smooth=0.5);
```

For more complex models, it is often easier to interpret model results from plots of predicted log-odds or probabilities than from estimated parameter values. Figure 11 shows plots of fitted logits with standard error bars for the arthritis data, fitting a models with main effects for Age, Sex, and Treatment. A probability scale at the left allows interpretation in terms of probabilities.

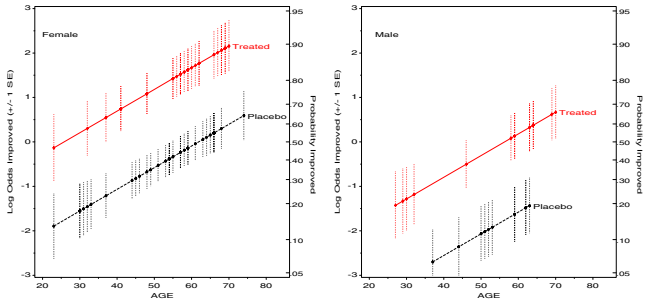

Figure 11: Estimated logits for sex, treatment and age. Corresponding probabilities of a "better" response are shown on the right scale.

4.2 Influence and diagnostic plots

A variety of diagnostic plots for logistic regression, designed to identify observations which have undue impact on the fitted model are provided by the INFLOGIS macro. An analogous macro, INFLGLIM provides similar plots for any generalized linear model which may be fit with PROC GENMOD.

One quite useful plot shows the estimated change in the χ^2 against the "hat" value measure of leverage, using the discrete analog of Cook's D as the size of a bubble symbol. For example, the INFLOGIS macro may be used as follows to give Figure 12.

```
%inflogis(data=arthrit,
  y=better, x=_sex_ _treat_ age, id=id,
  gy=DIFCHISQ, gx=HAT);    /* graph y, x */
```

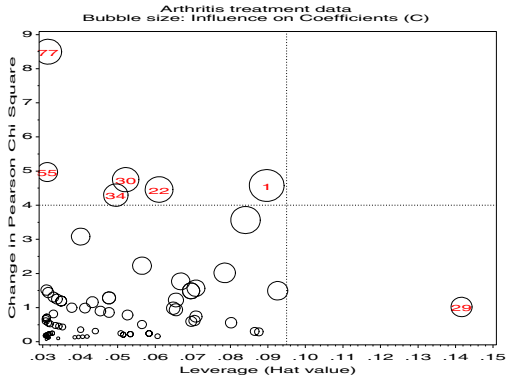

Figure 12: Changes in chi-square vs. leverage. Possibly influential cases are labeled by ID number.

Other diagnostic plots for logistic models include partial residual plots, added variable and constructed variables plots (ADDVAR macro).

4.3 Loglinear and logit models

Whereas logit models focus on the prediction of one response factor, loglinear models treat all variables symmetrically, and attempt

to model all important associations among them. Both types of models are most easily understood through visualizations, including mosaic displays and plots of associated logit models, provided with the CATPLOT macro. As with logistic regression, diagnostic plots and influence plots help to assure that the fitted model is an adequate summary of associations among variables.

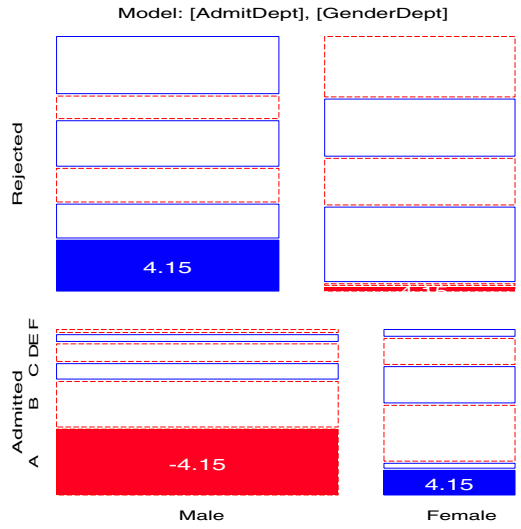

Figure 13: Mosaic display for Berkeley admissions data

The MOSAIC macro fits loglinear models, but it may also be used to visualize the results of models fit using PROC GENMOD. For example, Figure 13 displays the residuals of the model [AD][DG] fit to data on admission to graduate school at Berkeley classified by admission (A), department (D) and gender (G). Figure 13 is produced as follows:

```
proc genmod data=berkeley;
   class dept gender admit;
   model freq = admit|dept gender|dept /
                dist=poisson obstats residuals;
   make 'obstats' out=obstats;
data obstats;
   merge berkeley obstats;
%mosaic(data=obstats, vorder=Admit Gender Dept,
   count=freq, resid=streschi, cellfill=dev,
   title=Model: [AdmitDept] [GenderDept]);
```

This plot indicates that the model [AD][DG] fits well, except in Dept. A.

Because admission may be considered as a response, all loglinear models may be recast as an equivalent logit model. For example, the model [AD][DG][AG] is equivalent to

$$\text{logit}\,(\text{Admit}) = \alpha + \beta_i^{\text{Dept}} + \beta_j^{\text{Gender}} \quad . \tag{1}$$

That is, the logit model (1) asserts that department and gender have additive effects on the odds of admission. This model may be fit with PROC CATMOD,

```
proc catmod order=data data=berkeley;
   weight freq;
   response / out=predict;
   model admit = dept gender / ml noiter noprofile ;
```

and the fitted logits plotted with the CATPLOT macro. The PSCALE macro supplies an Annotate dataset to draw a probability scale at the right.

```
%pscale(lo=-4, hi=3, anno=pscale);
title 'Model:  logit(Admit) = Dept Gender'
      a=-90 'Probability (Admitted)' h=3.5 a=-90 ' ';
axis1 order=(-3 to 2) offset=(4)
      label=(a=90 'Log Odds (Admitted)');
axis2 label=('Department') offset=(4);
symbol1 i=none v=circle h=1.7 c=black;
symbol2 i=none v=dot    h=1.7 c=red  ;
%catplot(data=predict, xc=dept, y=_obs_, class=gender,
   type=FUNCTION, z=1.96, anno=pscale);
```

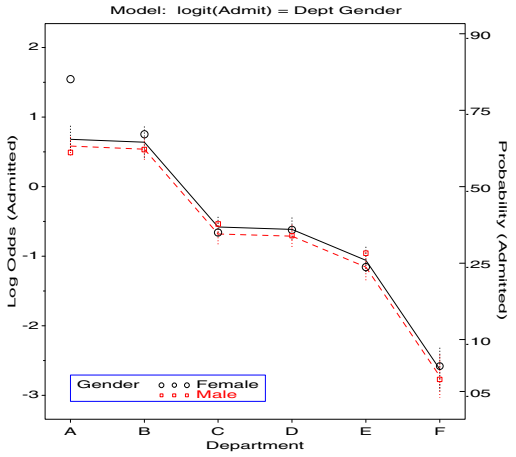

Figure 14: Observed (points) and fitted (lines) log odds of admission in the logit model corresponding to [AD][AG][DG]. The error bars show individual 95% confidence intervals around each fitted logit.

The effect of gender is very small and non-significant, implying that the simpler model, $\text{logit}\,(\text{Admit}) = \alpha + \beta_i^D$ is adequate, as we saw in Figure 13.

4.4 Diagnostic plots

Influence plots for loglinear models are provided by the INFLGLIM macro. The HALFNORM macro gives half-normal plots of the residuals from any generalized linear model. A simulated envelope, corresponding to an approximate 95assessment of whether the distribution of residuals corresponds to a good-fitting model.

We obtain an influence plot of adjusted Pearson residuals against hat values, showing Cook's D by bubble size, by

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

The plot clearly identifies the four deviant cells corresponding to Dept. A. Finally, we show a half-normal plot for this model in Figure 16, produced with the HALFNORM macro,

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

5 Other tools

In working with categorical data and making many custom graphs, a number of general tools for manipulating categorical data and annotating graphs were developed. Among these, the TABLE macro

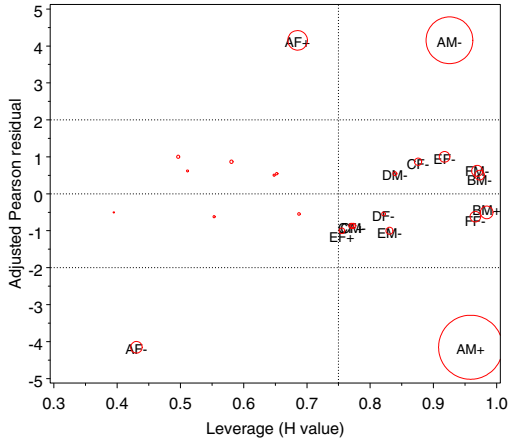

Figure 15: Influence plot for Berkeley admissions data, Model $[AD][GD]$. Bubble areas are proportional to Cook's D.

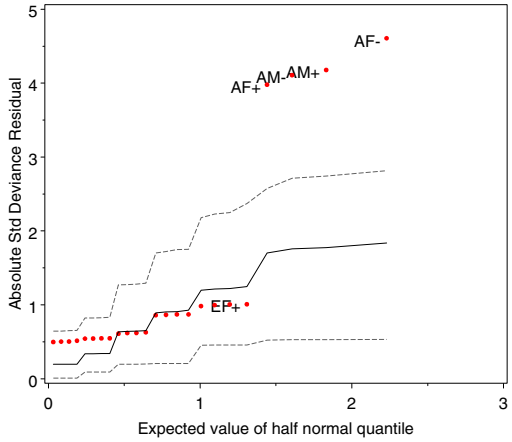

Figure 16: Half-normal residual plot for Berkeley admissions data, Model $[AD][GD]$

has been invaluable in collapsing, recoding or reordering the levels of discrete variables in data sets, and the `SORT` macro generalizes the idea of sorting to include sorting a data set according to the values of a user-specified format, or by the values of any summary statistic computed by `PROC UNIVARIATE`.

On the graphic side, I often found myself annotating graphs with points, lines, error bars, and point labels. It annoys me to code the same steps repeatedly, so a variety of utility macros for these tasks were developed. Likewise, a general macro, `PANELS`, for arranging any number of individual graphs in a rectangular array is provided.

As stated at the outset, my goal in writing *VCD* was to make it easy for people to apply these methods for graphical analysis of categorical to their own data, turning "theory into practice".

A Macros and Programs

The following macros and programs are described and illustrated in *VCD*. All require SAS/STAT and SAS/GRAPH; many require SAS/IML. They are available at `www.math.yorku.ca/SCS/vcd/`.

`ADDVAR`	Added variable plots for logistic regression
`AGREE`	Observer agreement chart (SAS/IML)
`BIPLOT`	Generalized biplot displays
`CATPLOT`	Plot results from `PROC CATMOD`
`CORRESP`	Plot `PROC CORRESP` results
`DISTPLOT`	Plots for discrete distributions
`DUMMY`	Create dummy variables
`FOURFOLD`	Fourfold displays for $2 \times 2 \times k$ tables (SAS/IML)
`GOODFIT`	Goodness-of-fit for discrete distributions
`HALFNORM`	Half-normal plots for generalized linear models
`INFLGLIM`	Influence plots for generalized linear models
`INFLOGIS`	Influence plots for logistic regression
`LAGS`	Calculate lagged frequencies for sequential analysis
`LOGODDS`	Plot empirical logits for binary data
`MOSAIC`	Mosaic displays (macro)
`MOSAICS`	SAS/IML modules for mosaic displays
`MOSMAT`	Mosaic matrices (macro)
`ORDPLOT`	Ord plot for discrete distributions
`PANELS`	Arrange multiple plots in a panelled display
`POISPLOT`	Poissonness plot
`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
`ROOTGRAM`	Hanging rootograms
`SIEVE`	Sieve diagrams (SAS/IML)
`SORT`	Sort a dataset by a statistic or formatted value
`TABLE`	Construct a grouped frequency table, with recoding
`TRIPLOT`	Trilinear plots for $n \times 3$ tables
`Utility`	Graphics utility macros: `BARS`, `EQUATE`, `GDISPLA`, `GENSYM`, `GSKIP`, `LABEL`, `POINTS`, `PSCALE`.

References

Friendly, M. Graphical methods for categorical data. *Proceedings of the SAS User's Group International Conference*, 17:1367–1373, 1992.

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*, 8:373–395, 1999.

Friendly, M. *Visualizing Categorical Data*. SAS Institute, Cary, NC, 2000.

Gilovich, T., Valone, R., and Tversky, A. The hot hand in basketball: On the misrepresentation of random sequences. *Cognitive Psychology*, 17:295–314, 1985.

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.

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