Categorical Data Analysis with Graphics

Michael Friendly
SCS Short Course*

May 16, 2003

Contents

1 Framework for Categorical Data Analysis

Methods of analysis of categorical data fall into two categories:

- **Non-parametric, randomization-based methods**
  - ▷ make minimal assumptions
  - ▷ useful for hypothesis-testing
  - ▷ SAS: `PROC FREQ`
    - ∗ Pearson Chi-square

*An earlier version of these notes is online at `http://www.math.yorku.ca/SCS/Courses/grcat/`

  - 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`, `PROC CATMOD`, `PROC GENMOD` , `PROC INSIGHT` (Fit YX)
    - estimate standard errors, covariances for model parameters
    - confidence intervals for parameters, predicted Pr{response}

Graphical methods for categorical data

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

You can see a lot, just by looking. Yogi Berra

Getting information from a table is like extracting sunlight from a cucumber.
Farquhar & Farquhar, 1891

Graphical methods for quantitative data are well-developed. From the basic display of data in a scatterplot, to diagnostic methods for assessing assumptions and finding transformations, to the final presentation of results, graphical techniques are commonplace adjuncts to most methods of statistical analysis. Graphical methods for categorical data are still in infancy. There are not many methods, and they are not widely used. Wondering why this is provokes several thoughts:

- **Exploratory methods**
  Many of the graphical methods described here make minimal assumptions about the data, like the non-parametric statistical methods. Their goal is to help the viewer see the data, detect patterns, and suggest hypotheses.

- **Graphic metaphor?**
  The basic metaphor for displaying quantitative data is **magnitude $\sim$ position along an axis**. Categorical data consist of counts of observations in discrete categories. Some of the methods described here (e.g., sieve diagram, mosaic display) suggest the metaphor

  count $\sim$ area

- **Generalizations?**
  The scatterplot is a basic tool for viewing raw (quantitative) data. It generalizes readily to three or more variables in the form of the scatterplot matrix – a matrix of pairwise scatterplots. The mosaic display is a simple graphic method for looking at cross-classified data which generalizes to more than two-way tables. Are there others?

- **Analogies?**
  Model-based methods for analyzing categorical data, such as logistic regression and log-linear models, are discrete analogs of methods of regression and analysis of variance for quantitative data. We can adapt some of the familiar graphical methods to categorical data.

- **Presentation plots for model-based methods**
  Results of model-based analysis are almost invariably presented in tables of estimated frequencies, parameter estimates, log-linear model effects, and so forth. Effect displays of estimated probabilities of response or log odds provide a useful alternative.

- **Practical power = Statistical power $\times$ Probability of Use**
  Statistical and graphical methods are of practical value to the extent that they are available and easy to use. Statistical methods for categorical data analysis have (nearly) reached that point. Graphical methods still have a long way to go. One aim for this workshop is to show what can now be done, with some examples of how to do it.

SAS programs

All of the graphs shown here were constructed with SAS/GRAPH software. Some of the SAS programs are available on SAS for Windows on the Hebb and Arts servers for the Psych Lab, and on `phoenix.yorku.ca`. You may also access some of these via the World Wide Web at

```
http://www.math.yorku.ca/SCS/friendly.html
http://www.math.yorku.ca/SCS/vcd/
http://euclid.psych.yorku.ca/ftp/sas/catdata/
```

Some of these are designed as general macro programs which can be used with any set of data; others are specific to the particular application.

The following programs are currently available:

`CATPLOT`	[Macro] Plot predicted logits from `PROC CATMOD`
`CATDEMO`	Demonstration of CATPLOT
`FOURFOLD`	IML modules for four-fold display of $2 \times 2 \times k$ tables
`FOURDEMO`	Demonstration of FOURFOLD
`GLOGIST1`	Logistic regression: binary response
`GLOGIST2`	Logistic regression: proportional odds model
`INFLGLIM`	[Macro] Influence plots for loglinear models
`INFLOGIS`	[Macro] Influence plots for logistic regression
`KAPPA`	Calculate Cohen's κ for agreement
`LABEL`	[Macro] Label points on a plot
`LAGS`	[Macro] Calculate lagged frequencies for sequential analysis
`MOSAICS`	IML modules for mosaic displays
`MOSADEMO`	Sample program for mosaic displays
`POISPLOT`	[Macro] Poissonness plot
`POISDEMO`	Demonstration of Poissonness plot
`PSCALE`	[Macro] Constructs an Annotate dataset for a probability scale
`ORDPLOT`	[Macro] Ord plots for discrete distributions
`ORDDEMO`	Demonstration of Ord plot
`SIEVE`	IML modules for sieve diagram
`SIEVE1`	Demonstration for sieve diagram

Most of these programs are illustrated in *Visualizing Categorical Data* (Friendly, 2000a). Other methods for graphical display of data are described in an earlier book, *SAS System for Statistical Graphics, First Edition* (Friendly, 1991). For SPSS users, there are online course notes, data sets and examples of loglinear model analysis by Brendan Halpin at `http://wivenhoe.staff8.ul.ie/~brendan/CDA/`.

2 Plots for discrete distributions

Discrete frequency distributions often involve counts of occurrences such as accidental fatalities, words in passages of text, 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 , $k = 0, 1, \ldots$.

The table below shows two such data sets:

- von Bortkiewicz's data on death of soldiers in the Prussian army from kicks by horses and mules (Andrews and Herzberg, 1985, p. 18). The data pertain to 10 army corps, each observed over 20 years. In 109 corps-years, no deaths occurred; 65 corps-years had one death, etc. (see Figure 1).

- In a study of the potential to identify authors from frequency distributions of various words, Mosteller and Wallace (1984) presented these data on the occurrence of the word *may* in 262 blocks of text (each about 200 words long) from issues of the *Federalist Papers* known to be written by James Madison. In 156 blocks, the word *may* did not occur; it occurred once in 63 blocks, etc. (see Figure 2).

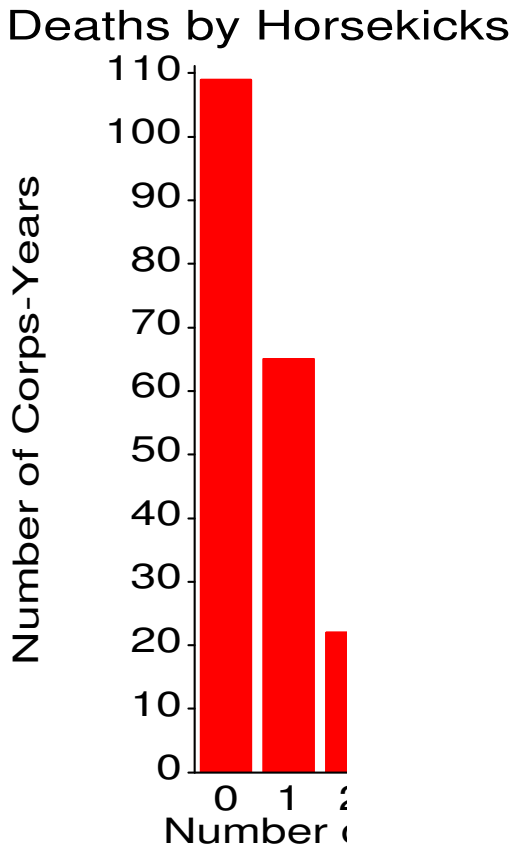


Figure 1: von Bortkiewicz's data

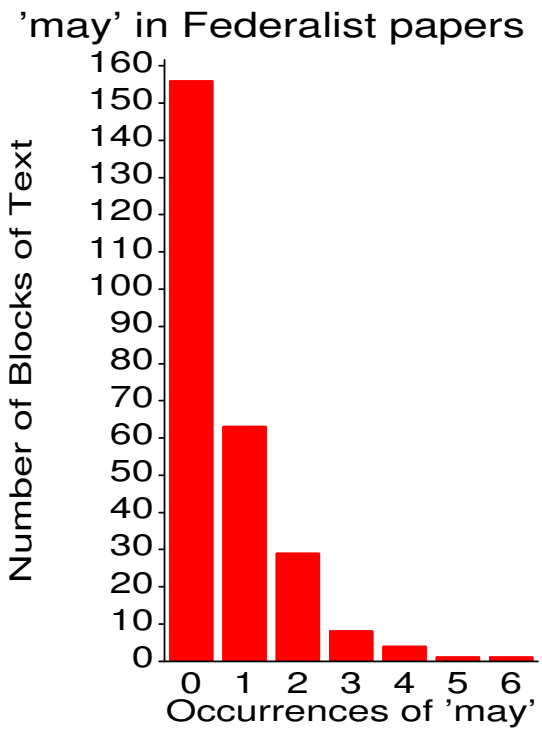


Figure 2: Mosteller & Wallace data

Deaths by Horsekick

k	nk
0	109
1	65
2	22
3	3
4	1
	N=200

Occurrences of 'may'

k	nk
0	156
1	63
2	29
3	8
4	4
5	1
6	1
	N=256

2.1 Fitting discrete distributions

Often interest is focussed on how closely such data follow a particular distribution, such as the Poisson, binomial, or geometric distribution. Usually this is examined with a classical goodness-of-fit chi-square test,

$$\chi^2 = \sum_{k=1}^{K} \frac{(n_k - N\hat{p}_k)^2}{N\hat{p}_k} \sim \chi^2(K-1)$$

where $\hat{p}_k$ is the estimated probability of each basic count, under the hypothesis that the data follows the chosen distribution.

For example, for the Poisson distribution, the probability function is

$$\Pr_{\lambda}\{X = k\} = p_k = \frac{e^{-\lambda}\,\lambda^k}{k!} \tag{1}$$

The maximum likelihood estimator of the parameter λ in (1) is just the mean of the distribution,

$$\hat{\lambda} = \frac{\sum k\, n_k}{N}$$

For the horsekick data, the mean is 122/200 = .610, and calculation of Poisson probabilities (PHAT), expected frequencies, and contributions to χ^2 are shown below.

k	nk	p	phat	exp	chisq
0	109	0.545	0.54335	108.670	0.00100
1	65	0.325	0.33144	66.289	0.02506
2	22	0.110	0.10109	20.218	0.15705
3	3	0.015	0.02056	4.111	0.30025
4	1	0.005	0.00313	0.627	0.22201
	===			=======	=======
	200			199.915	0.70537 $\sim$ χ^2 (4)

In this case the χ^2 shows an exceptionally good (unreasonably good?) fit. In the word frequency example, the fit of the Poisson turns out not to be close at all. However, even a close fit may show something interesting, if we know how to look; conversely, it is useful to know why or where the data differ from a chosen model.

2.1.1 The GOODFIT macro

The GOODFIT macro carries out Pearson χ^2 and likelihood-ratio goodness-of fit tests for the uniform, binomial, Poisson, negative binomial, logarithmic series, and geometric distributions, as well as any discrete distribution whose probabilities you can specify.The data may consist either of individual observations on a single variable, or a grouped frequency distribution in the form shown for the Horse Kick data.

The macro is used as follows:

```
%goodfit(data=SASdatasetname ,
   var=variablename ,
   freq=variablename ,
   dist=distribution ,
   parm=parameters ,
   sumat=value ,
   format=SASformat ,
   out=outputdatasetname ,
   outstat=statisticsdatasetname );
```

The data on the occurrences of the word *may* in Madison's Federalist Papers are fit to both the Poisson and Negative binomial distributions as shown below. In each case, the parameters are estimated from the data. The output for the Poisson distribution appears in Output 2.1 and Output 2.2. The results for the Negative binomial distribution appear in Output 2.3 and Output 2.4.

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

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

Output 2.1 Fitting the Poisson(λ) to the Federalist Papers data: Observed and fitted frequencies

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		

Output 2.2 Fitting the Poisson(λ) to the Federalist Papers data: Goodness of fit tests

Instances of 'may' in Federalist papers

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

Output 2.3 Fitting the Negative binomial(n, p) to the Federalist Papers data: Observed and fitted frequencies

COUNT	BLOCKS	PHAT	EXP	CHI	DEV
0	156	0.59047	154.702	0.10434	1.61446
1	63	0.25343	66.398	-0.41706	-2.57290
2	29	0.09826	25.743	0.64188	2.62853
3	8	0.03674	9.625	-0.52374	-1.72003
4	4	0.01348	3.532	0.24905	0.99777
5	1	0.00489	1.281	-0.24862	-0.70425
6	1	0.00176	0.461	0.79297	1.24381
	262	0.99902	261.743		

Output 2.4 Fitting the Negative binomial(n, p) to the Federalist Papers data: Goodness of fit tests

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

Analysis variable:       COUNT Number of Occurrences
Distribution:            NEGBIN
Estimated Parameters:    n, p = 1.2397, 0.6538


Pearson chi-square    = 1.6237622915
Prob > chi-square     = 0.8045151082

Likelihood ratio G2   = 1.9839511084
Prob > chi-square     = 0.7387108792

Degrees of freedom    = 4
```

2.1.2 Plots of observed and fitted frequencies

Plots of the observed and fitted frequencies can help to show both the shape of the theoretical distribution we have fitted and the pattern of any deviations between our data and theory.

Figure 3(a) shows the fit of the Poisson distribution to the Federalist papers data, using one common form of plot that is sometimes used for this purpose. In this plot, observed frequencies are shown by bars and fitted frequencies are shown by points, connected by a smooth (spline) curve.

Such a plot, however, is dominated by the largest frequencies, making it hard to assess the deviations among the smaller frequencies. To make the smaller frequencies more visible, Tukey (1977) suggest plotting the frequencies on a square-root scale, which he calls a *rootogram* (see Figure 3(b)). An additional improvement is to move the rootogram bars so their tops are at the expected frequences (giving a *hanging rootogram*, Figure 3(c)). This has the advantage that we can more easily judge the pattern of departures against the horizontal reference line at 0, than against the curve. A final variation is to emphasize the differences between the observed and fitted frequencies by drawing the bars to show the gaps between the 0 line and the (observed-expected) difference (Figure 3(d)).

These plots are produced by the `ROOTGRAM` macro using the `OUT=FIT` dataset from the `GOODFIT` macro:

```
title "Instances of 'may' in Federalist papers" ;
%include catdata(madison);
%goodfit(data=madison, var=count, freq=blocks, dist=poisson, out=fit);

title;
%rootgram(data=fit, var=count, obs=blocks, btype=0, func=none);  /* a */
%rootgram(data=fit, var=count, obs=blocks, btype=0);             /* b */
%rootgram(data=fit, var=count, obs=blocks);                      /* c */
%rootgram(data=fit, var=count, obs=blocks, btype=dev);           /* d */
```

2.2 Poissonness plot

The Poissonness plot (Hoaglin, 1980) is designed as a plot of some quantity against k, so that the result will be points along a straight line when the data follow a Poisson distribution. When the data deviate from a Poisson, the points will be curved. Hoaglin and Tukey (1985) develop similar plots for other discrete distributions, including the binomial, negative binomial, and logarithmic series distributions.

Assume, for some fixed λ, each observed frequency, n_k equals the expected frequency, $m_k = Np_k$. Then, setting $n_k = Np_k = e^{-\lambda}\,\lambda^k/k!$, and taking logs of both sides gives

$$\log(n_k) = \log\, N - \lambda + k\, \log\, \lambda - \log\, k!$$

which can be rearranged to

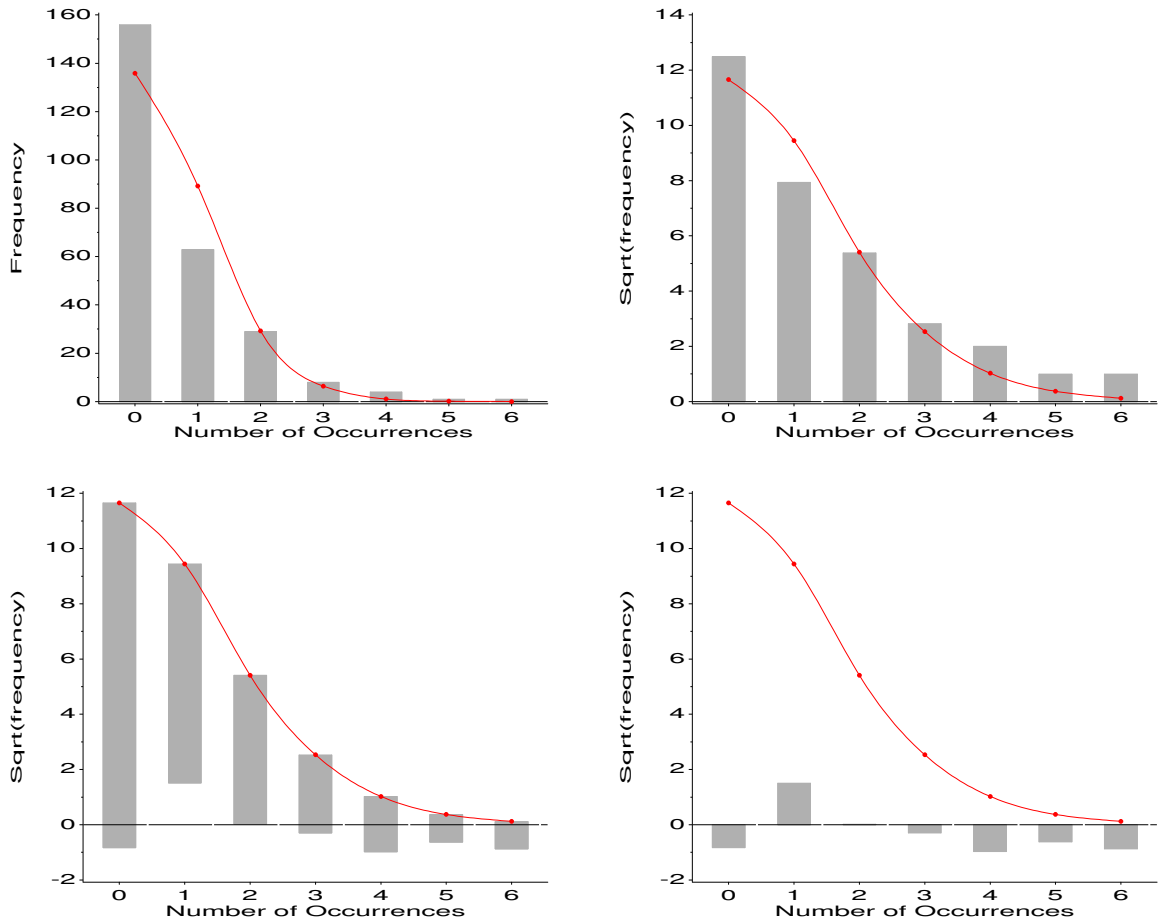


Figure 3: Plots of observed and fitted frequencies for the Federalist Papers data, Poisson model. Each panel shows the fitted frequencies as a smooth curve and observed frequencies as a bar. Panel (a) raw frequencies; panels (b)-(d) on a square-root scale, to emphasize smaller frequencies. Panel (c) is a hanging rootogram, where observed - fitted differences can be judged relative to the horizontal line. Panel (d) shows only the difference between the observed and fitted frequency.

$$\log\left(\frac{k!\, n_k}{N}\right) = -\lambda + (\log\ \lambda)\, k \tag{2}$$

The left side of (2) is called the ***count metameter***, and denoted $\phi\,(n_k) = k!\, n_k/N$. Hence, plotting $\phi(n_k)$ against k should give a line with

- intercept = $-\lambda$
- slope = $\log\ \lambda$

2.2.1 Features of the poissonness plot

- ***Resistance***: a single discrepant value of n_k affects only the point at value k.
- ***Comparison standard***: An approximate confidence interval can be found for each point, indicating its inherent variability and helping to judge whether each point is discrepant.
- ***Influence***: Extensions of the method result in plots which show the effect of each point on the estimate of the main parameter of the distribution (λ in the Poisson).

The calculations for the poissonness plot, including confidence intervals, is shown below for the horse kicks data. See the plots in Figure 4.

k	nk	$\phi(n_k)$ Y	CI center	CI width	Confidence Int lower	upper
0	109	-0.607	-0.617	0.130	-0.748	-0.487
1	65	-1.124	-1.138	0.207	-1.345	-0.931
2	22	-1.514	-1.549	0.417	-1.966	-1.132
3	3	-2.408	-2.666	1.318	-3.984	-1.348
4	1	-2.120	-3.120	2.689	-5.809	-0.432

2.2.2 Drawbacks

- A different formula for the count metameter, $\phi(n_k)$ is required for each discrete distribution.
- Systematic deviation from a linear relationship does not indicate which distribution provides a better fit.

2.3 Ord plots

An alternative plot suggested by Ord (1967) may be used to diagnose the form of the discrete distribution. Ord showed that a linear relationship of the form,

$$\frac{k\,p_k}{p_{k-1}} = a + b\,k \tag{3}$$

holds for each of the Poisson, binomial, negative binomial, and logarithmic series distributions. The slope, b, in (3) is zero for the Poisson, negative for the binomial, and positive for the negative binomial and logarithmic series distributions; the latter two are distinguished by their intercepts.

Thus, a plot of $k\,n_k/n_{k-1}$ against k, if linear, is suggested as a means to determine which distribution to apply.

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

2.3.1 Fitting the line

In the small number of cases I've tried, I have found that using a weighted least squares fit of $k\,n_k/n_{k-1}$ on k, using weights of $w_k = \sqrt{n_k - 1}$ produces reasonably good automatic diagnosis of the form of a probability distribution.

2.3.2 Examples

The table below shows the calculations for the horse kicks data, with the ratio $k\,p_k/p_{k-1}$ labeled y. The weighted least squares line, with weights w_k, has a slope close to zero, indicating the Poisson distribution. The estimate $\lambda = a = .656$ compares favorably with the value from the Poissonness plot.

Ord Plot: Deaths by Horsekicks

k	n_k	n_{k-1}	w_k	y	
0	109	.	10.3923	.	-- Weighted LS --
1	65	109	8.0000	0.5963	slope = -0.034
2	22	65	4.5826	0.6769	inter = 0.656
3	3	22	1.4142	0.4091	
4	1	3	0.0000	1.3333	

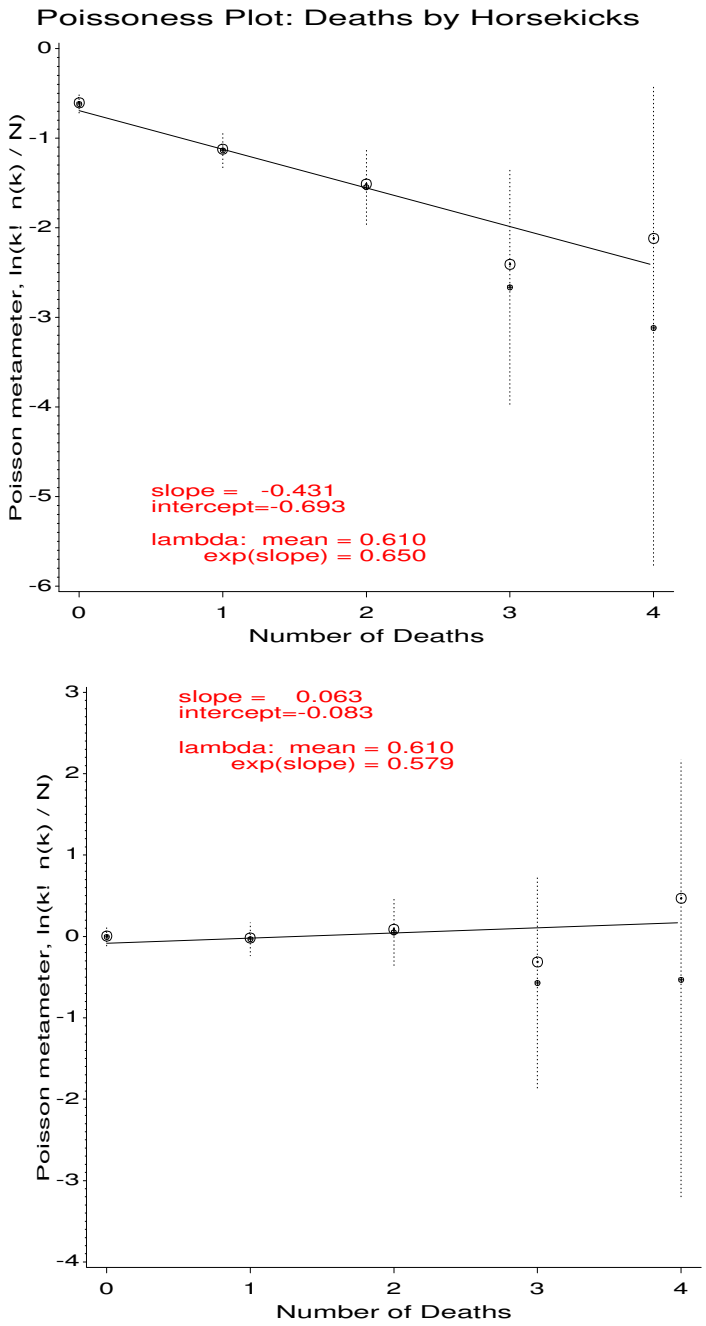


Figure 4: Poissonness plots for two discrete distributions. The Horse Kicks data fits the Poisson distribution reasonably well, but the May data does not.

For the word frequency data, the slope is positive, so either the negative binomial or log series are possible. The intercept is essentially zero, which is ambiguous. However, the logarithmic series requires $b \approx -a$, so the negative binomial is a better choice. Mosteller & Wallace did in fact find a reasonably good fit to this distribution.

Instances of 'may' in Federalist papers

k	n_k	n_{k-1}	w_k	y	
0	156	.	12.4499	.	-- Weighted LS --
1	63	156	7.8740	0.4038	slope = 0.424
2	29	63	5.2915	0.9206	inter = -0.023
3	8	29	2.6458	0.8276	
4	4	8	1.7321	2.0000	
5	1	4	0.0000	1.2500	
6	1	1	0.0000	6.0000	

Plots of data fitting several different discrete distributions are shown in Figure 5. In each case, the slope and intercept of the weighted least squares line correctly identifies the distribut

2.3.3 Drawbacks

- The Ord plot lacks resistance, since a single discrepant frequency affects the points for k and $k+1$.
- The sampling variance of $k\, n_k/n_{k-1}$ fluctuates widely. The use of weights w_k helps, but is purely a heuristic device.

2.3.4 The `ORDPLOT` macro

These plots are produced by the `ORDPLOT` macro. For the horsekicks data, the plot in Figure 5 is produced with the macro call

```
%ordplot(data=horskick,
         count=Deaths,label=Number of Deaths,freq=corpsyrs);
```

3 Tests of Association for Two-Way Tables

For two-way frequency tables, the typical analyses are based on Pearson's χ^2 (when the sample size is at least moderate: most expected frequencies 5 or more) or Fisher's exact test for small samples. These are tests of *general association*, where the null hypothesis is that the row and column variables are independent and alternative hypotheses is simply that the row and column variables are associated.

However, more powerful analyses are often available:

- When either the row or column variables are *ordinal*, tests which take the order into account are more specific and often have greater statistical power.
- When additional classification variables exist, it may be important to control for these variables, or to determine if the association between the row and column variables is the same (*homogeneous*) across the levels of the control variables.

Consider the data below, which compares a treatment for rheumatoid arthritis to a placebo (Koch and Edwards, 1988). The outcome reflects whether individuals showed no improvement, some improvement, or marked improvement.

		Outcome			
Treatment	Sex	None	Some	Marked	Total

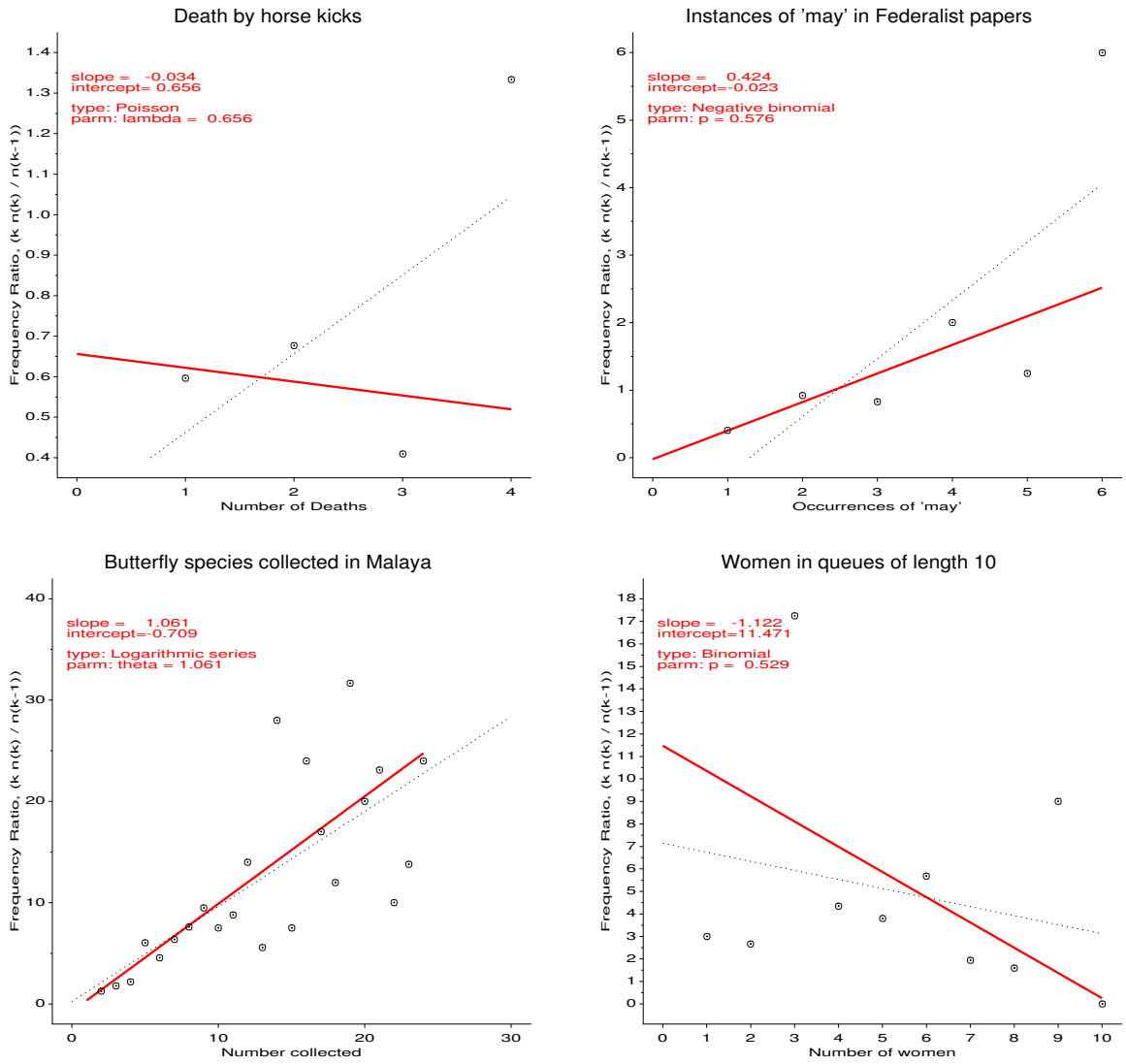


Figure 5: Ord plots for four discrete distributions. The slope and intercept of the weighted least squares line are used to identify the type of the distribution.

```
Active   |  Female |      6 |      5 |     16 |     27
         |  Male   |      7 |      2 |      5 |     14
---------+---------+--------+--------+--------+
Placebo  |  Female |     19 |      7 |      6 |     32
         |  Male   |     10 |      0 |      1 |     11
---------+---------+--------+--------+--------+
Total                   42       14       28       84
```

Here, the outcome variable is an ordinal one, and it is probably important to determine if the relation between treatment and outcome is the same for males and females.

3.1 Overall analysis

Since the main interest is in the relation between treatment and outcome, an overall analysis (which ignores sex) could be carried out using `PROC FREQ` as shown below.

```
title 'Arthritis Treatment: PROC FREQ Analysis';
data arth;
   input sex$ treat$ @;
   do improve = 'None  ', 'Some', 'Marked';
      input count @;
      output;
      end;
datalines;
Female  Active    6  5  16
Female  Placebo  19  7   6
Male    Active    7  2   5
Male    Placebo  10  0   1
;
*-- Ignoring sex;
proc freq order=data;
   weight count;
   tables treat * improve / cmh chisq nocol nopercent;
   run;
```

Notes:

- TREAT and IMPROVE are both character variables, which `PROC FREQ` orders alphabetically (i.e., 'Marked', 'None', 'Some') by default. Because I want to treat the IMPROVE variable as ordinal, I used `order=data` on the `PROC FREQ` statement to have the levels of IMPROVE ordered by their order of appearance in the dataset.
- The `chisq` option gives the usual χ^2 tests (Pearson, Fisher's, etc.). The `cmh` option requests the Cochran-Mantel-Haenszel tests for ordinal variables.

The output begins with the frequency table, including row percentages. The row percentages show a clear effect of treatment: for people given the Active treatment, 51% showed Marked improvement, while among those given the Placebo, 67% showed no improvement.

```
          TABLE OF TREAT BY IMPROVE
TREAT     IMPROVE

Frequency|
Row Pct  |None    |Some    |Marked  |  Total
---------+--------+--------+--------+
Active   |     13 |      7 |     21 |     41
         |  31.71 |  17.07 |  51.22 |
---------+--------+--------+--------+
Placebo  |     29 |      7 |      7 |     43
         |  67.44 |  16.28 |  16.28 |
---------+--------+--------+--------+
Total          42       14       28       84
```

The results for the `chisq` option is shown below. All tests show a significant association between treatment and outcome.

```
          STATISTICS FOR TABLE OF TREAT BY IMPROVE

Statistic                     DF      Value       Prob
------------------------------------------------------
Chi-Square                     2     13.055      0.001
Likelihood Ratio Chi-Square    2     13.530      0.001
Mantel-Haenszel Chi-Square     1     12.859      0.000
Phi Coefficient                      0.394
Contingency Coefficient              0.367
Cramer's V                           0.394
```

3.2 Tests for ordinal variables

For $r \times c$ tables, different tests are applicable depending on whether either or both of the row and column variables are ordinal. Tests which take the ordinal nature of a variable into account are provided by the `cmh` option on the `tables` statement. These tests are based on assigning numerical scores to the table categories; the default (table) scores treat the levels as equally spaced. They generally have higher power when the pattern of association is determined by the order of an ordinal variable.

For the arthritis data, these tests (`cmh` option) give the following output.

```
            SUMMARY STATISTICS FOR TREAT BY IMPROVE
   Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic   Alternative Hypothesis    DF       Value      Prob
-------------------------------------------------------------
   1        Nonzero Correlation        1      12.859     0.000
   2        Row Mean Scores Differ     1      12.859     0.000
   3        General Association        2      12.900     0.002
```

The three types of tests differ in the types of departure from independence they are sensitive to:

- ***General Association***. When the row and column variables are both nominal (unordered) the only alternative hypothesis of interest is that there is *some* association between the row and column variables. The CMH test statistic is similar to the (Pearson) Chi-Square and Likelihood Ratio Chi-Square in the Statistics table; all have $(r-1)(c-1)$ df.

- ***Mean score differences***. If the column variable is ordinal, assigning scores to the column variable produces a mean for each row. The association between row and column variables can be expressed as a test of whether these means differ over the rows of the table, with $r-1$ df. This is analogous to the Kruskal-Wallis non-parametric test (ANOVA based on rank scores).

- ***Linear association***. When **both** row and column variables are ordinal, we could assign scores to both variables and compute the correlation. The Mantel-Haenzsel χ^2 is equal to $(N-1)r^2$, where N is the total sample size. The test is most sensitive to a pattern where the row mean score changes linearly over the rows.

Notes:

- Different kinds of scores can be assigned using the `scores` option on the `tables` statement, but only the relative spacing of the scores is important.

- When only one variable is ordinal, make it the **last** one on the `tables` statement, because `PROC FREQ` only computes means across the column variable.

- When there are only r=2 rows (as here), the correlation and row means tests are equivalent.

3.3 Sample CMH Profiles

Two contrived examples may make the differences among these tests more apparent.

3.3.1 General Association

The table below exhibits a general association between variables A and B, but no difference in row means or linear association. (Figure 6 shows the pattern of association graphically.)

	b1	b2	b3	b4	b5	Total	Mean
a1	0	15	25	15	0	55	3.0
a2	5	20	5	20	5	55	3.0
a3	20	5	5	5	20	55	3.0
Total	25	40	35	40	25	165	

This is reflected in the PROC FREQ output:

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	0.000	1.000
2	Row Mean Scores Differ	2	0.000	1.000
3	General Association	8	91.797	0.000

3.3.2 Linear Association

This table contains a weak, non-significant general association, but significant row mean differences and linear associations (see Figure 7).

	b1	b2	b3	b4	b5	Total	Mean
a1	2	5	8	8	8	31	3.48
a2	2	8	8	8	5	31	3.19
a3	5	8	8	8	2	31	2.81
a4	8	8	8	5	2	31	2.52
Total	17	29	32	29	17	124	

Note that the χ^2-values for the row-means and non-zero correlation tests are very similar, but the correlation test is more highly significant.

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	10.639	0.001
2	Row Mean Scores Differ	3	10.676	0.014
3	General Association	12	13.400	0.341

The differences in sensitivity and power among these tests is analogous to the difference between general ANOVA tests and tests for linear trend in experimental designs with quantitative factors: The more specific test has greater power, but is sensitive to a narrower range of departure from the null hypothesis.

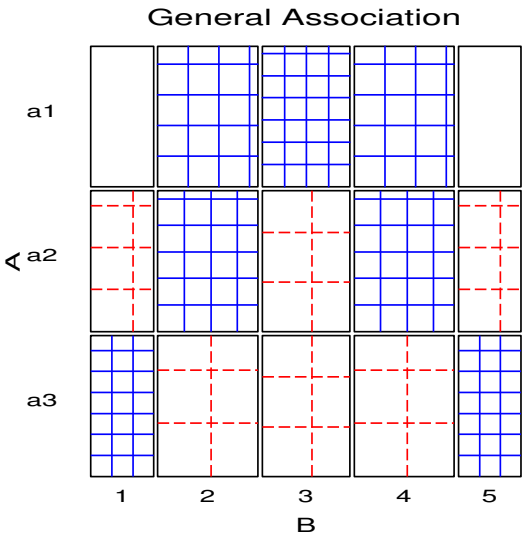


Figure 6: General association (sieve diagram)

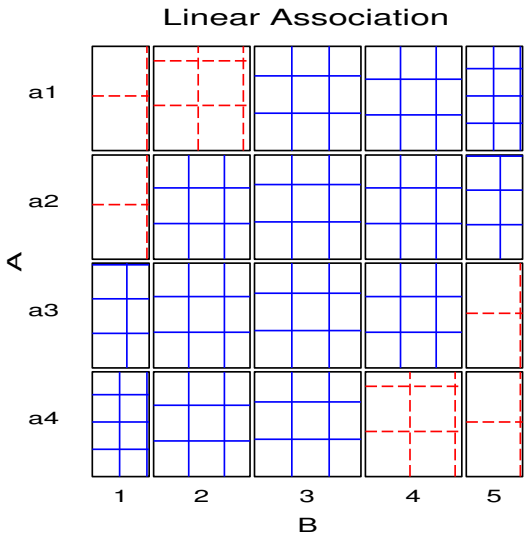


Figure 7: Linear association (sieve diagram)

3.4 Stratified Analysis

The overall analysis ignores other variables (like sex), by collapsing over them. It is possible that the treatment is effective only for one gender, or even that the treatment has opposite effects for men and women.

A stratified analysis:

- controls for the effects of one or more background variables. This is similar to the use of a blocking variable in an ANOVA design.
- is obtained by including more than two variables in the `tables` statement. List the stratification variables **first**. To examine the association between TREAT and IMPROVE, controlling for both SEX and AGE (if available):

```
tables age * sex * treat * improve;
```

The statements below request a stratified analysis with CMH tests, controlling for sex.

```
*-- Stratified analysis, controlling for sex;
proc freq order=data;
   weight count;
   tables sex * treat * improve / cmh chisq nocol nopercent;
   run;
```

PROC FREQ gives a separate table for each level of the stratification variables, plus overall (partial) tests controlling for the stratification variables.

```
          TABLE 1 OF TREAT BY IMPROVE
           CONTROLLING FOR SEX=Female

  TREAT     IMPROVE

  Frequency|
  Row Pct  |None    |Some    |Marked  |  Total
  ---------+--------+--------+--------+
  Active   |      6 |      5 |     16 |     27
           |  22.22 |  18.52 |  59.26 |
  ---------+--------+--------+--------+
```

```
     Placebo  |     19 |      7 |      6 |     32
              |  59.38 |  21.88 |  18.75 |
     ---------+--------+--------+--------+
     Total          25       12       22       59

      STATISTICS FOR TABLE 1 OF TREAT BY IMPROVE
              CONTROLLING FOR SEX=Female

Statistic                     DF     Value        Prob
------------------------------------------------------
Chi-Square                     2    11.296       0.004
Likelihood Ratio Chi-Square    2    11.731       0.003
Mantel-Haenszel Chi-Square     1    10.935       0.001
Phi Coefficient                      0.438
Contingency Coefficient              0.401
Cramer's V                           0.438
```

Note that the strength of association between treatment and outcome is quite strong for females. In contrast, the results for males (below) shows a non-significant association, even by the Mantel-Haenzsel test; but note that there are too few males for the general association χ^2 tests to be reliable (the statistic does not follow the theoretical χ^2 distribution).

```
             TABLE 2 OF TREAT BY IMPROVE
               CONTROLLING FOR SEX=Male

     TREAT     IMPROVE

     Frequency|
     Row Pct  |None    |Some    |Marked  |  Total
     ---------+--------+--------+--------+
     Active   |      7 |      2 |      5 |     14
              |  50.00 |  14.29 |  35.71 |
     ---------+--------+--------+--------+
     Placebo  |     10 |      0 |      1 |     11
              |  90.91 |   0.00 |   9.09 |
     ---------+--------+--------+--------+
     Total          17        2        6       25

      STATISTICS FOR TABLE 2 OF TREAT BY IMPROVE
               CONTROLLING FOR SEX=Male

Statistic                     DF     Value        Prob
------------------------------------------------------
Chi-Square                     2     4.907       0.086
Likelihood Ratio Chi-Square    2     5.855       0.054
Mantel-Haenszel Chi-Square     1     3.713       0.054
Phi Coefficient                      0.443
Contingency Coefficient              0.405
Cramer's V                           0.443

WARNING:  67% of the cells have expected counts less
           than 5. Chi-Square may not be a valid test.
```

The individual tables are followed by the (overall) partial tests of association controlling for sex. Unlike the tests for each strata, these tests **do not** require large sample size in the individual strata – just a large total sample size. Note that the χ^2 values here are slightly larger than those from the initial analysis that ignored sex.

```
                SUMMARY STATISTICS FOR TREAT BY IMPROVE
                          CONTROLLING FOR SEX

       Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

     Statistic   Alternative Hypothesis    DF       Value      Prob
     ------------------------------------------------------------
        1        Nonzero Correlation        1      14.632     0.000
        2        Row Mean Scores Differ     1      14.632     0.000
        3        General Association        2      14.632     0.001
```

3.4.1 Homogeneity of Association

In a stratified analysis it is often of interest to know if the association between the primary table variables is the same over all strata. For $k \times 2 \times 2$ tables this question reduces to whether the odds ratio is the same in all k strata, and `PROC FREQ` computes the Breslow-Day test for homogeneity when you use the `measures` option on the `tables` statement. `PROC FREQ` cannot perform tests of homogeneity for larger tables, but these can be easily done with the CATMOD procedure.

For the arthritis data, homogeneity means that there is no three-way sex * treatment * outcome association. This hypothesis can be stated as is the loglinear model,

$$[SexTreat][SexOutcome][TreatOutcome],$$

which allows associations between sex and treatment (e.g., more males get the Active treatment) and between sex and outcome (e.g. females are more likely to show marked improvement). In the `PROC CATMOD` step below, the `LOGLIN` statement specifies this log-linear model as `sex|treat|improve@2` which means "all terms up to 2-way associations".

```
title2 'Test homogeneity of treat*improve association';
data arth;
   set arth;
   if count=0 then count=1E-20;
proc catmod order=data;
   weight count;
   model sex * treat * improve = _response_ /
         ml noiter noresponse nodesign nogls ;
   loglin sex|treat|improve@2 / title='No 3-way association';
run;
   loglin sex treat|improve   / title='No Sex Associations';
```

(Frequencies of zero can be regarded as either "structural zeros"–a cell which could not occur, or as "sampling zeros"–a cell which simply did not occur. `PROC CATMOD` treats zero frequencies as "structural zeros", which means that cells with `count = 0` are excluded from the analysis. The DATA step above replaces the one zero frequency by a small number.)

In the output from `PROC CATMOD`, the likelihood ratio χ^2 (the badness-of-fit for the No 3-Way model) is the test for homogeneity across sex. This is clearly non-significant, so the treatment-outcome association can be considered to be the same for men and women.

```
   Test homogeneity of treat*improve association
               No 3-way association
   MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source                    DF   Chi-Square      Prob
-----------------------------------------------------
SEX                        1        14.13    0.0002
TREAT                      1         1.32    0.2512
SEX*TREAT                  1         2.93    0.0871
IMPROVE                    2        13.61    0.0011
SEX*IMPROVE                2         6.51    0.0386
TREAT*IMPROVE              2        13.36    0.0013

LIKELIHOOD RATIO           2         1.70    0.4267
```

Note that the associations of sex with treatment and sex with outcome are both small and of borderline significance, which suggests a stronger form of homogeneity, the log-linear model [Sex] [TreatOutcome] which says the only association is that between treatment and outcome. This model is tested by the second `loglin` statement given above, which produced the following output. The likelihood ratio test indicates that this model might provide a reasonable fit.

```
                No Sex Associations
   MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source                    DF   Chi-Square      Prob
-----------------------------------------------------
SEX                        1        12.95    0.0003
TREAT                      1         0.15    0.6991
IMPROVE                    2        10.99    0.0041
TREAT*IMPROVE              2        12.00    0.0025

LIKELIHOOD RATIO           5         9.81    0.0809
```

4 Plots for two-way frequency tables

Several schemes for representing contingency tables graphically are based on the fact that when the row and column variables are independent, the estimated expected frequencies, e_{ij}, are products of the row and column totals (divided by the grand total). Then, each cell can be represented by a rectangle whose area shows the cell frequency, f_{ij}, or deviation from independence.

4.1 Sieve diagrams

Table 1 shows data on the relation between hair color and eye color among 592 subjects (students in a statistics course) collected by Snee (1974). The Pearson χ^2 for these data is 138.3 with 9 degrees of freedom, indicating substantial departure from independence. The question is how to understand the *nature* of the association between hair and eye color.

For any two-way table, the expected frequencies under independence can be represented by rectangles whose widths are proportional to the total frequency in each column, f_{+j}, and whose heights are proportional to the total frequency in each row, f_{i+}; the area of each rectangle is then proportional to e_{ij}. Figure 8 shows the expected frequencies for the hair and eye color data.

Riedwyl and Schüpbach (1983, 1994) proposed a ***sieve diagram*** (later called a ***parquet diagram***) based on this principle. In this display the area of each rectangle is proportional to expected frequency and observed frequency is shown by the number of squares in each rectangle. Hence, the difference between observed and expected frequency appears as the density of shading, using color to indicate whether the deviation from independence is positive or

Table 1: Hair-color eye-color data

Eye Color	**Hair Color** Black	Brown	Red	Blond	Total
Green	5	29	14	16	64
Hazel	15	54	14	10	93
Blue	20	84	17	94	215
Brown	68	119	26	7	220
Total	108	286	71	127	592

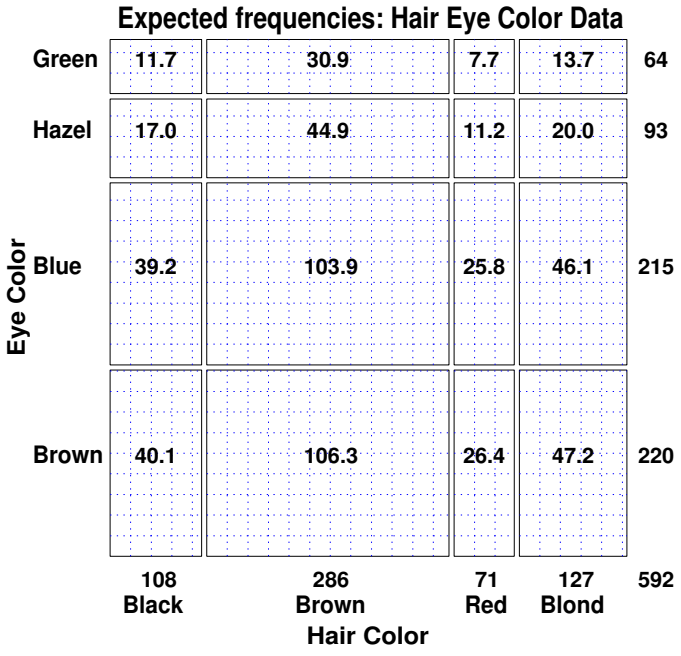


Figure 8: Expected frequencies under independence. Each box has area equal to its expected frequency, and is cross-ruled proportionally to the expected frequency.

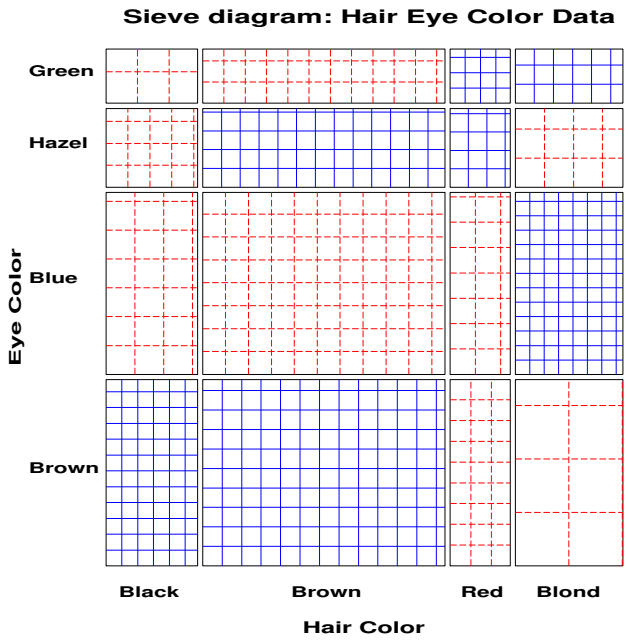


Figure 9: Sieve diagram for hair-color, eye-color data

negative. (In monochrome versions, positive deviations are shown by solid lines, negative by broken lines.) The sieve diagram for hair color and eye color is shown in Figure 9.

Figure 10 shows data on vision acuity in a large sample of women (n=7477). The diagonal cells show the obvious: people tend to have the same visual acuity in both eyes, and there is strong lack of indepence. The off diagonal cells show a more subtle pattern which suggests symmetry, and a diagonals model.

Figure 11 shows the frequencies with which draft-age men with birthdays in various months were assigned priority values for induction into the US Army in the 1972 draft lottery. The assignment was supposed to be random, but the figure shows a greater tendency for those born in the latter months of the year to be assigned smaller priority values.

4.2 Association plot for two-way tables

In the sieve diagram the foreground (rectangles) shows expected frequencies; deviations from independence are shown by color and density of shading. The association plot (Cohen, 1980, Friendly, 1991), puts deviations from independence in the foreground: the area of each box is made proportional to observed - expected frequency.

For a two-way contingency table, the signed contribution to Pearson χ^2 for cell i, j is

$$d_{ij} = \frac{f_{ij} - e_{ij}}{\sqrt{e_{ij}}} = \text{ std. residual} \qquad \chi^2 = \sum\sum_{ij} (d_{ij})^2$$

In the ***association plot***, each cell is shown by a rectangle:

- (signed) height $\sim d_{ij}$
- width = $\sqrt{e_{ij}}$.

so, the area of each cell is proportional to the raw residual, $f_{ij} - e_{ij}$. The rectangles for each row in the table are positioned relative to a baseline representing independence ($d_{ij} = 0$) shown by a dotted line. Cells with observed $>$ expected frequency rise above the line (and are colored black); cells that contain less than the expected frequency fall below it (and are shaded red); see Figure 12.

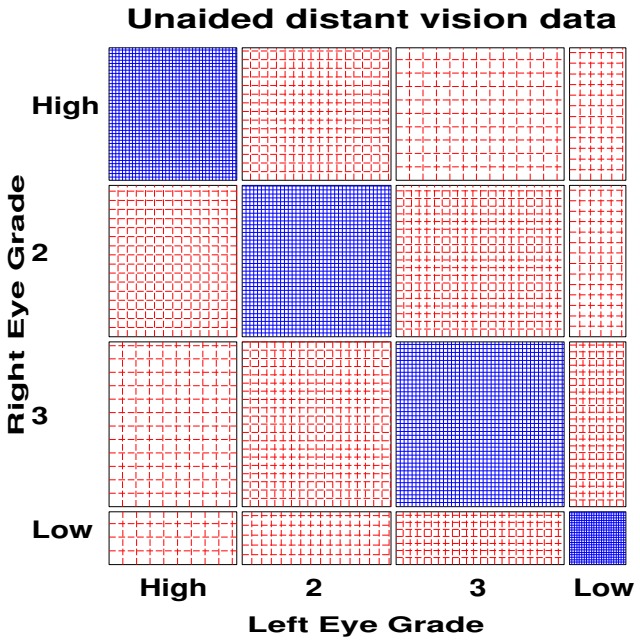


Figure 10: Vision classification data for 7477 women

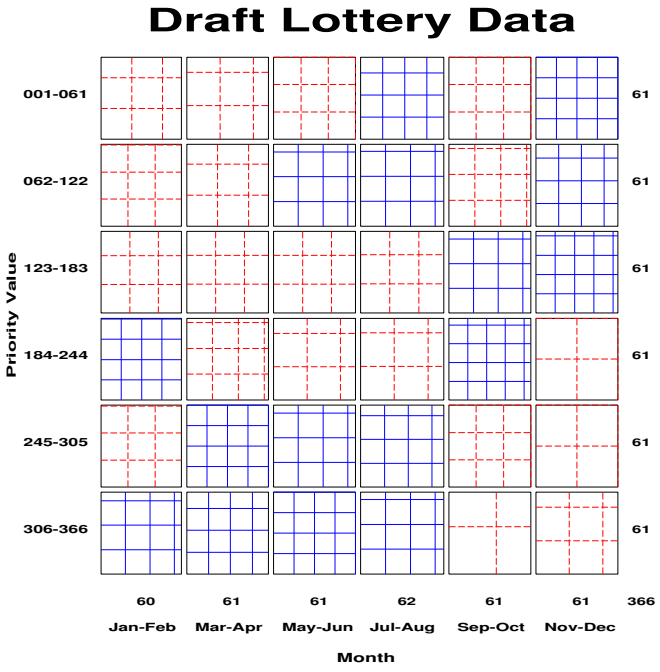


Figure 11: Data from the US Draft Lottery

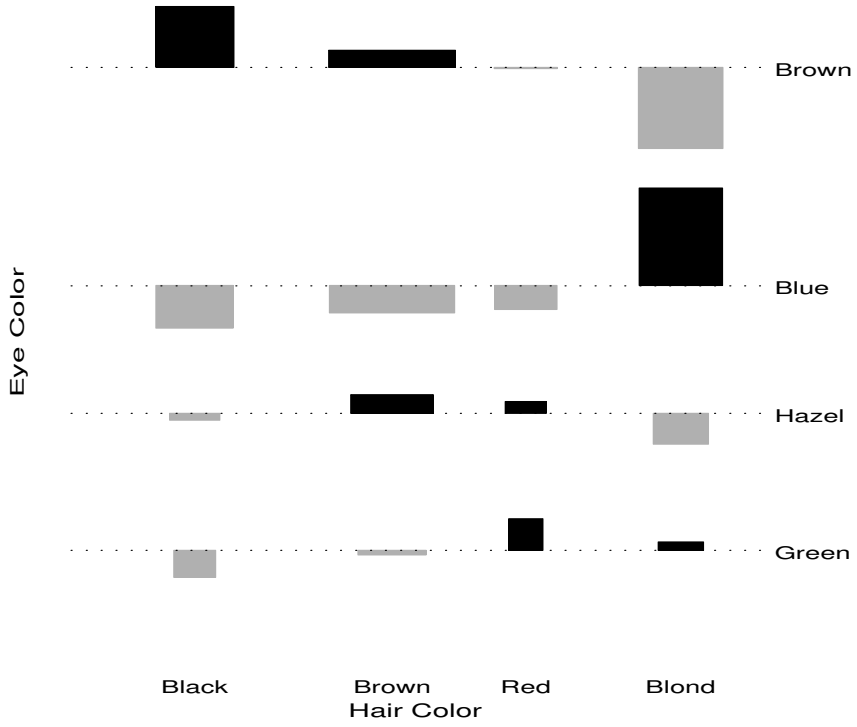


Figure 12: Association plot for hair-color, eye-color.

5 Portraying agreement: Observer Agreement chart

Inter-observer agreement is often used as a method of assessing the reliability of a subjective classification or assessment procedure. For example, two (or more) clinical psychologists might classify patients on a scale with categories: normal, mildly impaired, severely impaired.

5.1 Measuring agreement

- Strength of agreement vs. strength of association: Observers ratings can be strongly associated without strong agreement.
- Marginal homogeneity: If observers tend to use the categories with different frequency, this will affect measures of agreement.

5.1.1 Intraclass correlation

An analysis of variance framework leads to the **intraclass correlation** as a measure of inter-rater reliability, particularly when there are more than two raters. This approach is not covered here, but various applications are described by Shrout and Fleiss (1979).

5.1.2 Cohen's Kappa

A commonly used measure of agreement, Cohen's kappa (κ) compares the observed agreement, $P_o = \sum p_{ii}$, to agreement expected by chance if the two observer's ratings were independent, $P_c = \sum p_{i+}\, p_{+i}$.

$$\kappa = \frac{P_o - P_c}{1 - P_c} \tag{4}$$

κ has the following properties:

- For perfect agreement, $\kappa = 1$.

- The minimum κ may be < 0, and lower bound depends on marginal totals.
- Unweighted κ only counts strict agreement (same category assigned by both observers). A weighted version of κ is used when one wishes to allow for partial agreement. For example, exact agreements might be given full weight, one-category difference given weight 1/2. (This makes sense only when the categories are ordered, as in severity of diagnosis.)

5.1.3 Example

The table below summarizes responses of 91 married couples to a questionnaire item,

> Sex is fun for me and my partner (a) Never or occasionally, (b) fairly often, (c) very often, (d) almost always.

	Wife's Rating				
Husband's Rating	Never fun	Fairly often	Very Often	Almost always	SUM
Never fun	7	7	2	3	19
Fairly often	2	8	3	7	20
Very often	1	5	4	9	19
Almost always	2	8	9	14	33
SUM	12	28	18	33	91

Unweighted κ gives the following results

```
 Observed and Expected Agreement (under independence)

Observed agreement           0.3626
Expected agreement           0.2680

Cohen's Kappa (Std. Error)   0.1293  (0.1343)
```

Two commonly-used pattern of weights are those based on equal spacing of mismatch weights (Cicchetti-Alison weights), and ***Fleiss-Cohen weights***, based on an inverse-square spacing. The Fleiss-Cohen weights attach greater importance to near disagreements.

Integer Weights				Fleiss-Cohen Weights			
1	2/3	1/3	0	1	8/9	5/9	0
2/3	1	2/3	1/3	8/9	1	8/9	5/9
1/3	2/3	1	2/3	5/9	8/9	1	8/9
0	1/3	2/3	1	0	5/9	8/9	1

For the data at hand, these weights give a somewhat higher assessment of agreement (perhaps too high).

	Obs Agree	Exp Agree	Kappa	Std Error	Lower 95%	Upper 95%
Unweighted	0.363	0.268	0.1293	0.134	-0.1339	0.3926
Integer Weights	0.635	0.560	0.1701	0.065	0.0423	0.2978
Fleiss-Cohen Wts	0.814	0.722	0.3320	0.125	0.0861	0.5780

5.1.4 Computing Kappa with SAS

`PROC FREQ` provides the κ statistic with the `agree` option, as shown in the following example. The default weights for weighted κ are the inverse integer Cicchetti-Alison weights. Use `agree (wt=FC)` for Fleiss-Cohen weights. Use the `printkwt` option option to print the weights.

```
title 'Kappa for Agreement';
data fun;
   label husband = 'Husband rating'
         wife    = 'Wife Rating';
   do husband = 1 to 4;
   do wife    = 1 to 4;
      input count @@;
      output;
      end; end;
datalines;
 7     7     2      3
 2     8     3      7
 1     5     4      9
 2     8     9     14
;
proc freq;
   weight count;
   tables husband * wife / noprint agree;
run;
```

This produces the following output:

```
                        Kappa for Agreement
              STATISTICS FOR TABLE OF HUSBAND BY WIFE

                          Test of Symmetry
                          ----------------
       Statistic = 3.878        DF = 6        Prob = 0.693

                         Kappa Coefficients
       Statistic        Value     ASE   95% Confidence Bounds
       -------------------------------------------------------
       Simple Kappa     0.129   0.069      -0.005    0.264
       Weighted Kappa   0.237   0.078       0.084    0.391

       Sample Size = 91
```

5.1.5 Kappa for multiple strata— Overall agreement and homogeneity

When ratings are made for individuals from several populations, it it often of interest to test whether the agreement between raters differs over strata and to test the overall agreeement for all strata.

In SAS Version 6.11 and later, these tests are carried out when a stratified analysis is requested in the TABLES statement with the AGREE option.

In the example below, patients in New Orleans and in Winnipeg were diagnosed for multiple sclerosis by two neurologists, with ratings on a 1–4 scale from "certain" to "doubtful".

```
proc format;
     value rating 1="Certain MS" 2="Probable" 3="Possible" 4="Doubtful MS";
data msdiag;
     do patients='Winnipeg  ', 'New Orleans';
        do N_rating = 1 to 4;
           do W_rating = 1 to 4;
              input count @;
              output;
              end;
           end;
        end;
   format N_rating W_rating rating.;
```

```
   label N_rating = 'New Orleans neurologist'
         W_rating = 'Winnipeg nurologist';
cards;
38  5  0  1     33 11  3  0     10 14  5  6     3  7  3 10
 5  3  0  0      3 11  4  0      2 13  3  4     1  2  4 14
;
```

Using `PATIENTS` as a stratifying variable, `PROC FREQ` calculates κ for each patient population separately, as well as for combined, controlling for population.

```
proc freq data=msdiag;
     weight count;
     tables patients * N_rating * W_rating / norow nocol nopct agree;
run;
```

The output from this analysis (excluding the frequency tables) appears in Output 5.1. The tests of symmetry (using Bowker's test) test whether non-agreements are equally likely to be in higher or lower rating categories; this hypothesis is rejected for the Winnipeg patients. The individual values of κ are both significantly greater than zero (from their 95% confidence intervals), and there is no evidence that either patient sample is rated more reliably.

5.2 Bangdiwala's Observer Agreement Chart

The observer agreement chart Bangdiwala (1987) provides a simple graphic representation of the strength of agreement in a contingency table, and a measure of strength of agreement with an intuitive interpretation.

The agreement chart is constructed as an $n \times n$ square, where n is the total sample size. Black squares, each of size $n_{ii} \times n_{ii}$, show observed agreement. These are positioned within larger rectangles, each of size $n_{i+} \times n_{+i}$ as shown in Figure 13. The large rectangle shows the maximum possible agreement, given the marginal totals. Thus, a visual impression of the strength of agreement is

$$B_N = \frac{\text{area of dark squares}}{\text{area of rectangles}} = \frac{\sum_i^k n_{ii}^2}{\sum_i^k n_{i+}\, n_{+i}} \tag{5}$$

5.2.1 Partial agreement

Partial agreement is allowed by including a weighted contribution from off-diagonal cells, b steps from the main diagonal.

$$\begin{array}{ccccccccccc} & & & & n_{i-b,i} & & & & & & \\ & & & & \vdots & & & & & & \\ n_{i,i-b} & \cdots & & & n_{i,i} & & \cdots & & n_{i,i+b} & & \\ & & & & \vdots & & & & & & \\ & & & & n_{i-b,i} & & & & & & \end{array} \qquad \begin{array}{ccccc} & & w_2 & & \\ & & w_1 & & \\ w_2 & w_1 & 1 & w_1 & w_2 \\ & & w_1 & & \\ & & w_2 & & \end{array}$$

This is incorporated in the agreement chart (Figure 14) by successively lighter shaded rectangles whose size is proportional to the sum of the cell frequencies, denoted A_{bi}, shown schematically above. A_{1i} allows 1-step disagreements, A_{2i} includes 2-step disagreements, etc. From this, one can define a weighted measure of agreement, analogous to weighted κ.

$$B_N^w = \frac{\text{weighted sum of areas of agreement}}{\text{area of rectangles}} = 1 - \frac{\sum_i^k \left[n_{i+}n_{+i} - n_{ii}^2 - \sum_{b=1}^q w_b A_{bi}\right]}{\sum_i^k n_{i+}\, n_{+i}}$$

where w_b is the weight for A_{bi}, the shaded area b steps away from the main diagonal, and q is the furthest level of partial disagreement to be considered.

Output 5.1 Cohen's κ for MS diagnosis, stratified by patients

```
      STATISTICS FOR TABLE 1 OF N_RATING BY W_RATING
            CONTROLLING FOR PATIENTS=New Orlean

                    Test of Symmetry
                    ----------------
Statistic = 9.765        DF = 6        Prob = 0.135

                   Kappa Coefficients

Statistic        Value     ASE   95% Confidence Bounds
-------------------------------------------------------
Simple Kappa     0.297   0.079       0.143    0.450
Weighted Kappa   0.477   0.073       0.334    0.620

Sample Size = 69

    STATISTICS FOR TABLE 2 OF N_RATING BY W_RATING
           CONTROLLING FOR PATIENTS=Winnipeg

                    Test of Symmetry
                    ----------------
Statistic = 46.749       DF = 6        Prob = 0.001

                   Kappa Coefficients

Statistic        Value     ASE   95% Confidence Bounds
-------------------------------------------------------
Simple Kappa     0.208   0.050       0.109    0.307
Weighted Kappa   0.380   0.052       0.278    0.481

Sample Size = 149

      SUMMARY STATISTICS FOR N_RATING BY W_RATING
               CONTROLLING FOR PATIENTS

              Overall Kappa Coefficients

Statistic        Value     ASE   95% Confidence Bounds
-------------------------------------------------------
Simple Kappa     0.234   0.042       0.151    0.317
Weighted Kappa   0.412   0.042       0.330    0.495

          Tests for Equal Kappa Coefficients

                     Chi-Square        DF         Prob
-------------------------------------------------------
Simple Kappa              0.901         1        0.343
Weighted Kappa            1.189         1        0.276

Total Sample Size = 218
```

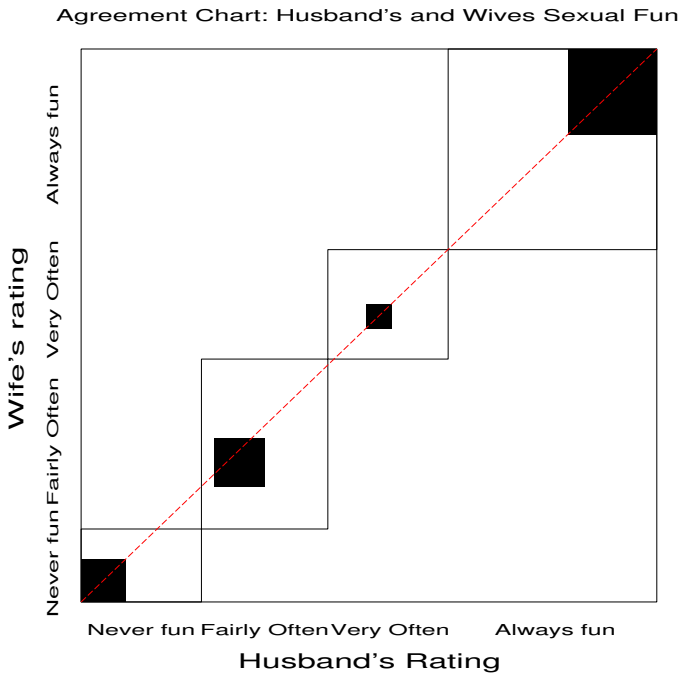


Figure 13: Agreement chart for husbands and wives sexual fun. The B_N measure (5) is the ratio of the areas of the dark squares to their enclosing rectangles, counting only exact agreement. $B_N = 0.146$ for these data.

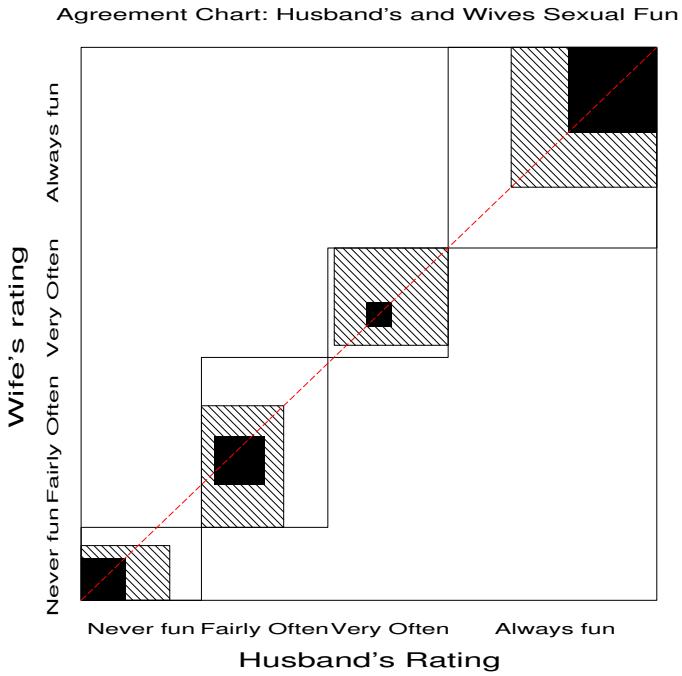


Figure 14: Weighted agreement chart. The B_N^w measure is the ratio of the areas of the dark squares to their enclosing rectangles, weighting cells one step removed from exact agreement with $w_1 = 8/9 = .889$. $B_N^w = 0.628$ for these data.

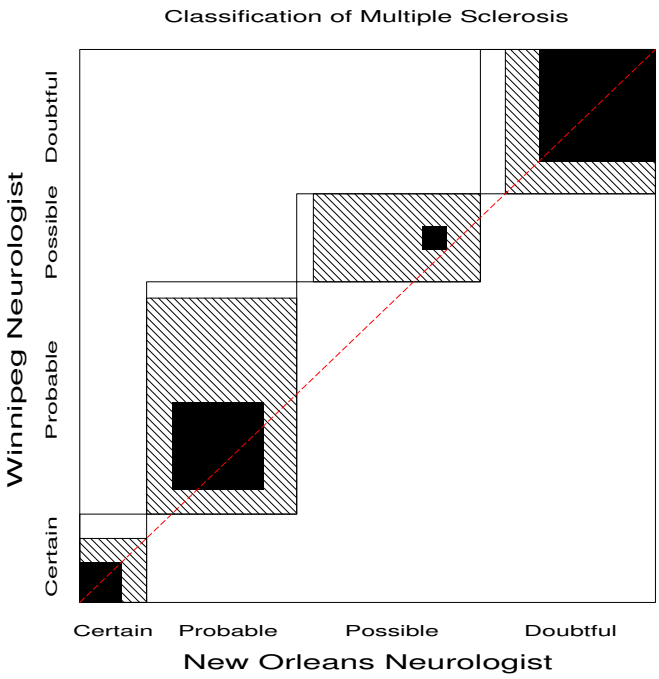


Figure 15: Weighted agreement chart. Departure of the middle squares from the diagonal indicates lack of marginal homogeneity.

5.3 Observer bias

With an ordered scale, it may happen that one observer consistently tends to classify the objects into higher or lower categories than the other. This produces differences in the marginal totals, n_{i+}, and n_{+i}. While special tests exist for ***marginal homogeneity***, the observer agreement chart shows this directly by the relation of the dark squares to the diagonal line: When the marginal totals are the same, the squares fall along the diagonal.

5.3.1 Example

The table below shows the classification of 69 New Orleans patients regarding multiple sclerosis diagnosis by neurologists in New Orleans and Winnipeg. The agreement chart (Figure 15) shows the two intermediate categories lie largely above the line, indicating that the Winnipeg neurologist tends to classify patients into more severe diagnostic categories.

New Orleans	Winnipeg Neurologist				
Neurologist	Certain	Probable	Possible	Doubtful	SUM
Certain MS	5	3	0	0	8
Probable MS	3	11	4	0	18
Possible MS	2	13	3	4	22
Doubtful MS	1	2	4	14	21
SUM	11	29	11	18	69

5.3.2 Testing marginal homogeneity

We can test the hypothesis that the marginal totals in the four diagnostic categories are equal for both neurologists using the CATMOD procedure. The following statements read the frequencies, creating a data set `ms` with variables `win_diag` and `no_diag` for the diagnostic categories assigned by the Winnepeg and New Orleans neurologists,

respectively. Note that zero frequencies are changed to a small number so that CATMOD will not treat them as structural zeros and eliminate these cells from the table.

```
title "Classification of Multiple Sclerosis: Marginal Homogeneity";
proc format;
   value diagnos 1='Certain ' 2='Probable'  3='Possible'  4='Doubtful';

data ms;
 format win_diag no_diag diagnos.;
   do win_diag = 1 to 4;
   do no_diag  = 1 to 4;
      input count @@;
      if count=0 then count=1e-10;
      output;
      end; end;
datalines;
   5     3     0      0
   3    11     4      0
   2    13     3      4
   1     2     4     14
;
```

In this analysis the diagnostic categories for the two neurologists are repeated measures, since each patient is rated twice. To test whether the marginal frequencies of ratings is the same we specify `response marginals`. (The `oneway` option displays the marginal frequencies, not shown here.)

```
title "Classification of Multiple Sclerosis: Marginal Homogeneity";
proc catmod data=ms;
   weight count;
   response marginals;
   model win_diag * no_diag = _response_ / oneway;
   repeated neuro 2 / _response_= neuro;
```

The test of marginal homogeneity is the test of `NEURO` in this model:

ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	3	222.62	0.0000
NEURO	3	10.54	0.0145
RESIDUAL	0	.	.

Because the diagnostic categories are ordered, we can actually obtain a more powerful test by assigning scores to the diagnostic category and testing if the mean scores are the same for both neurologists. To do this, we specify `response means`.

```
title2 'Testing means';
proc catmod data=ms order=data;
   weight count;
   response means;
   model win_diag * no_diag = _response_;
   repeated neuro 2 / _response_= neuro;
```

```
                ANALYSIS-OF-VARIANCE TABLE

Source                    DF   Chi-Square      Prob
-------------------------------------------------
INTERCEPT                  1       570.61    0.0000
NEURO                      1         7.97    0.0048

RESIDUAL                   0            .         .
```

5.4 Four-fold display for 2 x 2 tables

For a 2×2 table, the departure from independence can be measured by the sample *odds ratio*, $\theta = (f_{11}/f_{12}) \div (f_{21}/f_{22})$. The ***four-fold display*** (Friendly, 1994a,c) shows the frequencies in a 2×2 table in a way that depicts the odds ratio. In this display the frequency in each cell is shown by a quarter circle, whose radius is proportional to $\sqrt{f_{ij}}$, so again area is proportional to count. An association between the variables (odds ratio $\neq 1$) is shown by the tendency of diagonally opposite cells in one direction to differ in size from those in the opposite direction, and we use color and shading to show this direction. If the marginal proportions in the table differ markedly, the table may first be standardized (using iterative proportional fitting) to a table with equal margins but the same odds ratio.

```
              RAW DATA              STANDARDIZED
            Admit    Reject         Admit  Reject
          ----------------------------------------
Male       1198     1493            57.6  42.4     100
Female      557     1278            42.4  57.6     100
                                   --------------
                                    100   100

 Odds ratio:   1.84                   1.84
```

Figure 16 shows aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and gender. At issue is whether the data show evidence of sex bias in admission practices (Bickel et al., 1975). The figure shows the cell frequencies numerically, but margins for both sex and admission are equated in the display. For these data the sample odds ratio, Odds (Admit|Male) / (Admit|Female) is 1.84 indicating that males are almost twice as likely in this sample to be admitted. The four-fold display shows this imbalance clearly.

5.4.1 Confidence rings for the odds ratio

The fourfold display is constructed so that the four quadrants will align vertically and horizontally when the odds ratio is 1. Confidence rings for the observed odds ratio provide a visual test of the hypothesis of no association ($H_0 : \theta = 1$). They have the property that rings for adjacent quadrants overlap iff the observed counts are consistent with this null hypothesis.

The 99% confidence intervals in Figure 16 do not overlap, indicating a significant association between sex and admission. The width of the confidence rings give a visual indication of the precision of the data.

5.4.2 2 x 2 x k tables

In a $2 \times 2 \times k$ table, the last dimension often corresponds to "strata" or populations, and it is typically of interest to see if the association between the first two variables is homogeneous across strata. For such tables, simply make one fourfold panel for each strata. The standardization of marginal frequencies is designed allow easy visual comparison of the pattern of association across two or more populations.

The admissions data shown in Figure 16 were obtained from six departments, so to determine the source of the apparent sex bias in favor of males, we make a new plot, Figure 17, stratified by department.

Surprisingly, Figure 17 shows that, for five of the six departments, the odds of admission is approximately the same for both men and women applicants. Department A appears to differs from the others, with women approximately 2.86 ($= (313/19)/(512/89)$) times as likely to gain admission. This appearance is confirmed by the confidence rings, which in Figure 17 are joint 99% intervals for θ_c, $c = 1, \ldots, k$.

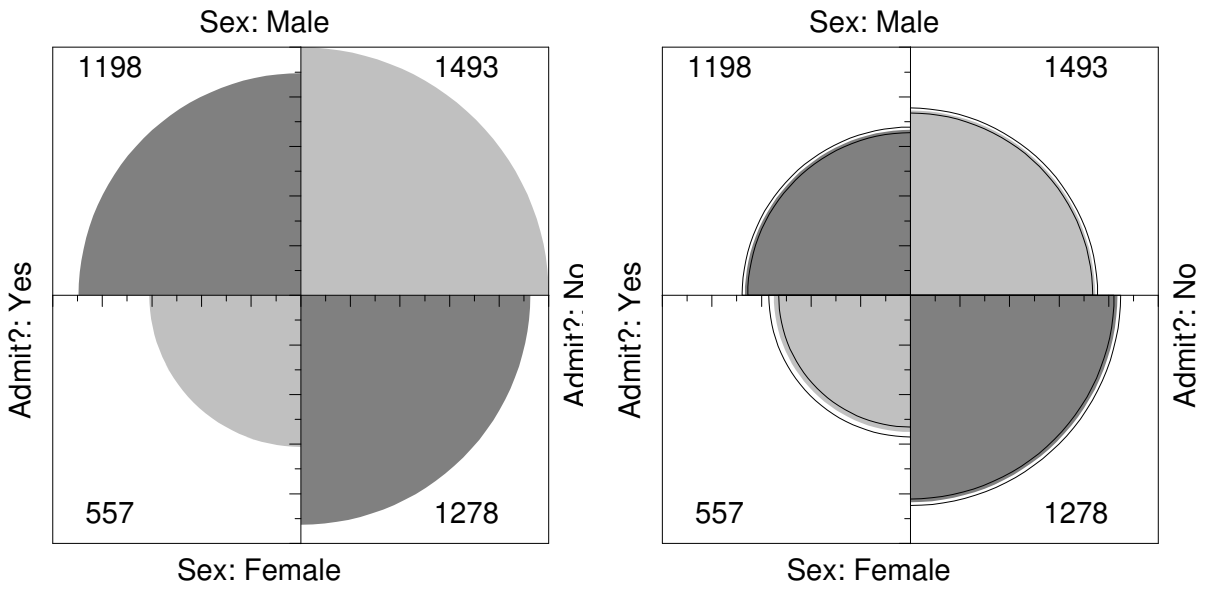


Figure 16: Four-fold display for Berkeley admissions. The area of each shaded quadrant shows the frequency, raw frequencies in the left panel, standardized to equate the margins for sex and admission in the right panel. Circular arcs show the limits of a 99% confidence interval for the odds ratio.

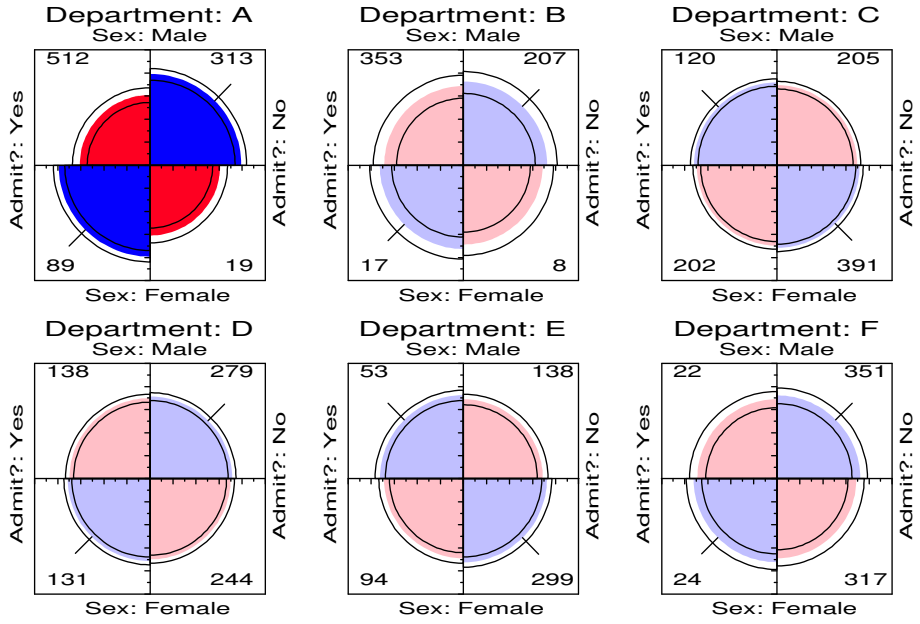


Figure 17: Fourfold display of Berkeley admissions, by department. In each panel the confidence rings for adjacent quadrants overlap if the odds ratio for admission and sex does not differ significantly from 1. The data in each panel have been standardized as in Figure 16.

(This result, which contradicts the display for the aggregate data in Figure 16, is a classic example of Simpson's paradox. The resolution of this contradiction can be found in the large differences in admission rates among departments as we shall see shortly.)

6 Mosaic displays for n-way tables

The mosaic display (Hartigan and Kleiner, 1981, 1984, Friendly, 1994b), represents the counts in a contingency table directly by tiles whose size is proportional to the cell frequency. This display:

- generalizes readily to n-way tables
- can be used to display the deviations from various log-linear models.

In the ***column proportion mosaic***, the width of each box is proportional to the total frequency in each column of the table. The height of each box is proportional to the cell frequency, and the dotted line in each row shows the expected frequencies under independence. Thus the deviations from independence, $f_{ij} - e_{ij}$, are shown by the areas between the rectangles and the dotted lines for each cell.

6.1 Condensed mosaic

The amount of empty space inside the mosaic plot may make it harder to see patterns, especially when there are large deviations from independence. In these cases, it is more useful to separate the rectangles in each column by a small constant space, rather than forcing them to align in each row. This is done in the ***condensed mosaic display***. Again, the area of each box is proportional to the cell frequency, and complete independence is shown when the tiles in each row all have the same height.

6.1.1 Detecting patterns

In Hartigan & Kleiner's original version (Hartigan and Kleiner, 1981), all the tiles are unshaded and drawn in one color, so only the relative sizes of the rectangles indicate deviations from independence. We can increase the visual impact of the mosaic by:

- using color and shading to reflect the size of the residual and
- reordering rows and columns to make the pattern more coherent.

6.2 Multi-way tables

The condensed form of the mosaic plot generalizes readily to the display of multi-dimensional contingency tables. Imagine that each cell of the two-way table for hair and eye color is further classified by one or more additional variables–sex and level of education, for example. Then each rectangle can be subdivided horizontally to show the proportion of males and females in that cell, and each of those horizontal portions can be subdivided vertically to show the proportions of people at each educational level in the hair-eye-sex group.

6.3 Fitting models

When three or more variables are represented in the mosaic, we can fit several different models of independence and display the residuals from that model. We treat these models as null or baseline models, which may not fit the data particularly well. The deviations of observed frequencies from expected, displayed by shading, will often suggest terms to be added to to an explanatory model which achieves a better fit.

- **Complete independence**: The model of complete independence asserts that all joint probabilities are products of the one-way marginal probabilities:

$$\pi_{ijk} = \pi_{i++}\, \pi_{+j+}\, \pi_{++k} \tag{6}$$

  for all i, j, k in a three-way table. This corresponds to the log-linear model $[A]\,[B]\,[C]$. Fitting this model puts all higher terms, and hence all association among the variables into the residuals.

- **Joint independence**: Another possibility is to fit the model in which variable C is jointly independent of variables A and B,

$$\pi_{ijk} = \pi_{ij+}\, \pi_{++k}. \tag{7}$$

  This corresponds to the log-linear model is $[AB][C]$. Residuals from this model show the extent to which variable C is related to the combinations of variables A and B but they do not show any association between A and B.

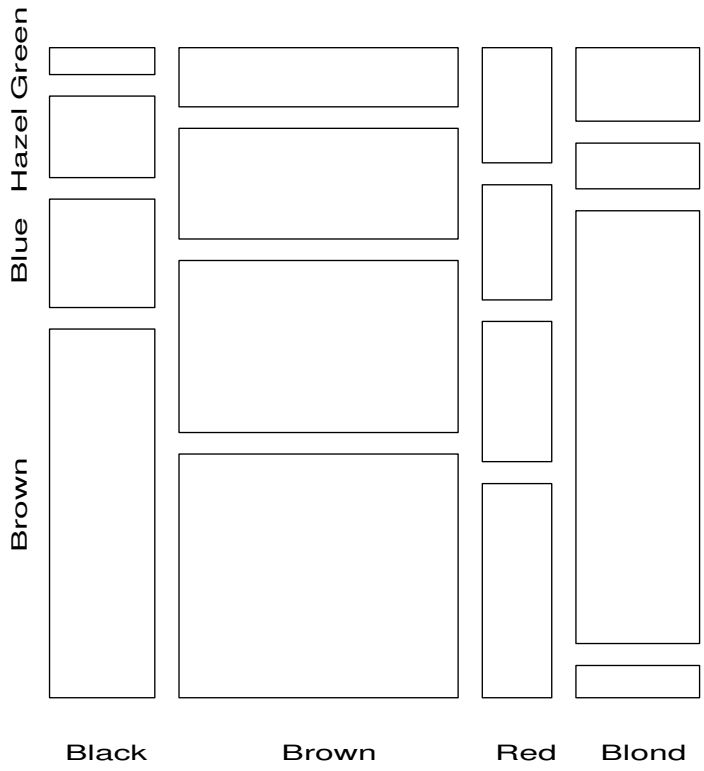


Figure 18: Condensed mosaic for Hair-color, Eye-color data. Each column is divided according to the conditional frequency of eye color given hair color. The area of each rectangle is proportional to observed frequency in that cell.

6.4 Sequential plots and models

The series of mosaic plots fitting models of joint independence to the marginal subtables can be viewed as partitioning the hypothesis of complete independence in the full table.

For a three-way table, the the hypothesis of complete independence, $H_{A\otimes B\otimes C}$ can be expressed as

$$H_{A\otimes B\otimes C} = H_{A\otimes B} \cap H_{AB\otimes C}, \tag{8}$$

where $H_{A\otimes B}$ denotes the hypothesis that A and B are independent in the marginal subtable formed by collapsing over variable C, and $H_{AB\otimes C}$ denotes the hypothesis of joint independence of C from the AB combinations. When expected frequencies under each hypothesis are estimated by maximum likelihood, the likelihood ratio G^2s are additive:

$$G^2_{A\otimes B\otimes C} = G^2_{A\otimes B} + G^2_{AB\otimes C}. \tag{9}$$

For example, for the hair-eye data, the mosaic displays for the Hair Eye marginal table (Figure 19) and the [HairEye] [Sex] table (Figure 20) can be viewed as representing the partition

```
Model                      df        G

Hair Eye                9       146.44
[Hair, Eye] [Sex]          15        19.86
-----------------------------------------
[Hair] [Eye] [Sex]         24       166.60
```

The partitioning scheme in (9) extends readily to higher-way tables.

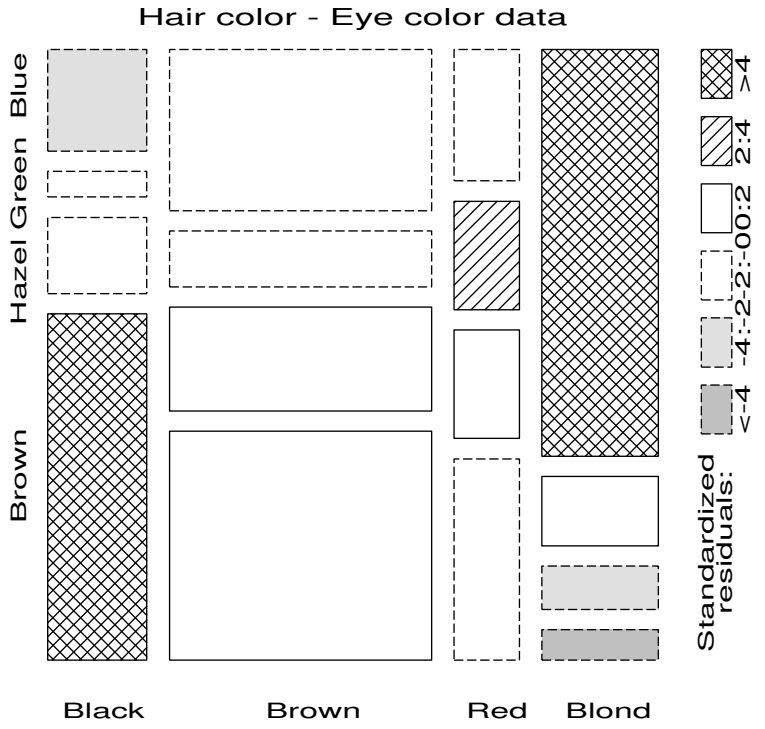


Figure 19: Condensed mosaic, reordered and shaded. Deviations from independence are shown by color and shading. The two levels of shading density correspond to standardized deviations greater than 2 and 4 in absolute value. This form of the display generalizes readily to multi-way tables.

7 Correspondence analysis

Correspondence analysis is an exploratory technique related to to principal components analysis which finds a multidimensional representation of the association between the row and column categories of a two-way contingency table. This technique finds scores for the row and column categories on a small number of dimensions which account for the greatest proportion of the χ^2 for association between the row and column categories, just as principal components account for maximum variance. For graphical display two or three dimensions are typically used to give a reduced rank approximation to the data.

For a two-way table the scores for the row categories, namely x_{im}, and column categories, y_{jm}, on dimension $m = 1, \ldots, M$ are derived from a singular value decomposition of residuals from independence, expressed as $d_{ij}/\sqrt{n}$, to account for the largest proportion of the χ^2 in a small number of dimensions. This decomposition may be expressed as

$$\frac{f_{ij} - e_{ij}}{\sqrt{n\, e_{ij}}} = \sum_{m=1}^{M} \lambda_m \, x_{im} \, y_{jm}, \tag{10}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$, and $M = \min(I-1,\, J-1)$. In M dimensions, the decomposition (10) is exact. A rank-d approximation in d dimensions is obtained from the first d terms on the right side of (10), and the proportion of χ^2 accounted for by this approximation is

$$n \sum_{m}^{d} \lambda_m^2 / \chi^2$$

Thus, correspondence analysis is designed to show how the data deviate from expectation when the row and column variables are independent, as in the association plot and mosaic display. However, the association plot and mosaic display depict every *cell* in the table, and for large tables it may be difficult to see patterns. Correspondence

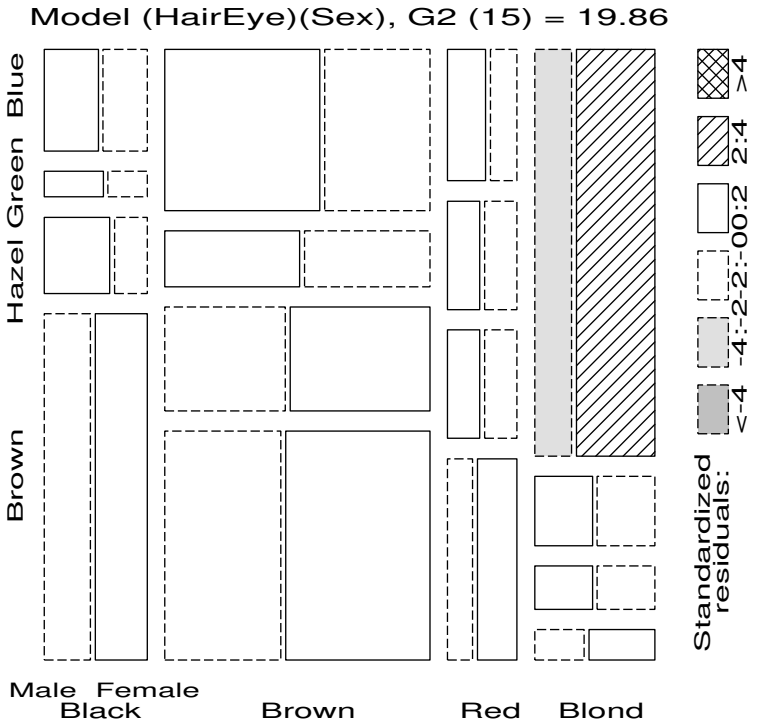


Figure 20: Three-way mosaic display for hair color, eye color, and sex. The categories of sex are crossed with those of hair color, but only the first occurrence is labeled. Residuals from the model of joint independence, $[HE]\,[S]$ are shown by shading. G^2 = 19.86 on 15 df. The lack of fit is attributable mainly to an overabundance of females among blue-eyed blonds.

analysis shows only row and column *categories* in the two (or three) dimensions which account for the greatest proportion of deviation from independence.

7.1 PROC CORRESP

In SAS Version 6, correspondence analysis is performed using `PROC CORRESP` in SAS/STAT. `PROC CORRESP` can read two kinds of input:

- a two-way contingency table
- raw category responses on two or more classification variables

An OUTC= data set from `PROC CORRESP` contains the row and column coordinates, which can be plotted with `PROC PLOT` or `PROC GPLOT`. The procedure has many options for scaling row and column coordinates, and for printing various statistics which aid interpretation. Only the basic use of the procedure is illustrated here.

7.1.1 Example: Hair and Eye Color

The program below reads the hair and eye color data into the data set COLORS, and calls the CORRESP procedure. This example illustrates the use of `PROC PLOT` and the Annotate facility with `PROC GPLOT` to produce a labeled display of the correspondence analysis solution. To input a contingency table in the CORRESP step, the hair colors (columns) are specified as the variables in the VAR statement, and the eye colors (rows) are indicated as the ID variable.

```
data colors;
```

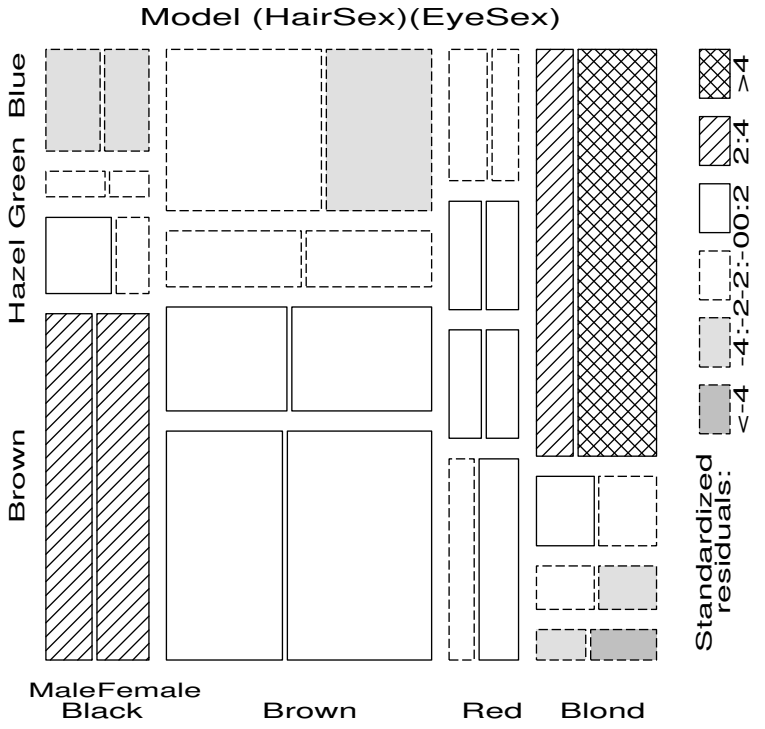


Figure 21: Mosaic display for hair color, eye color, and sex. This display shows residuals from the model of complete independence, $[H]\,[E]\,[S]$, G^2 = 166.6 on 24 df.

```
   input BLACK BROWN RED BLOND   EYE$;
   datalines;
           68  119   26    7     Brown
           20   84   17   94     Blue
           15   54   14   10     Hazel
            5   29   14   16     Green
;
proc corresp data=colors out=coord short;
    var black brown red blond;
    id eye;
proc print data=coord;
```

The printed output from the CORRESP procedure is shown below. The section labeled “Inertia, ... ” indicates that over 98% of the χ^2 for association is accounted for by two dimensions, with most of that attributed to the first dimension.

```
                The Correspondence Analysis Procedure

                 Inertia and Chi-Square Decomposition

   Singular  Principal Chi-
   Values    Inertias  Squares Percents   18   36   54   72   90
                                        ----+----+----+----+----+---
   0.45692   0.20877   123.593  89.37% *************************
   0.14909   0.02223    13.158   9.51% ***
   0.05097   0.00260     1.538   1.11%
             -------   -------
             0.23360    138.29 (Degrees of Freedom = 9)

                          Row Coordinates
                                Dim1          Dim2

                Brown       -.492158      -.088322
                Blue        0.547414      -.082954
                Hazel       -.212597      0.167391
                Green       0.161753      0.339040

                         Column Coordinates
                                Dim1          Dim2

                BLACK       -.504562      -.214820
                BROWN       -.148253      0.032666
                RED         -.129523      0.319642
                BLOND       0.835348      -.069579
```

The singular values, λ_i, in Eqn. (10), are also the (canonical) correlations between the optimally scaled categories. Thus, if the DIM1 scores for hair color and eye color are assigned to the 592 observations in the table, the correlation of these variables would be 0.4569. The DIM2 scores give a second, orthogonal scaling of these two categorical variables, whose correlation would be 0.1491.

A plot of the row and column points can be constructed from the OUT= data set COORD requested in the `PROC CORRESP` step. The variables of interest in this example are shown in below. Note that row and column points are distinguished by the variable `_TYPE_`.

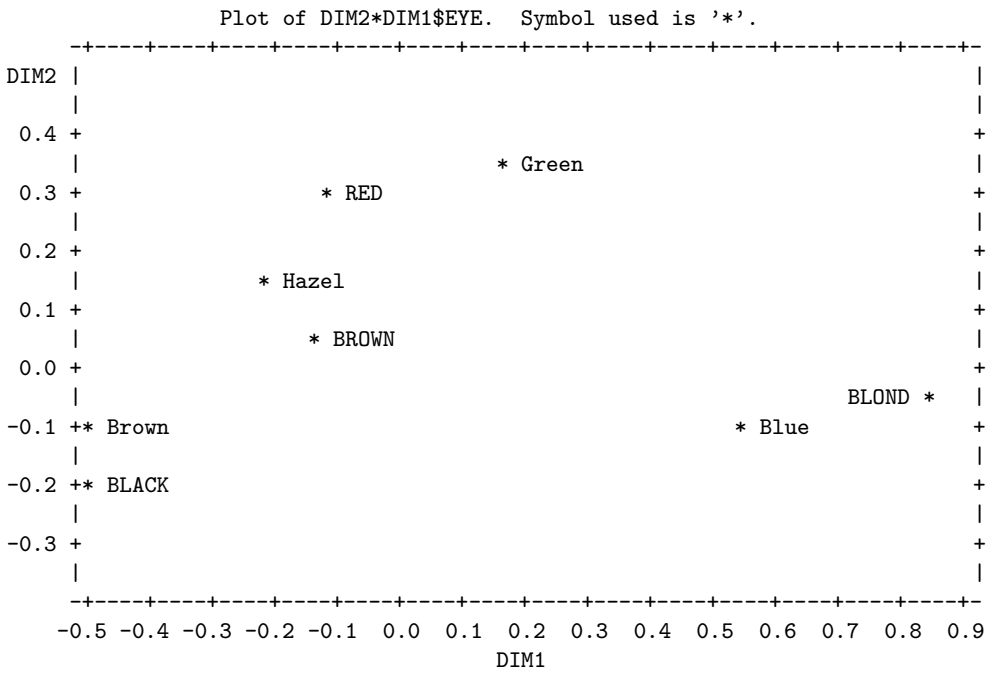
```
 OBS    _TYPE_      EYE       DIM1        DIM2

  1     INERTIA                .           .
  2     OBS         Brown    -0.49216    -0.08832
  3     OBS         Blue      0.54741    -0.08295
  4     OBS         Hazel    -0.21260     0.16739
  5     OBS         Green     0.16175     0.33904
  6     VAR         BLACK    -0.50456    -0.21482
  7     VAR         BROWN    -0.14825     0.03267
  8     VAR         RED      -0.12952     0.31964
  9     VAR         BLOND     0.83535    -0.06958
```

The interpretation of the correspondence analysis results is facilitated by a *labelled* plot of the row and column points. As of Version 6.08, points can be labeled in `PROC PLOT`. The following statements produce a labelled plot. The plot should be scaled so that the number of data units/inch are the same for both dimensions. Otherwise, the distances in this plot would not be represented accurately. In `PROC PLOT`, this is done with the `vtoh` option, which specifies the aspect ratio (vertical to horizontal) of your printer.

```
proc plot vtoh=2;
```

```
   plot dim2 * dim1 = '*'$ eye / box haxis=by .1 vaxis=by .1;
run;
```



A labeled high-resolution display of the correspondence analysis solution (Figure 22) is constructed with the CORRESP macro. To equate the axes, you can use the defaults, specify specific axis control parameters, or use AXIS statements together with the HAXIS= and VAXIS= parameters to the CORRESP macro.

```
axis1 length=3.2 in  order=(-.5 to .5 by .5) label=(a=90 r=0);
axis2 length=6.4 in  order=(-1. to 1. by .5);

%corresp(data=haireye,
   id=eye,
   var=BLACK BROWN RED BLOND,
   vaxis=axis1, haxis=axis2, dimlab=Dim);
```

7.2 Multi-way tables

A three- or higher-way table can be analyzed by correspondence analysis in several ways. One approach is called "stacking". A three-way table, of size $I \times J \times K$ can be sliced into I two-way tables, each $J \times K$. If the slices are concatenated vertically, the result is one two-way table, of size $(I \times J) \times K$. In effect, the first two variables are treated as a single composite variable, which represents the main effects and interaction between the original variables that were combined. van der Heijden and de Leeuw (1985) discuss this use of correspondence analysis for multi-way tables and show how each way of slicing and stacking a contingency table corresponds to the analysis of a specified log-linear model.

In particular, for the three-way table that is reshaped as a table of size $(I \times J) \times K$, the correspondence analysis solution analyzes residuals from the log-linear model [AB] [C]. That is, for such a table, the $I \times J$ rows represent the joint combinations of variables A and B. The expected frequencies under independence for this table are

$$e_{[ij]k} = \frac{f_{[ij]+}\, f_{[+]k}}{n} = \frac{f_{ij+}\, f_{++k}}{n}$$

which are the ML estimates of expected frequencies for the log-linear model [AB] [C]. The χ^2 that is decomposed is the Pearson χ^2 for this log-linear model. When the table is stacked as $I \times (J \times K)$ or $J \times (I \times K)$, correspondence analysis decomposes the residuals from the log-linear models [A] [BC] and [B] [AC], respectively. van der Heijden and de Leeuw (1985) show how a generalized form of correspondence analysis can be interpreted as decomposing the difference between two specific log-linear models.

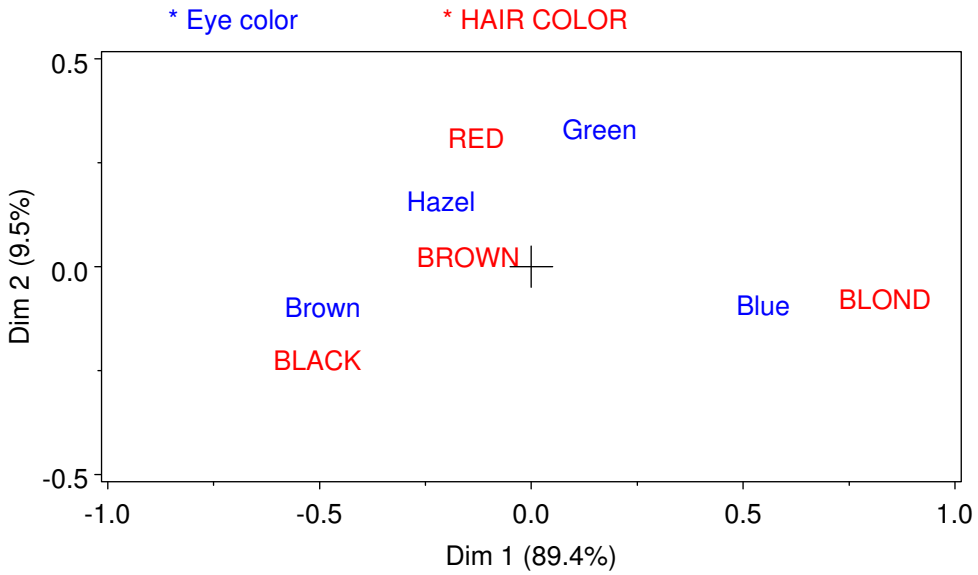


Figure 22: Correspondence analysis solution for Hair color, Eye color data

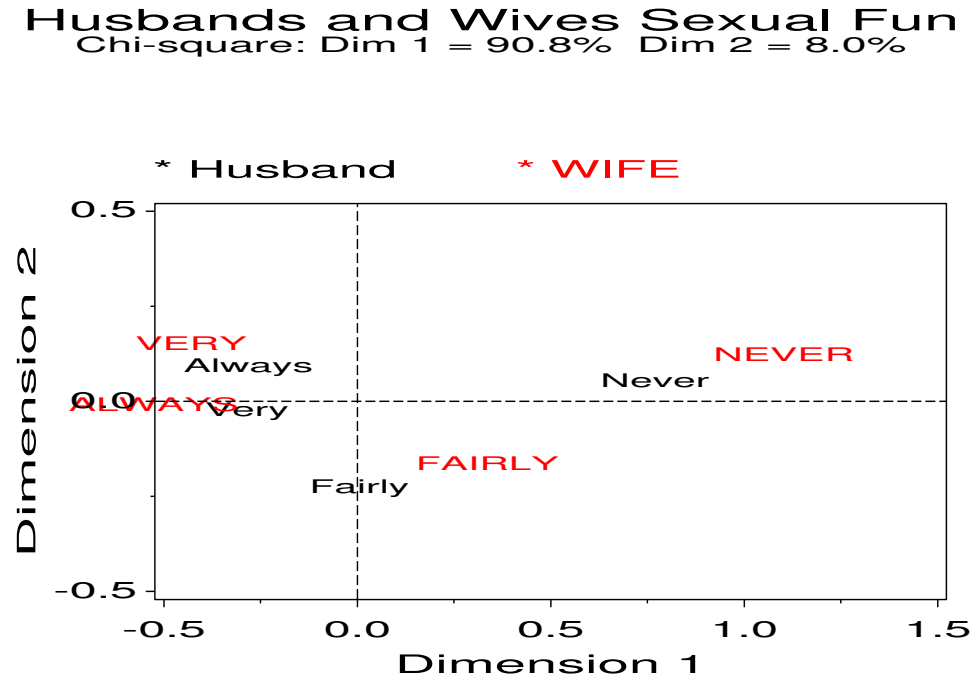


Figure 23: Correspondence analysis solution for Sexual fun data

7.2.1 Example: Suicide Rates

To illustrate the use of correspondence analysis for the analysis for three-way tables, we use data on suicide rates in West Germany, classified by age, sex, and method of suicide used. The data, from Heuer (1979, Table 1), have been discussed by van der Heijden and de Leeuw (1985) and others.

Sex	Age	POISON	GAS	HANG	DROWN	GUN	JUMP
M	10-20	1160	335	1524	67	512	189
M	25-35	2823	883	2751	213	852	366
M	40-50	2465	625	3936	247	875	244
M	55-65	1531	201	3581	207	477	273
M	70-90	938	45	2948	212	229	268
F	10-20	921	40	212	30	25	131
F	25-35	1672	113	575	139	64	276
F	40-50	2224	91	1481	354	52	327
F	55-65	2283	45	2014	679	29	388
F	70-90	1548	29	1355	501	3	383

The table below shows the results of all possible hierarchical log-linear models for the suicide data. It is apparent that none of these models has an acceptable fit to the data. Given the enormous sample size ($n = 48,177$), even relatively small departures from expected frequencies under any model would appear significant, however.

Model	df	L.R. G^2	G.F. χ^2
[M] [A] [S]	49	10119.60	9908.24
[M] [AS]	45	8632.0	8371.3
[A] [MS]	44	4719.0	4387.7
[S] [MA]	29	7029.2	6485.5
[MS] [AS]	40	3231.5	3030.5
[MA] [AS]	25	5541.6	5135.0
[MA] [MS]	24	1628.6	1592.4
[MA] [MS] [AS]	20	242.0	237.0

Correspondence analysis applied to the [AS] by [M] table helps to show the nature of the association between method of suicide and the joint age-sex combinations and decomposes the $\chi^2 = 8371$ for the log-linear model [AS] [M]. To carry out this analysis, the variables `age` and `sex` are listed together in the `TABLES` parameter; the option `CROSS=ROW` generates all combinations of Age and Sex.

```
%include catdata(suicide);

axis1 order=(-.7 to .7 by .7) length=6.5 in label=(a=90 r=0);
axis2 order=(-.7 to .7 by .7) length=6.5 in;
%corresp(data=suicide,
    tables=%str(age sex, method), weight=count,
options=cross=row short, vaxis=axis1, haxis=axis2);
```

The results show that over 93% of the association can be represented well in two dimensions.

```
                Inertia and Chi-Square Decomposition

Singular  Principal Chi-
Values    Inertias  Squares Percents   12   24   36   48   60
                                      ----+----+----+----+----+---
0.32138   0.10328   5056.91  60.41% ************************
0.23736   0.05634   2758.41  32.95% **************
0.09378   0.00879    430.55   5.14% **
0.04171   0.00174     85.17   1.02%
0.02867   0.00082     40.24   0.48%
          -------   -------
          0.17098   8371.28 (Degrees of Freedom = 45)
```

The plot of the scores for the rows (sex-age combinations) and columns (methods) in Figure 24 shows residuals from the log-linear model [AS] [M]. Thus, it shows the two-way associations of sex × method, age × method, and the three-way association, sex × age × method which are set to zero in the model [AS] [M]. The possible association between sex and age is not shown in this plot.

Dimension 1 in the plot separates males and females. This dimension indicates a strong difference between suicide profiles of males and females. The second dimension is mostly ordered by age with younger groups at the top and older groups at the bottom. Note also that the positions of the age groups are approximately parallel for the two sexes. Such a pattern indicates that sex and age do not interact in this analysis. The relation between the age - sex groups and methods of suicide can be interpreted in terms of similar distance and direction from the origin, which represents the marginal row and column profiles. Young males are more likely to commit suicide by gas or a gun, older males by hanging, while young females are more likely to ingest some toxic agent and older females by jumping or drowning.

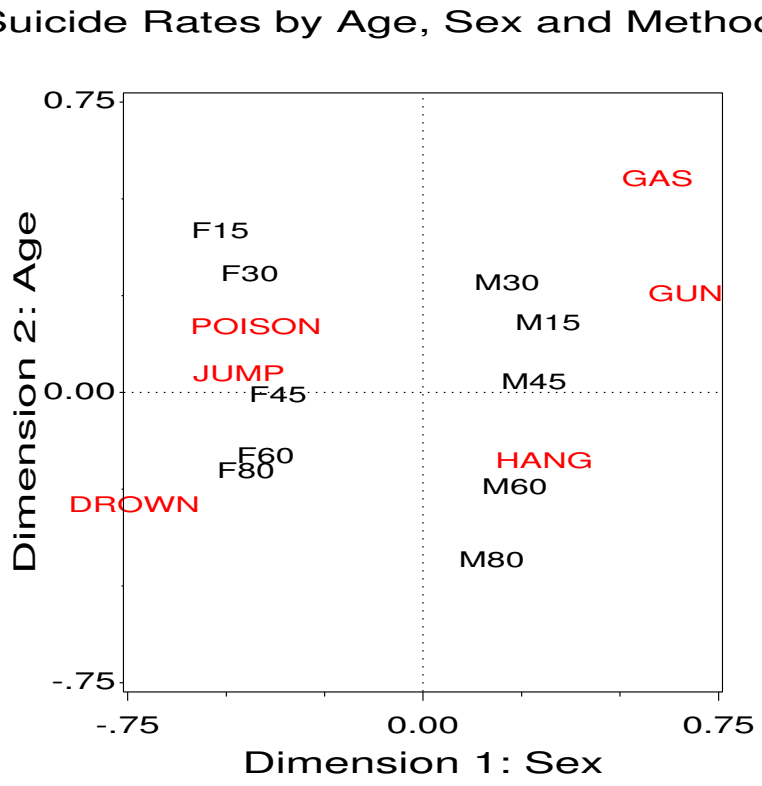


Figure 24: Two-dimensional correspondence analysis solution for the [AS] [M] multiple table.

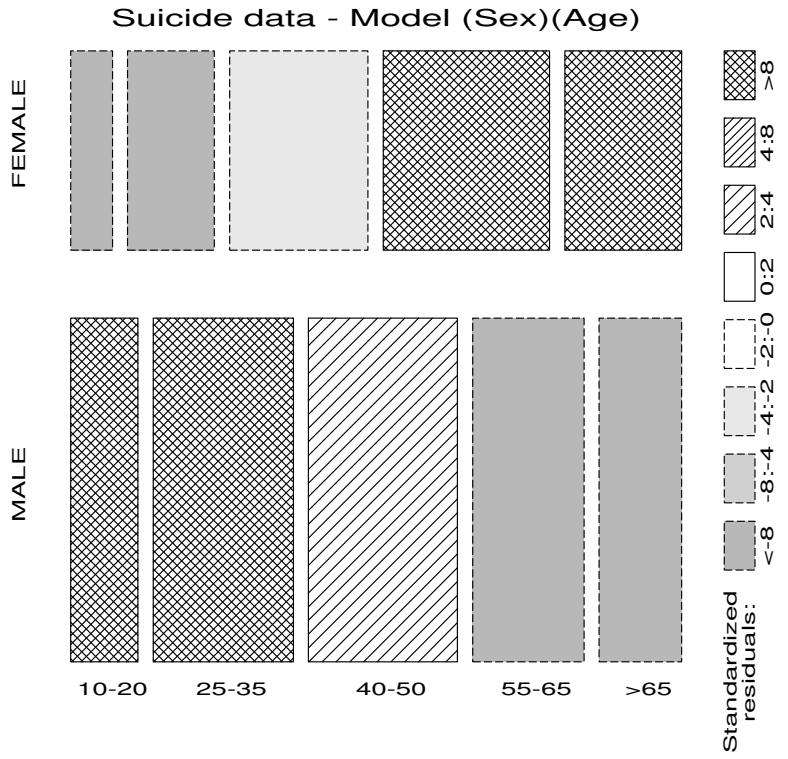


Figure 25: Mosaic display for sex and age. The frequency of suicide shows opposite trends with age for males and females.

8 Logistic Regression

8.1 Logistic Regression Model

Logistic regression[1] describes the relationship between a ***dichotomous*** response variable and a set of explanatory variables. The explanatory variables may be continuous or (with dummy variables) discrete.

Consider the data below on the effects of a particular treatment for arthritis. The response is improvement; sex and treatment are the explanatory variables.

		Improvement		
Sex	Treatment	None	Some/Marked	Total
F	Active	6	21	27
F	Placebo	19	13	32
M	Active	7	7	14
M	Placebo	10	1	11

Let π_{ij} be the probability that a patient of sex i who receives treatment j will show some or marked improvement. It is convenient to model this probability in terms of the ***log odds*** of improvement, called the ***logit***,

$$\text{logit}(\pi_{ij}) \equiv \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) \tag{11}$$

Number Observed Odds Observed

[1] Some material in this section borrows from Koch and Stokes (1991).

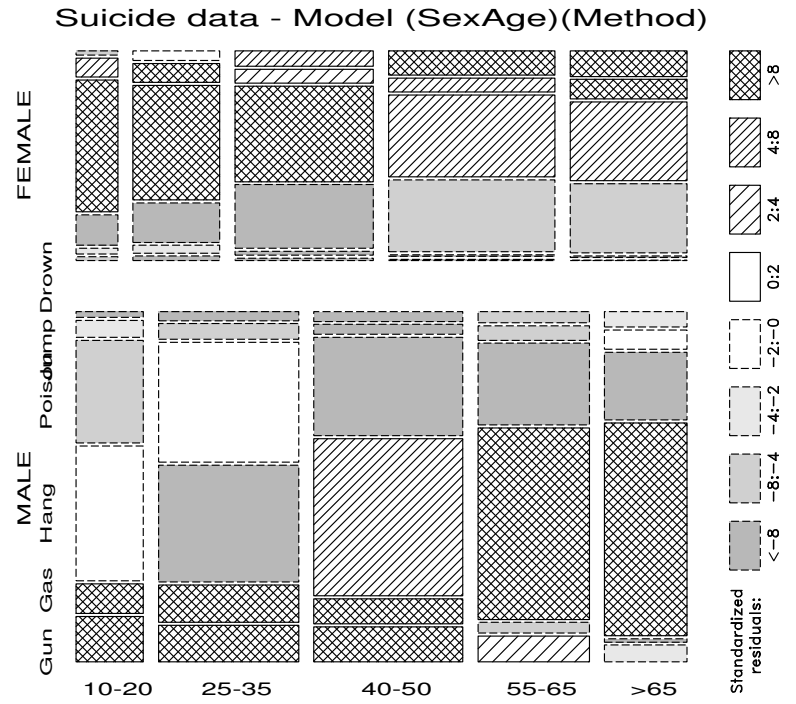


Figure 26: Mosaic display showing deviations from model [AS] [M]. The methods have been reordered according to their positions on Dimension 1 of the correspondence analysis solution for the [AS] [M] table.

SEX	Treatmt	Better	Total	P{better}	Better	Logit
Female	Active	21	27	0.7778	3.50	1.2529
Female	Placebo	13	32	0.4062	0.68	-0.3797
Male	Active	7	14	0.5000	1.00	0.0000
Male	Placebo	1	11	0.0909	0.10	-2.3026

The ***logistic regression model*** fits the log odds by a linear function of the explanatory variables (as in multiple regression).

$$\text{logit}(\pi_{ij}) = \alpha + \mathbf{x}'_{ij}\,\beta \tag{12}$$

For example, a simple model might assume additive ("main") effects for sex and treatment on the log odds of improvement.

$$\text{logit}\,(\pi_{ij}) = \alpha + \beta_1\,x_1 + \beta_2\,x_2 \tag{13}$$

where

- x_1 and x_2 are dummy variables representing sex and treatment, respectively:

$$x_1 = \begin{cases} 0 & \text{if male} \\ 1 & \text{if female} \end{cases} \qquad x_2 = \begin{cases} 0 & \text{if placebo} \\ 1 & \text{if active} \end{cases}$$

- α is the log odds of improvement for the baseline group—males receiving placebo.

- β_1 is the increment in log odds for being female. Therefore, e^{β_1} gives the odds of improvement for females relative to males.

- β_2 is the increment in log odds for being in the active treatment group. e^{β_2} gives the odds of improvement for the active treatment group relative to placebo.

Thus, the parameters defined here are *incremental effects*. The intercept corresponds to a baseline group (males given the placebo); the other parameters are incremental effects for the other groups compared to the baseline group.

8.1.1 Predicted probabilities

For plotting and interpreting results from logistic regression, it is usually more convenient to express fitted values on the scale of probabilities. The inverse transformation of (11) and (12) is the logistic function,

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \mathbf{x}'_{ij}\,\beta)}{1 + \exp(\alpha + \mathbf{x}'_{ij}\,\beta)} \tag{14}$$

For the example, when α, β_1, and β_2 have been estimated, the predicted odds and probabilities are:

Sex	Treatment	Odds Improved	Pr{Improved}
Female	Active	$e^{\alpha+\beta_1+\beta_2}$	$e^{\alpha+\beta_1+\beta_2}/(1+e^{\alpha+\beta_1+\beta_2})$
Female	Placebo	$e^{\alpha+\beta_1}$	$e^{\alpha+\beta_1}/(1+e^{\alpha+\beta_1})$
Male	Active	$e^{\alpha+\beta_2}$	$e^{\alpha+\beta_2}/(1+e^{\alpha+\beta_2})$
Male	Placebo	e^{α}	$e^{\alpha}/(1+e^{\alpha})$

The formulas for predicted probabilities may look formidable, but the numerical results are easy to interpret.

8.2 Fitting Logistic Regression Models

Logistic regression models can be fit using `PROC LOGISTIC`, `PROC CATMOD`, `PROC GENMOD` and SAS/INSIGHT. The examples below illustrate the use of `PROC LOGISTIC`. The input data set for `PROC LOGISTIC` can be in one of two forms: *frequency form* – one observation per group, with a variable containing the frequency for that group. A `FREQ` statement is used to provide the frequency variable. The other form is *case form* – one observation per case.

The following DATA step creates a data set in frequency form named `arthrit`. The dummy variables `_SEX_` and `_TREAT_` corresponding to x_1 and x_2 are created, as is the dichotomous response variable, `better`.

The first logistic regression model includes effects for sex and treatment, specified by the dummy variables on the `MODEL` statement Note that by default, `PROC LOGISTIC` orders the response values in *increasing* order, and sets up the model so that it is predicting the probability of the *smallest* ordered value, Pr{better=0}, which means it would be modelling the probability of No improvement. The `descending` option (available with Version 6.08) reverses this order, so that predicted results will be for Pr{better=1}.[2]

```
data arthrits;
   input sex$ trtment$ improve$ count;
   _treat_ = (trtment='Active');
   _sex_   = (sex='F');
   better  = (improve='some');
datalines;
F Active  none   6
M Active  none   7
F Active  some  21
M Active  some   7
F Placebo none  19
M Placebo none  10
F Placebo some  13
M Placebo some   1
;
proc logistic data=arthrits descending;
```

[2] In earlier releases, a format is used to order the values of `better` so that the first value corresponds to Improved.

```
freq count;
model better = _sex_ _treat_ / scale=none aggregate;
```

The options `scale=none aggregate` provide goodness of fit tests for the model. The goodness of fit tests are based on the difference between the actual model fitted and the saturated model (containing an interaction of sex and treatment in this example), which would fit perfectly. The following results are produced:

```
            Deviance and Pearson Goodness-of-Fit Statistics
                                                        Pr >
       Criterion        DF       Value    Value/DF    Chi-Square

       Deviance          1      0.2776      0.2776        0.5983
       Pearson           1      0.2637      0.2637        0.6076

                    Number of unique profiles: 4

                Testing Global Null Hypothesis: BETA=0

                           Intercept
              Intercept       and
  Criterion     Only      Covariates   Chi-Square for Covariates

  AIC           118.449     104.222      .
  SC            118.528     104.460      .
  -2 LOG L      116.449      98.222    18.227 with 2 DF (p=0.0001)
  Score            .           .      16.797 with 2 DF (p=0.0002)
```

The Chi-square tests for BETA=0 above test the joint effect of sex and treatment. Individual effects in the model are tested by Wald χ^2s, the squared ratio of each parameter divided by its standard error. These tests, shown below, indicate that both sex and treatment effects are highly significant.

```
                  Analysis of Maximum Likelihood Estimates

              Parameter Standard    Wald       Pr >    Standardized    Odds
  Variable DF  Estimate   Error  Chi-Square Chi-Square   Estimate     Ratio

  INTERCPT 1    -1.9037   0.5982    10.1286     0.0015           .     .
  _SEX_    1     1.4687   0.5756     6.5092     0.0107     0.372433   4.343
  _TREAT_  1     1.7817   0.5188    11.7961     0.0006     0.493956   5.940
```

8.2.1 Model interpretation

The fitted model,

$$\text{logit}(\pi_{ij}) = -1.90 + 1.47\,\text{sex} + 1.78\,\text{treat} \tag{15}$$

is most easily interpreted by considering the odds ratios corresponding to the parameters:

- 1.47 is the increment to log odds of a better outcome for females; the odds ratio $e^{1.47} = 4.34$ indicates that females are 4.3 times as likely to achieve a better outcome than males.
- 1.78 is the increment to log odds for the treatment group; the odds ratio $e^{1.78} = 5.94$ indicates that the treated group is nearly 6 times as likely to achieve a better outcome than the placebo group.

8.2.2 Plotting results from PROC LOGISTIC

`PROC LOGISTIC` also calculates predicted probabilities and logits, and these results may be obtained in an output data set, from which plots can be made.

```
proc logistic data=arthrit;
   freq count;
   model  better = _sex_  _treat_ ;
    output out=results p=predict l=lower u=upper xbeta=logit;
```

The OUTPUT statement produces a data set containing estimated logit values for each group, and corresponding predicted probabilities of improvement and confidence limits (UPPER, LOWER) for these probabilities.

The output data set RESULTS contains the following variables. There are two observations for each group (for none and some improvement). The PREDICT variable gives the predicted probability of an improved outcome according to model (15), using the inverse transformation (14) of logit to probability.

SEX	TRTMENT	IMPROVE	COUNT	PREDICT	LOWER	UPPER	LOGIT
F	Active	none	6	0.794	0.620	0.900	1.347
M	Active	none	7	0.470	0.257	0.694	-0.122
F	Active	some	21	0.794	0.620	0.900	1.347
M	Active	some	7	0.470	0.257	0.694	-0.122
F	Placebo	none	19	0.393	0.248	0.560	-0.435
M	Placebo	none	10	0.130	0.044	0.325	-1.904
F	Placebo	some	13	0.393	0.248	0.560	-0.435
M	Placebo	some	1	0.130	0.044	0.325	-1.904

To plot the predicted probabilities of improvement and confidence limits from the RESULTS data set, we select the observations for improve='some'. A plot can be done as a bar chart with PROC GCHART or as a line graph with PROC GPLOT. Confidence limits can be added to either with the SAS/GRAPH Annotate facility. The statements below show how a grouped horizontal bar chart (Figure 27) is constructed.

```
data results;
   set results;
   if improve='some';
   label predict='Prob. Improved';
data limits;
   set results;
   xsys='2'; ysys='2';
   midpoint=trtment;
   group=sex;
   x = lower;  function='MOVE   '; output;
   text='|';   function='LABEL  '; output;
   x = upper;  function='DRAW   '; output;
   text='|';   function='LABEL  '; output;
proc gchart data=results;
   hbar trtment / sumvar=predict group=sex gspace=3
                  anno=limits
                  raxis=axis1
                  maxis=axis2
                  gaxis=axis3;
   axis1 order=(0 to 1 by .2) minor=none
         label=(h=1.5) value=(h=1.3);
   axis2 label=(h=1.3 'Treat') value=(h=1.1);
   axis3 label=(h=1.3) value=(h=1.2);
   pattern1 v=solid c=cyan;
   pattern2 v=solid c=rose;
```

8.3 Quantitative predictors

Logistic regression can be generalized to include continuous explanatory variables. The main differences are:

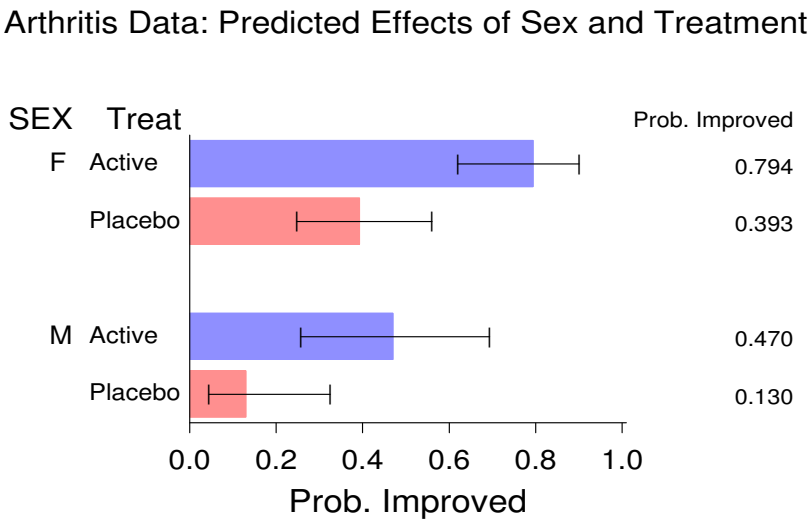


Figure 27: Predicted probabilities of improvement

- the data are usually entered in case form, with one observation per person.
- the statistics for goodness-of-fit are computed differently.
- the output data set for plotting now contains one observation per person.

The DATA step below creates a SAS data set named `arthrit`, in case form. It is equivalent to the earlier example, but contains the age of each person in the sample. Programming statements are used to create dummy variables `_sex_` and `_treat_` as before. Variables for testing for interactions among sex, treatment and age are also created. A preliminary analysis (described later) is used to test whether any of these variables interact. That test shows that all interactions can safely be ignored in this example. That test is one suitable test for goodness of fit of the main effects model.

```
data arthrit;
   length treat$7. sex$6. ;
   input id treat$ sex$ age improve @@ ;
   better  = (improve > 0);            /* Dichotomous response    */
   _treat_ = (treat ='Treated') ;      /* Dummy var for treatment */
   _sex_   = (sex = 'Female');         /*           and sex       */
   agesex  = age*_sex_ ;               /* Dummy var for testing   */
   agetrt  = age*_treat_;              /*   interactions          */
   sextrt  = _sex_*_treat_;
   age2    = age*age ;
  cards ;
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
```

The next logistic model includes (main) effects for age, sex, and treatment. The `lackfit` option requests a lack-of-fit test proposed by Hosmer and Lemeshow (1989). This test divides subjects into deciles based on predicted probabilities, then computes a χ^2 from observed and expected frequencies. In this example, we plot both a confidence interval for Pr{Improved} and logit $\pm$ s.e.. To make these intervals roughly comparable, we choose $\alpha = .33$ to give a 67% confidence interval.

```
title2 h=1.3 f=duplex 'Estimated Effects of Age, Treatment and Sex';
proc logistic data=arthrit;
   format better outcome.;
   model  better = _sex_  _treat_  age / lackfit;
   output out=results p=predict l=lower u=upper
                      xbeta=logit stdxbeta=selogit / alpha=.33;
```

The results include:

```
               Testing Global Null Hypothesis: BETA=0
                           Intercept
                Intercept      and
 Criterion        Only     Covariates   Chi-Square for Covariates

 AIC             118.449     100.063       .
 SC              120.880     109.786       .
 -2 LOG L        116.449      92.063     24.386 with 3 DF (p=0.0001)
 Score              .           .        22.005 with 3 DF (p=0.0001)
```

```
                    Analysis of Maximum Likelihood Estimates
                Parameter Standard    Wald       Pr >    Standardized    Odds
 Variable DF    Estimate   Error  Chi-Square Chi-Square   Estimate     Ratio
 INTERCPT 1     -4.5033   1.3074    11.8649     0.0006           .       .
 _SEX_    1      1.4878   0.5948     6.2576     0.0124     0.377296   4.427
 _TREAT_  1      1.7598   0.5365    10.7596     0.0010     0.487891   5.811
 AGE      1      0.0487   0.0207     5.5655     0.0183     0.343176   1.050
```

To interpret the parameter estimate for age, note that the odds ratio of 1.050 means that the odds of improvement increases 5% per year. This means that over 10 years, the odds of improvement would be multiplied by $e^{.487} = 1.63$, a 63% increase.

```
               Hosmer and Lemeshow Goodness-of-Fit Test
                                                 BETTER = not
                         BETTER = improved          improved
                       --------------------    --------------------
 Group       Total     Observed    Expected    Observed    Expected
     1           9            0        1.09           9        7.91
     2           8            3        1.48           5        6.52
     3           8            2        2.00           6        6.00
     4           8            2        3.04           6        4.96
     5           8            5        3.64           3        4.36
     6           9            5        4.82           4        4.18
     7           8            4        4.78           4        3.22
     8           8            5        5.77           3        2.23
     9           8            7        6.62           1        1.38
    10          10            9        8.77           1        1.23
      Goodness-of-fit Statistic = 5.5549 with 8 DF (p=0.6970)
```

The Hosmer and Lemeshow goodness-of-fit χ^2 indicates that the model has an acceptible fit to the data.

Plots are constructed from the data set RESULTS. The first few observations are shown below.

ID	TREAT	AGE	SEX	IMPROVE	PREDICT	LOWER	UPPER	LOGIT	SELOGIT
57	Treated	27	Male	1	0.194	0.103	0.334	-1.427	0.758
9	Placebo	37	Male	0	0.063	0.032	0.120	-2.700	0.725
46	Treated	29	Male	0	0.209	0.115	0.350	-1.330	0.728
14	Placebo	44	Male	0	0.086	0.047	0.152	-2.358	0.658
...									

Predicted probabilities and confidence limits are contained in the variables PREDICT, UPPER, and LOWER. To show the effects of sex, treatment, and age on Pr{Improvement}, separate plots are drawn for each sex, using the statement BY SEX; in the PROC GPLOT step. Most of the work consists of drawing the confidence limits with the Annotate facility. Plots of the predicted logits can be made in a similar way using the variables LOGIT and SELOGIT.

```
proc sort data=results;
   by sex treat age;
data bars;
   set results;
   by sex treat;
   length text$8;
   xsys='2'; ysys='2';
   if treat='Placebo' then color='BLACK';
                      else color='RED';
   x  = age; line=33;
   y  = upper;  function='MOVE   '; output;
   text='-';    function='LABEL  '; output;
   y  = lower;  function='DRAW   '; output;
   text='-';    function='LABEL  '; output;
   if last.treat then do;
      y = predict;
      x = age+1; position='6'; size=1.4;
      text = treat; function='LABEL'; output;
      end;
   if first.sex then do;
      ysys ='1'; y=90;
      xsys ='1'; x=10; size=1.4;
      text = sex; function='LABEL'; output;
      end;

goptions hby=0;
proc gplot;
   plot predict * age = treat / vaxis=axis1 haxis=axis2
                                nolegend anno=bars;
   by sex;
   axis1 label=(h=1.3 a=90 f=duplex 'Prob. Improvement (67% CI)')
         value=(h=1.2) order=(0 to 1 by .2);
   axis2 label=(h=1.3 f=duplex)
         value=(h=1.2) order=(20 to 80 by 10)
         offset=(2,5);
   symbol1 v=+ h=1.4 i=join l=3 c=black;
   symbol2 v=$ h=1.4 i=join l=1 c=red;
```

8.3.1 Scales

Note that the response variable is easy to understand on the probability scale, but the curvilinear relationship of Pr{Improve} to age, sex, and treatment is more complex. On the other hand, the response is somewhat more difficult to understand on the logit scale (log odds), but the linear relationship of the logit to age, sex, and treatment is simpler.

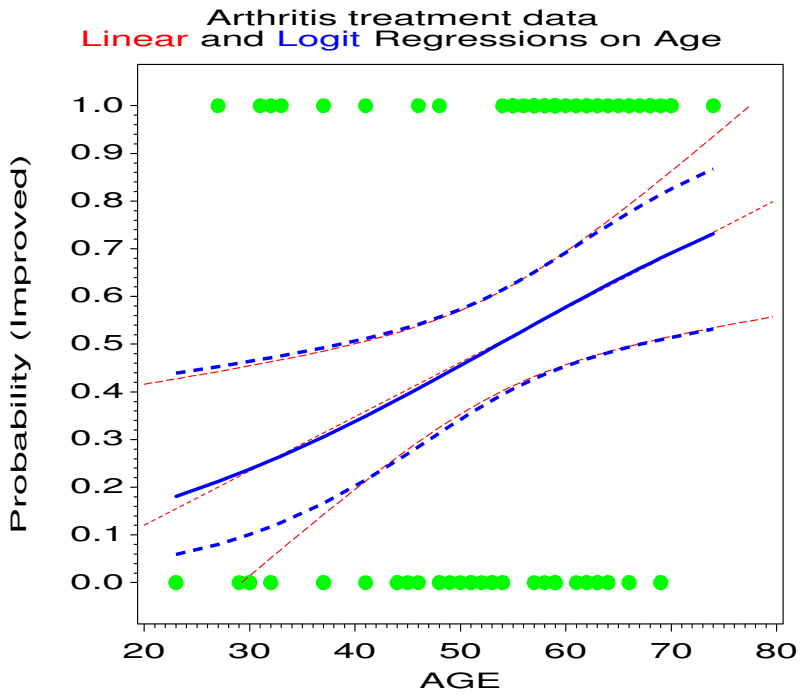


Figure 28: Quantitative predictor: Linear and Logit regression on age. The curves show predicted probabilities of improvement and 95% confidence bands (ignoring sex and treatment). The points show the observations. Except in the extremes, the linear and logistic models give similar predicted values.

One compromise is to plot the results on the logit scale, and add a second scale showing corresponding probability values. In SAS, this can be done with an Annotate data set. The example below is designed for a plot with logit values ranging from -5 to +2. For each of a set of simple probability values (`prob`), the probability is labeled on an axis at the right of the figure corresponding to the logit value at the left.

```
data pscale;
   xsys = '1';                              * percent values for x;
   ysys = '2';                              * data values for y;
   do prob = .01,.05,.1,.25,.5,.75,.9;
      logit = log( prob / (1-prob) );       * convert to logit;
      y = logit;
      x = 99 ;  function='MOVE   '; output;      * tick marks;
      x = 100;  function='DRAW   '; output;
      text = put(prob,4.3); position='6';        * values;
      style='DUPLEX';
      function='LABEL  '; output;
      end;
   ysys='1'; x=100.1; y=99.5;                    * axis label;
   text='Prob';           output;
```

In a `PROC GPLOT` step, the Annotate data set `pscale` is used as shown below:

```
proc gplot data=results;
   plot logit * x / annotate=pscale ... ;
```

8.3.2 The *Challenger* disaster

The Space Shuttle *Challenger* exploded shortly after take-off in January 1986. Subsequent investigation determined that the cause was failure of the O-ring seals used to isolate the fuel supply from burning gases.

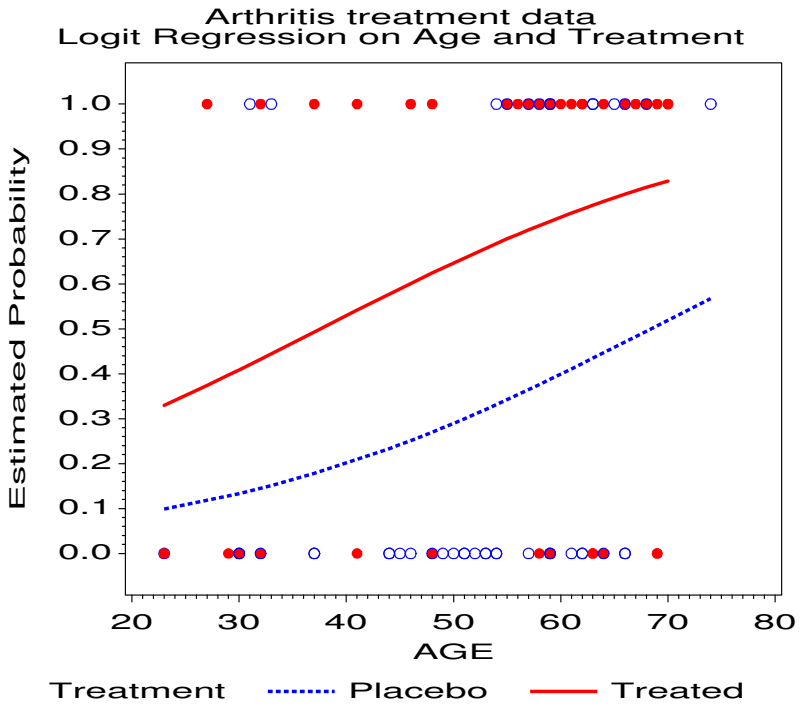


Figure 29: Logit regression on age and treatment.

Engineers from Morton Thiokol, manufacturers of the rocket motors, had been worried about the effects of unseasonably cold weather on the the O-ring seals and recommended aborting the flight. NASA staff analysed the data on the relation between ambient temperature and the number of O-ring failures (out of 6), but they had excluded observations where no O-rings failed, believing that they were uninformative. Unfortunately, those observations had occurred when the launch temperature was relatively warm ($65 - 80^\circ$F.); the coldest temperature at any previous launch was 53°. When *Challenger* was launched, the temperature was 31°.

The data relating O-ring failures to temperature were depicted as in Figure 31, perhaps one of the most misleading graphs in history. Examination of this graph seemed to indicate that there was no relation between ambient temperature and failure. Thus, the decision to launch the *Challenger* was made, in spite of the initial concerns of the Morton Thiokol engineers.

The DATA step below reads the data on the number of O-ring failures and temperature for the 23 flights for which information was available before the *Challenger* launch. Our interest here is in predicting the likelihood of failures at low temperatures.

```
title 'NASA Space Shuttle O-Ring Failures';
data nasa;
   input failures temp @@;
   orings = 6;
   label failures = 'Number of O-ring failures'
      temp = 'Temperature (deg F)';
   cards;
   2  53    1  57    1  58    1  63
   0  66    0  67    0  67    0  67
   0  68    0  69    0  70    0  70
   1  70    1  70    0  72    0  73
   0  75    2  75    0  76    0  76
   0  78    0  79    0  80
;
```

To obtain predicted probabilities for observations not in the original sample, create an additional data set which contains values for the independent variables in the extrapolation sample, and join these observations to the actual data set. The response variable (`failures`) will be missing for the extrapolation sample.

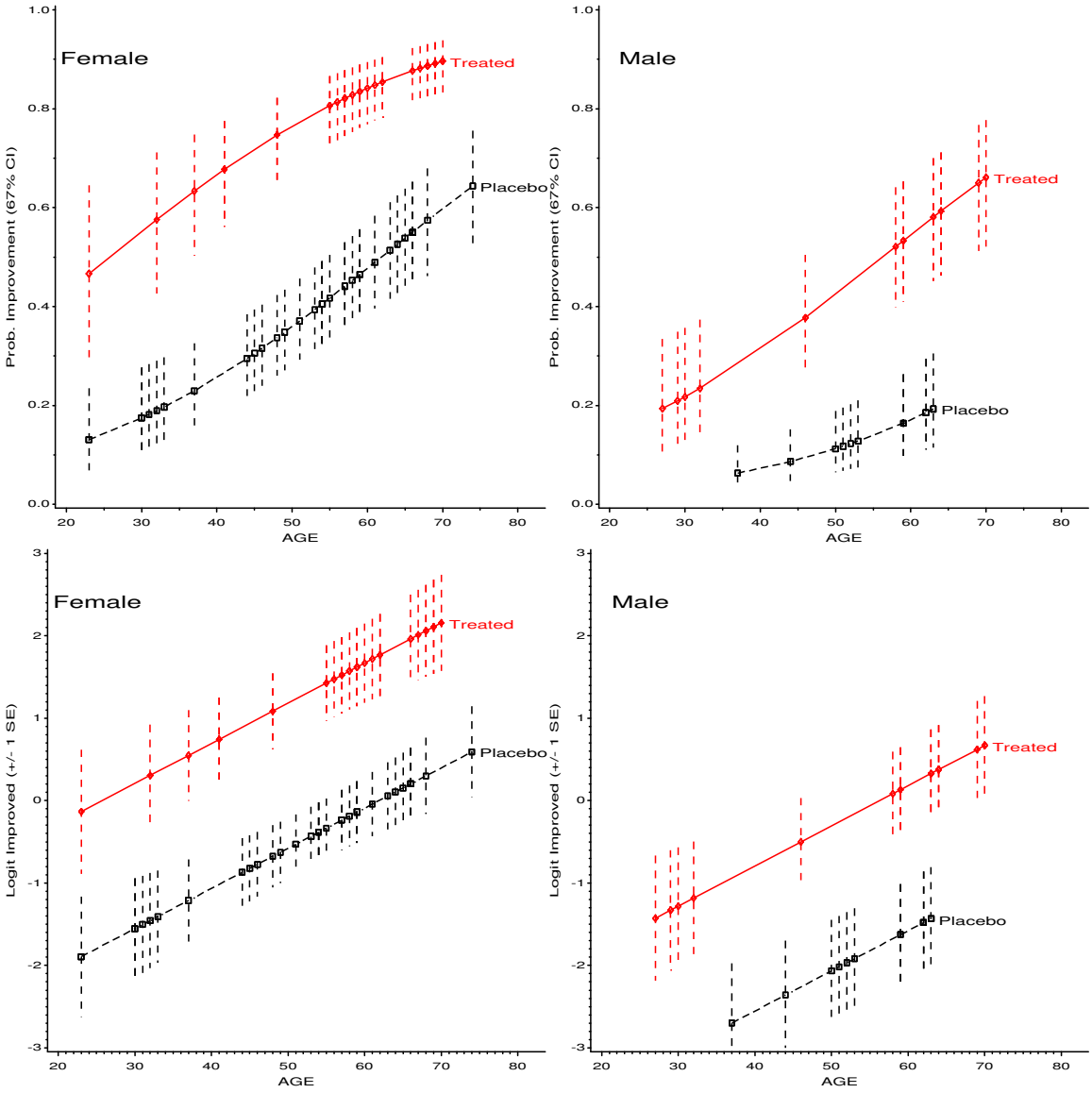


Figure 30: Estimated effects of age, treatment and sex. The effects are simpler (linear and additive) on the logit scale, but more easily interpreted in terms of probabilities. One solution is to plot on the logit scale, and provide a separate (nonlinear) scale of probabilities.

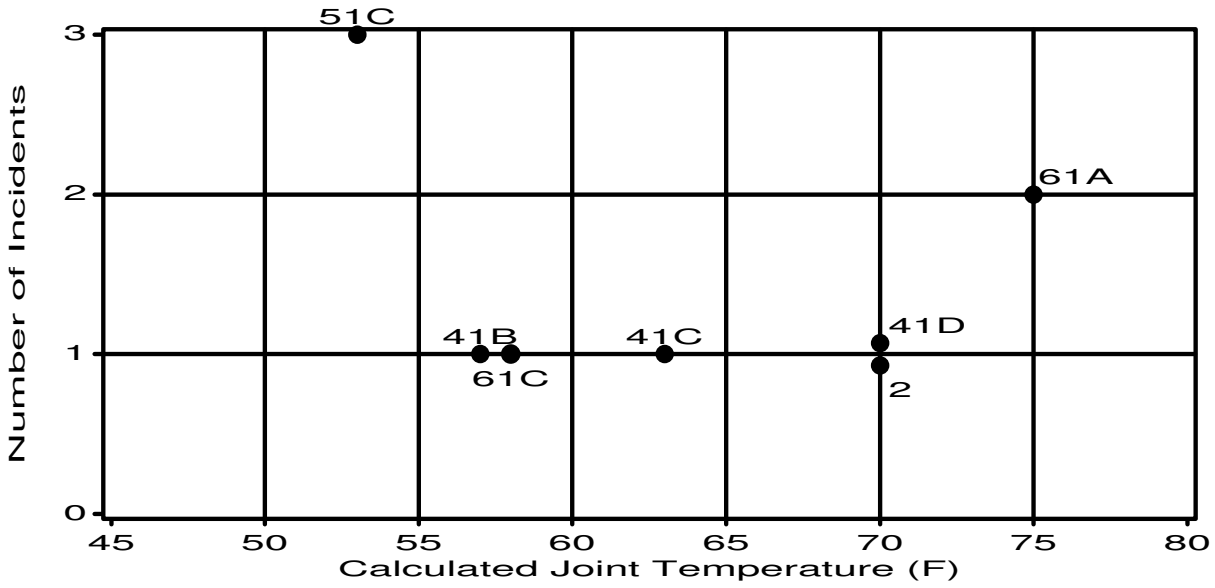


Figure 31: NASA Space Shuttle pre-launch graph. Observations with no failures omitted.

```
*-- Obtain predicted values for 30-80 degrees;
data temp;
   input temp @@;
cards;
31 30 35 40 45 50 55 60 65 70 75 80
;
data nasa2;
   set nasa temp;
```

In the PROC LOGISTIC step, we use the ***events/trials*** syntax to indicate the number of failures and number of trials. The observations in the extrapolation sample are not used in fitting the model, yet the procedure produces predicted probabilities and logits (as long as the independent variable(s) are non-missing).

```
proc logistic data=nasa2 nosimple;
   model failures/orings = temp ;
   output out=results p=predict l=lower u=upper;
proc print;
```

The printed output, shown in Output 8.1 indicates that the 12 new observations were not used in the analysis. The odds ratio, 0.891, is interpreted to mean that each increase of 1° in temperature decreases the odds of a failure by 11%!

The output data set `results` contains the predicted probability of a failure at each temerature and upper and lower confidence 95% limits for this probability. We can plot the predicted and observed values as shown below. A vertical reference line at 31° is used to highlight the conditions at the *Challenger* launch.

```
proc sort data=results;
   by predict;
data results;
   set results;
   obs = failures / orings;

proc gplot data=results;
   plot (obs predict lower upper) * temp /
      href=31 lhref=33
      overlay frame vaxis=axis1 vminor=1;
   symbol1 v=dot i=none c=blue h=2;
```

Output 8.1 Logistic regression for NASA O-ring data

```
                NASA Space Shuttle O-Ring Failures                1

                        The LOGISTIC Procedure

Data Set: WORK.NASA2
Response Variable (Events): FAILURES  Number of O-ring failures
Response Variable (Trials): ORINGS
Number of Observations: 23
Link Function: Logit


                           Response Profile

                     Ordered  Binary
                       Value  Outcome      Count

                           1  EVENT            9
                           2  NO EVENT       129

WARNING: 12 observation(s) were deleted due to missing values for
         the response or explanatory variables.


Model Fitting Information and Testing Global Null Hypothesis BETA=0

                          Intercept
             Intercept       and
Criterion      Only      Covariates  Chi-Square for Covariates

AIC             68.540      64.416       .
SC              71.468      70.271       .
-2 LOG L        66.540      60.416      6.124 with 1 DF (p=0.0133)
Score             .           .         6.804 with 1 DF (p=0.0091)


              Analysis of Maximum Likelihood Estimates

             Parameter Standard    Wald       Pr >    Standardized    Odds
Variable DF  Estimate   Error  Chi-Square Chi-Square   Estimate     Ratio

INTERCPT 1     5.0940   3.0552     2.7798     0.0955            .     .
TEMP     1    -0.1158   0.0471     6.0491     0.0139    -0.437656   0.891
```

```
symbol2 v=none i=spline c=black w=5;
symbol3 v=none i=spline c=red l=33 r=2 w=3;
axis1 label=(a=90 'Estimated Failure probability') offset=(3);
```

The graph is shown in Figure 32. There's not much data at low temperatures (the confidence band is quite wide), but the predicted probability of failure is uncomfortably high. Would you take a ride on *Challenger* when the weather is cold?

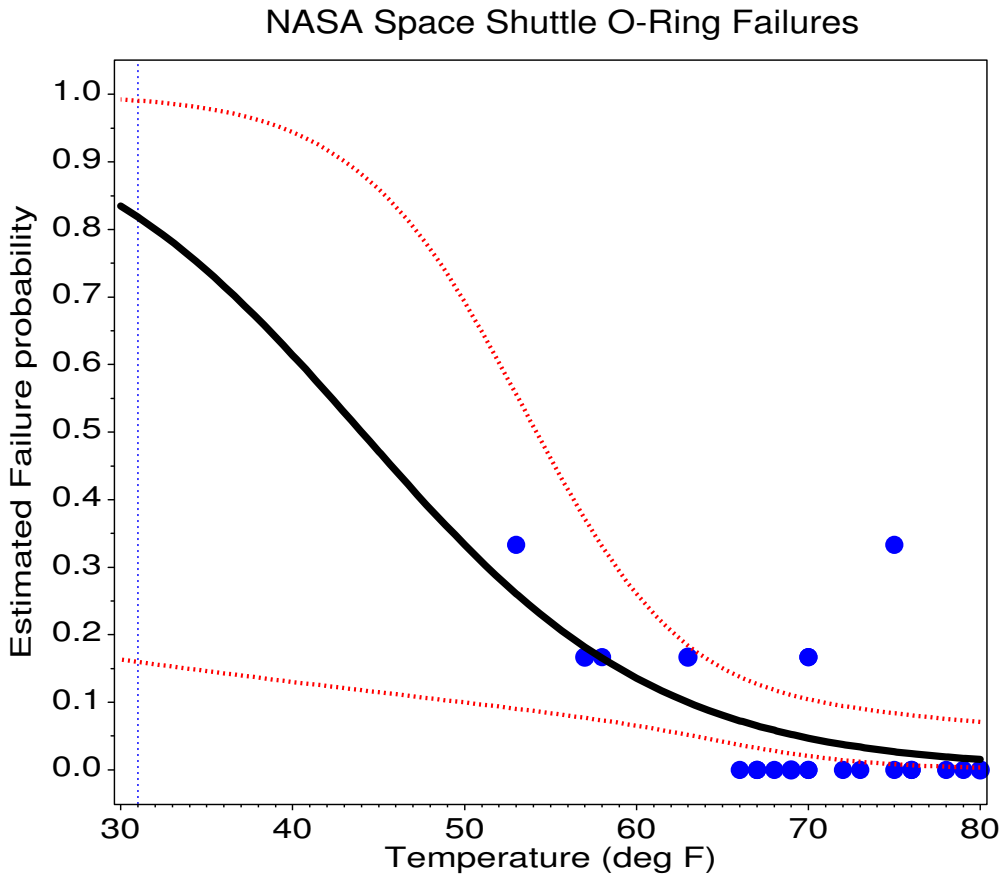


Figure 32: NASA Space Shuttle O-ring Failure, Observed and Predicted probabilities

8.4 Models with interaction

The examples for the arthritis data have involved only main effects of sex, age, and treatment. (Tests for interactions among these factors show them all to be insignificant.) However, the plotting of estimated logits or predicted probabilities from `PROC LOGISTIC` is no more complicated in models with interactions.

In fact, since the predicted probabilities and logits are calculated by the procedure and output to the data set `RESULTS`, the results plotted depend purely on the `MODEL` statement. The plotting steps remain the same, assuming you want to make separate plots for males and females of the age by treatment effects.

8.4.1 Coefficient method

A different way to construct these same plots is based on calculating the fitted logit values from the coefficients for the variables in the model. For the arthritis main effect model, the fitted relationship is

$$\text{logit}(\text{p}) = -4.5033 + 1.4878\,\text{sex} + 1.7598\,\text{treat} + 0.0487\,\text{age}$$

With this method, each independent variable is manipulated over its range. The response can be graphed on the probability scale by transforming the logit: $p = e^{\text{logit}}/(1+e^{\text{logit}})$. For example, the fitted logits and corresponding probabilities can be calculated in this data step:

```
data fitted;
  do _sex_ = 0 to 1;
     do _treat_ = 0 to 1;
        do age = 25 to 75 by 5;
           logit= -4.5033 + 1.4878*_sex_ + 1.7598*_treat_
                          + 0.0487*age;
           prob = exp(logit) / (1 + exp(logit));
           output;
           end;
        end;
     end;
```

Fox (1987) explains how this method may be used to construct adjusted effects plots for particular interactions, adjusting for other variables not represented in the plot.

8.4.2 Testing for interactions

One way to test for interactions with PROC LOGISTIC is to start with a main-effects model, and use a forward-selection method to find interaction terms which significantly reduce the residual χ^2.

The interaction effects were defined in the data step arthrit as the dummy variables, agesex, agetrt, and sextrt. The variable age2 = age**2 can be used to test whether the relationship between age and logit(better) is quadratic rather than linear.

The PROC LOGISTIC step below requests a forward-selection procedure. Setting START=3 requests that the model building begin with the first three variables (the main effects) listed in the model statement. SLENTRY=1 (significance level to enter) forces all variables to enter the model eventually.

```
proc logistic data=arthrit;
   format better outcome.;
   model  better = _sex_ _treat_ age        /* main effects  */
                   agesex agetrt sextrt     /* interactions  */
                   age2                     /* quadratic age */
          / selection=forward
            start=3
            slentry=1;
```

The variables included in each model for the selection procedure are listed in a note at the beginning of each set of results:

```
   Step  0. The following variables were entered:
            INTERCPT  _SEX_     _TREAT_   AGE
```

Results for this step are identical to those of the main effects model given earlier. Near the end of this step, the residual χ^2 is printed, which corresponds to a joint test for the other four variables. This test is an appropriate test of goodness of fit of the main effects model.

```
   Residual Chi-Square = 4.0268 with 4 DF (p=0.4024)
```

Other tests printed show none of the interaction terms is significant individually.

8.5 Ordinal Response: Proportional Odds Model

The proportional odds model extends logistic regression to handle an ordinal response variable. The response variable improve in the arthritis data actually has 3 levels, corresponding to None, Some, or Marked improvement.

		Improvement			
Sex	Treatment	None	Some	Marked	Total
F	Active	6	5	16	27
F	Placebo	19	7	6	32
M	Active	7	2	5	14
M	Placebo	10	0	1	11

One way to model these data is to consider two logits for the dichotomies between adjacent categories:

$$\text{logit}\,(\theta_{ij1}) = \log \frac{\pi_{ij1}}{\pi_{ij2} + \pi_{ij3}} = \text{logit (None vs. [Some or Marked])}$$

$$\text{logit}\,(\theta_{ij2}) = \log \frac{\pi_{ij1} + \pi_{ij2}}{\pi_{ij3}} = \text{logit ([None or Some] vs. Marked)}$$

Consider a logistic regression model for each logit:

$$\text{logit}(\theta_{ij1}) = \alpha_1 + \mathbf{x}'_{ij}\,\beta_1 \tag{16}$$

$$\text{logit}(\theta_{ij2}) = \alpha_2 + \mathbf{x}'_{ij}\,\beta_2 \tag{17}$$

The proportional odds assumption is that ***the regression functions are parallel*** on the logit scale, i.e., that $\beta_1 = \beta_2$.

For the arthritis example, with additive effects for sex and treatment on the both log odds,

$$\text{logit}\,(\theta_{ij1}) = \alpha_1 + \beta_1\,x_1 + \beta_2\,x_2 \tag{18}$$

$$\text{logit}\,(\theta_{ij2}) = \alpha_2 + \beta_1\,x_1 + \beta_2\,x_2 \tag{19}$$

where:

- x_1 and x_2 are dummy variables representing sex and treatment.
- α_1 is the log odds of no improvement (vs. some or marked) for males receiving the placebo.
- α_2 is the log odds of no improvement or some improvement (vs. marked) for males receiving the placebo.
- β_1 is the increment to *both* log odds for being female. Therefore, e^{β_1} gives the odds of improvement for females relative to males.
- β_2 is the increment to both log odds for being in the active treatment group. e^{β_2} gives the odds of improvement for the active treatment group relative to placebo.

The corresponding models including effects of age, as well as treatment and sex are similar to (18) and (19), with the addition of a term $\beta_3\,age$.

8.5.1 Plotting results from PROC LOGISTIC

Plotting results for the proportional odds model is similar to the earlier examples for a binary response variable. The main differences are:

- The validity of the analysis depends on the correctness of the proportional odds assumption. A test of this assumption appears in the output from `PROC LOGISTIC`.

- The results from `PROC LOGISTIC` are cast in terms of predicted probabilities and fitted logits for response *less than* each of the cutpoints. To plot Pr{Improve}, we must reverse the sense of the probabilities and logits.

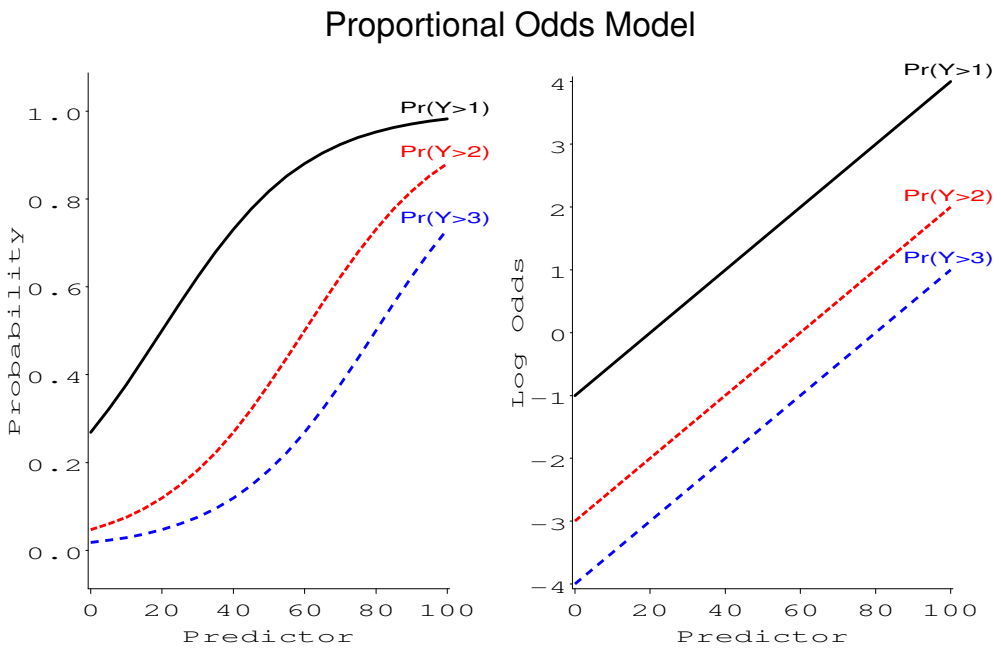


Figure 33: Proportional odds model. The model assumes that the regression functions for different response categories are parallel on the logit scale.

8.5.2 Example

This example fits the effects of treatment, sex and age for the proportional odds model with the arthritis data. Note that the dependent variable is `improve`, with values 0, 1, and 2.

```
   *-- Proportional Odds Model: Effects of treat, sex and age;
proc logistic data=arthrit nosimple;
   model  improve = _sex_  _treat_ age ;
   output out=results p=predict l=lower u=upper
                      xbeta=logit stdxbeta=selogit / alpha=.33;
```

The response profile displays the ordering of the outcome variable. Note that logits are formed from top to bottom, i.e., none vs. some or marked, none or some vs. marked. The output also shows the proportional odds assumption is reasonable.[3]

```
                          Response Profile
                    Ordered
                      Value  IMPROVE     Count

                          1        0        42
                          2        1        14
                          3        2        28
```

```
          Score Test for the Proportional Odds Assumption
              Chi-Square = 2.4917 with 3 DF (p=0.4768)
```

The parameter estimates relate to the odds of a poorer response (they are all negative):

[3]Peterson and Harrell (1990) showed that the score test for the proportional odds assumption is very liberal—rejecting the assumption far too often. One should adopt a stringent α-level for this test.

```
             Analysis of Maximum Likelihood Estimates

              Parameter  Standard     Wald        Pr >    Standardized
  Variable     Estimate    Error   Chi-Square  Chi-Square   Estimate

  INTERCP1       3.7837    1.1530     10.7683      0.0010           .
  INTERCP2       4.6827    1.1949     15.3569      0.0001           .
  _SEX_         -1.2517    0.5321      5.5343      0.0186    -0.317412
  _TREAT_       -1.7453    0.4772     13.3770      0.0003    -0.483871
  AGE           -0.0382    0.0185      4.2358      0.0396    -0.268666
```

The output data set RESULTS contains, for each observation, the predicted probability, Pr{Not Improved} and estimated logit for both types of odds. These are distinguished by the variable _LEVEL_. To plot probabilities for both types of improvement in a single graph, the values of TREAT and _LEVEL_ are combined in a single variable. To plot Pr{Improve}, we must reverse the direction of the variables in a data step:

```
data results;
   set results;
   treatl = treat||put(_level_,1.0);
   if _level_=0 then better = (improve > 0);
                else better = (improve > 1);
   *-- Change direction of probabilities & logit;
   predict = 1 - predict;
   lower = 1 - lower;
   upper = 1 - upper;
   logit = -logit;
```

The first few observations in the data set RESULTS after these changes are shown below:

```
   ID  TREAT   SEX IMPROVE _LEVEL_ PREDICT   LOWER   UPPER   LOGIT

   57 Treated Male    1       0      0.267   0.417   0.157  -1.008
   57 Treated Male    1       1      0.129   0.229   0.069  -1.907
    9 Placebo Male    0       0      0.085   0.149   0.048  -2.372
    9 Placebo Male    0       1      0.037   0.069   0.019  -3.271
   46 Treated Male    0       0      0.283   0.429   0.171  -0.932
   46 Treated Male    0       1      0.138   0.238   0.076  -1.831
     ...
```

As in the earlier example, an Annotate data set is used to add more descriptive labels and confidence intervals to the plots. (This adds somewhat more work, but I prefer the plots labeled this way, rather than with legends at the bottom.)

```
proc sort data=results;
   by sex treatl age;
data bars;
   set results;
   by sex treatl;
   length text$8;
   xsys='2'; ysys='2';
   if treat='Placebo' then color='BLACK';
                      else color='RED';
   x  = age; line=33;
   *-- plot confidence limits  ;
   y  = upper;  function='MOVE   '; output;
   text='-';    function='LABEL  '; output;
   y  = lower;  function='DRAW   '; output;
   text='-';    function='LABEL  '; output;
   if last.treatl then do;
```

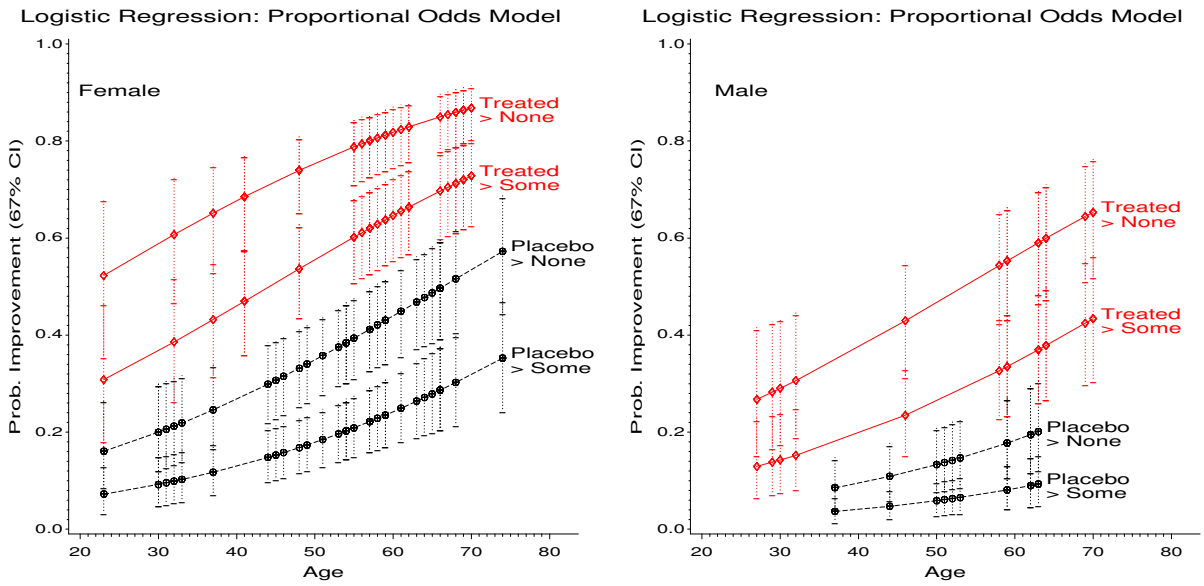


Figure 34: Predicted probabilities for the proportional odds model.

```
      y = predict;
      x = age+1; position='C'; size=1.4;
      text = treat; function='LABEL'; output;
      position='F';
      if _level_ = 0
         then text='> None';
         else text='> Some';;
      output;
      end;
   if first.sex then do;
      ysys ='1'; y=90;
      xsys ='1'; x=10; size=1.5;
      text = sex; function='LABEL'; output;
      end;

goptions hby=0;
proc gplot;
   plot predict * age = treatl / vaxis=axis1 haxis=axis2
                                 nolegend anno=bars    ;
   by sex;
   axis1 label=(h=1.4 a=90 f=duplex 'Prob. Improvement (67% CI)')
         value=(h=1.2) order=(0 to 1 by .2);
   axis2 label=(h=1.4 f=duplex)
         value=(h=1.2) order=(20 to 80 by 10)
         offset=(2,5);
   symbol1 v=+ h=1.4 i=join l=3 c=black;
   symbol2 v=+ h=1.4 i=join l=3 c=black;
   symbol3 v=$ h=1.4 i=join l=1 c=red;
   symbol4 v=$ h=1.4 i=join l=1 c=red;
```

8.6 Polytomous Response: Nested Dichotomies

When the response, y, takes on $m > 2$ discrete values, there are several ways to model the response probabilities. Let $\pi_{ij} \equiv \pi_j(\mathbf{x}_i)$ be the probability of response j for case or group i. Because $\sum_j \pi_{ij} = 1$, only $m - 1$ of these probabilities are independent.

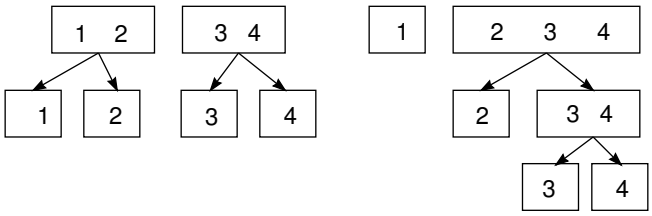


Figure 35: Nested dichotomies. The boxes show two different ways a four-category response can be represented as three nested dichotomies.

The proportional odds model is attractive if the response is ordinal, *and* the proportional odds assumption holds. However, if the response is purely nominal (e.g., vote Tory, Liberal, Reform, NDP), or if the proportional odds assumption is untenable, one particularly simple strategy is to fit separate models to a set of $m-1$ dichotomies derived from the polytomous response.[4]

- Each dichotomy can be fit using the familiar binary-response logistic model, and,
- the $m-1$ models will be statistically independent (so that likelihood-ratio G^2 statistics will be additive) *if* the dichotomies are chosen as ***nested*** dichotomies.

Nested dichotomies are successive binary partitions of the response categories into nested sets. For example, the response categories {1,2,3,4} could be divided first as {1,2} vs. {3,4}. Then these two dichotomies could be divided as {1} vs. {2}, and {3} vs. {4}. Alternatively, these response categories could be divided as {1} vs. {2,3,4}, then {2} vs {3,4}, and finally {3} vs. {4}. (For levels of a factor in an ANOVA design, nested dichotomies correspond to orthogonal contrasts.)

8.6.1 Example: Women's Labour-Force Participation

To illustrate, Fox (1984, 1997) presented data on women's labour-force participation in Canada, where women were classified as not working outside the home (n=155), working part-time (n=42) or working full-time (n=66). Predictor variables were presence/absence of children, and husband's income; a third variable, region of Canada, is not considered here. For these data, it makes sense to model the log odds for two nested dichotomies:

- Working vs. NotWorking
- Working fulltime vs working parttime.

The data are read in as follows shown below. The 3-level variable `labour` is used to define two dichotomous variables, `working`, and `fulltime`. Note that `fulltime` is defined (has non-missing values) only for working women.

```
proc format;
   value labour    /* labour-force participation */
      1 ='working full-time'  2 ='working part-time'
      3 ='not working';
   value kids     /* presence of children in the household */
      0 ='Children absent'  1 ='Children present';
data wlfpart;
   input case labour husinc children region;
   working = labour < 3;
   if working then
      fulltime = (labour = 1);
datalines;
  1  3  15  1  3
  2  3  13  1  3
```

[4] An alternative strategy is to choose one response category (the last, say) as the "base category", and model the ***generalized logits*** for each of categories $j = 1, 2, \ldots, (m-1)$ compared to category m. For a 3-category response, e.g., there are 2 generalized logits, $\text{logit}_{i1} = \log(\pi_{i1}/\pi_{i3})$ and $\text{logit}_{i1} = \log(\pi_{i2}/\pi_{i3})$. These models can be fit using `PROC CATMOD`.

```
 3  3  45  1  3
 4  3  23  1  3
 5  3  19  1  3
 6  3   7  1  3
 7  3  15  1  3
 8  1   7  1  3
 9  3  15  1  3
 ... more data lines ...
```

An initial analysis attempts to fit the proportional odds model, with the 3-level `labour` variable as the response:

```
proc logistic data=wlfpart nosimple;
    model labour  = husinc children ;
   title2 'Proportional Odds Model for Fulltime/Parttime/NotWorking';
```

However, the proportional odds assumption is rejected by the score test:

```
          Score Test for the Proportional Odds Assumption

             Chi-Square = 18.5641 with 2 DF (p=0.0001)
```

Hence, we fit models for each of the `working` and `fulltime` dichotomies. The `descending` option is used so that in each case the probability of a 1 response (working, or fulltime) will be the event modelled.

```
proc logistic data=wlfpart nosimple descending;
    model working = husinc children ;
   output out=resultw p=predict xbeta=logit;
   title2 'Nested Dichotomies';
run;
proc logistic data=wlfpart nosimple descending;
    model fulltime = husinc children ;
   output out=resultf p=predict xbeta=logit;
```

The `output` statements create the datasets `resultw` and `resultf` for plotting the predicted probabilities and logits. The printed output for the working dichotomy includes:

```
                            Response Profile
                         Ordered
                           Value  WORKING     Count

                               1        1       108
                               2        0       155

                 Testing Global Null Hypothesis: BETA=0
                              Intercept
               Intercept          and
   Criterion     Only        Covariates    Chi-Square for Covariates

   AIC           358.151       325.733       .
   SC            361.723       336.449       .
   -2 LOG L      356.151       319.733    36.418 with 2 DF (p=0.0001)
   Score            .             .       35.713 with 2 DF (p=0.0001)
```

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	1.3358	0.3838	12.1165	0.0005	.	.
HUSINC	1	-0.0423	0.0198	4.5751	0.0324	-0.168541	0.959
CHILDREN	1	-1.5756	0.2923	29.0651	0.0001	-0.398992	0.207

To interpret the parameter estimates, note that the odds ratio of 0.959 for husband's income means a 4% decrease in the odds of working with each $1000 increase in husband's income; an additional $10,000 means a decrease in the odds of working by $e^{-.423} = .655$. The effect of having children sililarly corresponds to an odds of working of .207 compared to those without children.

The output for the fulltime vs. parttime dichotomy is shown below. Note that nonworking women are excluded in this analysis.

Response Profile

Ordered Value	FULLTIME	Count
1	1	66
2	0	42

WARNING: 155 observation(s) were deleted due to missing values for the response or explanatory variables.

Testing Global Null Hypothesis: BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	146.342	110.495	.
SC	149.024	118.541	.
-2 LOG L	144.342	104.495	39.847 with 2 DF (p=0.0001)
Score	.	.	35.150 with 2 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	3.4778	0.7671	20.5537	0.0001	.	.
HUSINC	1	-0.1073	0.0392	7.5063	0.0061	-0.424867	0.898
CHILDREN	1	-2.6515	0.5411	24.0135	0.0001	-0.734194	0.071

Thus, the full 3-category response has been fitted by two models,

$$\log\left(\frac{\Pr(\text{working})}{\Pr(\text{not working})}\right) = 1.336 - 0.042\,\text{H\$} - 1.576\,\text{kids} \quad (20)$$

$$\log\left(\frac{\Pr(\text{fulltime})}{\Pr(\text{parttime})}\right) = 3.478 - 0.107\,\text{H\$} - 2.652\,\text{kids} \quad (21)$$

Note that the second equation gives the predicted log odds for fulltime vs. parttime work *conditional* on working.

Moreover, we can add the likelihood ratio or Wald tests across the two models, so the overall test of the hypothesis that neither husband's income nor presence of children predicts working status (the 3-level response)

has a $G^2 = 36.42 + 39.85 = 66.27$ on 2+2=4 df ($p < .0001$). Similarly, the hypothesis that husband's income does not predict working status has a Wald-test $G^2 = 4.58 + 7.51 = 12.09$ on 2 df ($p < .001$).

Comparison of the regression coefficients in the two sub-models (in relation to the size of their standard errors) indicates why the proportional odds model was not tenable. The proportional odds model requires that the coefficients for husband's income and children in the two models are equal. We can see that both variables have a greater effect on the odds of fulltime vs. parttime work than on the odds of working vs. not working.

As usual, these effects can be seen and interpreted more easily in a graph. The odds of working outside the home decrease as husband's income increase and when there are children present. However, among working women, the odds of fulltime vs. parttime work decrease at a faster rate with husband's income; women with children are less likely to work fulltime.

To construct this graph, first join the separate results datasets into one.

```
*-- Merge the results datasets to create one plot;
data both;
   set resultw(in=inw)
       resultf(in=inf);
   if inw then do;
      if children=1 then event='Working, with Children ';
      else event='Working, no Children ';
   end;
   else do;
      if children=1 then event='Fulltime, with Children ';
      else event='Fulltime, no Children ';
  end;
```

Then, we can plot the log odds (or predicted probability) against husband's income, using `event` as to determine the curves to be joined and labelled (the Annotate data step for the labels is not shown here).

```
proc gplot data=both;
   plot logit * husinc = event /
        anno=lbl nolegend frame vaxis=axis1;
   axis1 label=(a=90 'Log Odds') order=(-5 to 4);
   title2 'Working vs Not Working and Fulltime vs. Parttime';
   symbol1 v=dot    h=1.5 i=join l=3 c=red;
   symbol2 v=dot    h=1.5 i=join l=1 c=black;
   symbol3 v=circle h=1.5 i=join l=3 c=red;
   symbol4 v=circle h=1.5 i=join l=1 c=black;
```

8.6.2 Generalized Logits

The generalized logit approach models the probabilities of the m response categories directly as a set of $m-1$ logits comparing each of the first $m-1$ categories to the last category, which serves as the baseline. The logits for any other pair of categories can be retrieved from the $m-1$ fitted ones.

When there are k predictors, $x_1, x_2, \ldots, x_k$, which may, in principle, be quantitative or categorical, the generalized logit model expresses the logits as

$$L_{jm} \equiv \log \frac{\pi_{ij}}{\pi_{im}} = \beta_{0j} + \beta_{1j}\, x_{i1} + \beta_{2j}\, x_{i2} + \cdots + \beta_{kj}\, x_{ik} \quad \text{for } j = 1, 2, \ldots, m-1 \tag{22}$$

$$= \beta_j \prime \mathbf{x}_i \tag{23}$$

Thus, there is one set of fitted coefficients, β_j for each response category except the last. Each coefficient, β_{hj}, gives the effect on the log odds of a unit change in the predictor x_h that an observation belongs to category j, as opposed to category m.

The probabilities themselves are given by

$$\pi_{ij} = \frac{\exp(\beta_j \prime \mathbf{x}_i)}{\sum_{i=1}^{m} \exp(\beta_j \prime \mathbf{x}_i)} \tag{24}$$

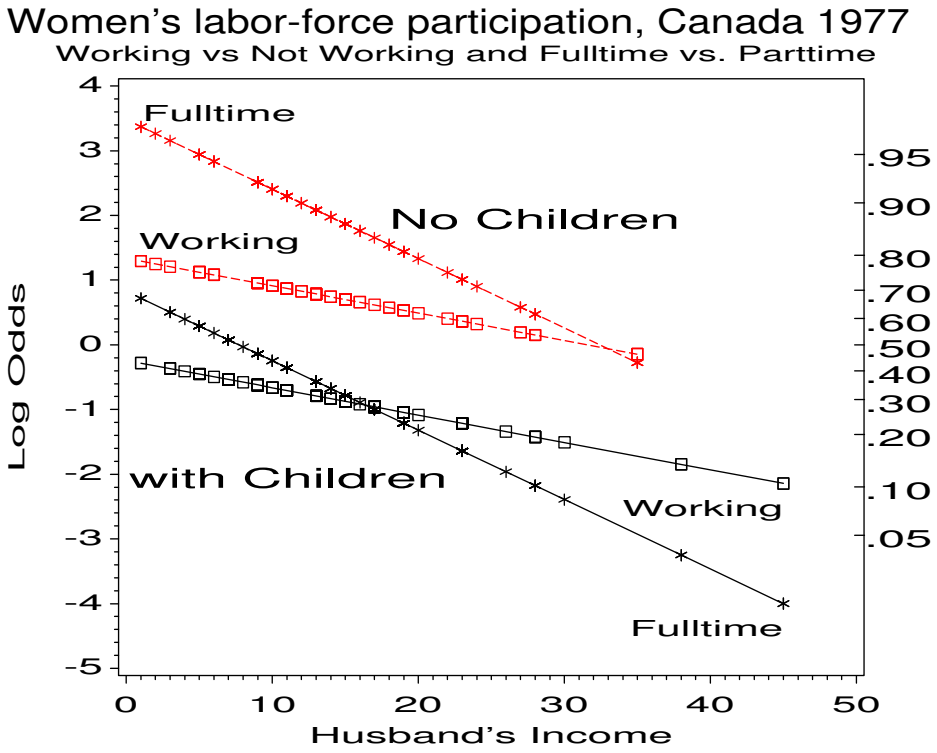


Figure 36: Log odds of working vs. not working, and of fulltime work vs. parttime work.

Parameters in the $m-1$ equations (22) can be used to determine the parameters or the predicted log odds for any pair of response categories by subtraction. For instance, for an arbitrary pair of categories, a and b, and two predictors, x_1 and x_2,

$$L_{ab} = \log \frac{\pi_{ia}/\pi_{im}}{\pi_{ib}/\pi_{im}} \quad (25)$$

$$= \log \frac{\pi_{ia}}{\pi_{im}} - \log \frac{\pi_{ib}}{\pi_{im}} \quad (26)$$

$$= (\beta_{0a} - \beta_{0n}) + (\beta_{1a} - \beta_{1b})x_{1i} + (\beta_{2a} - \beta_{2b})x_{2i} \quad (27)$$

Thus, for example, the coefficient for x_{1i} in L_{ab} is just $(\beta_{1a} - \beta_{1b})$. Similarly, the predicted logit for any pair of categories can be calculated as

$$\hat{L}_{ab} = \hat{L}_{am} - \hat{L}_{bm} \quad (28)$$

The generalized logit model cannot be fit using `PROC LOGISTIC` [5] but it can be fit using `PROC CATMOD`.[6] An output data set provides all predicted probabiliites, and the fitted logits.

To illustrate, we fit the generalized logit model to the women's labor force participation data using the statements below. Husband's income is treated as a quantitative variable by declaring it on the `direct` statement. `PROC CATMOD` does not provide an overall test of the whole model, however this can be carried out with a contrast statement to test $H_0 : \beta = 0$.

```
proc catmod data=wlfpart;
   direct husinc;
   model labour = husinc children / noprofile noiter;
```

[5] This model can now (SAS V8.2) be fit in `PROC LOGISTIC` with the `LINK=GLOGIT` option on the `MODEL` statement.

[6] When one or more of the predictor variables are continuous, however, you may have difficulty due to zero cell frequencies, because `PROC CATMOD` treats the data as a contingency table. In this case, it may help to reorder the response variable so that the response category with the highest frequency is the last, baseline category. Alternatively, the continuous variable(s) can be collapsed into categories so that populations with zero frequencies do not occur.

```
response logits / out=results;
contrast 'Husinc,Children=0'
   husinc   1,
   children 1;
```

The maximum likelihood ANOVA table shows that there are two parameters fit for each regressor. With a continuous predictor, the likelihood-ratio test of goodness-of-fit, which compares the current model to the saturated model, is unreliable, because the contingency table is very sparse.

```
          MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

       Source                    DF   Chi-Square      Prob
       ---------------------------------------------------
       INTERCEPT                  2        15.91    0.0004
       HUSINC                     2        12.82    0.0016
       CHILDREN                   2        53.98    0.0000

       LIKELIHOOD RATIO          86       138.67    0.0003
```

The table of parameter estimates, shown below, contains the coefficients for the two fitted logits,

$$\log\left(\frac{\Pr(\text{fulltime})}{\Pr(\text{not working})}\right) = 0.7035 - 0.0972\,\text{H\$} + 1.2793\,\text{kids} \tag{29}$$

$$\log\left(\frac{\Pr(\text{parttime})}{\Pr(\text{not working})}\right) = -1.4216 + 0.00689\,\text{H\$} - 0.0107\,\text{kids} \tag{30}$$

The predicted log odds for working fulltime as opposed to parttime are therefore given by

$$\log\left(\frac{\Pr(\text{fulltime})}{\Pr(\text{not working})}\right) = 2.1251 - 0.1041\,\text{H\$} + 1.29\,\text{kids}$$

These coefficients are not directly comparable to those in (21) and (21) for the nested dichotomies models.

```
             ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

                                          Standard    Chi-
  Effect            Parameter  Estimate    Error    Square   Prob
  ------------------------------------------------------------------
  INTERCEPT                 1    0.7035   0.4140      2.89  0.0892
                            2   -1.4216   0.4528      9.86  0.0017
  HUSINC                    3   -0.0972   0.0281     11.98  0.0005
                            4   0.00689   0.0235      0.09  0.7689
  CHILDREN                  5    1.2793   0.1811     49.90  0.0000
                            6   -0.0107   0.2345      0.00  0.9635
```

A plot of the predicted probabilities of the three categories of `labour` is easily obtained from the `results` data set produced by `PROC CATMOD`. This data set contains both fitted probabilities (`_type_='PROB'`) and fitted logits (`_type_='FUNCTION'`), so we select the `_type_='PROB'` observations with a `where` statement.

```
proc gplot data=results;
   where (_type_='PROB');
   plot _pred_ * husinc = labour /
      vaxis=axis1 hm=1 vm=1 anno=labels nolegend;
   by children;
   axis1 order=(0 to .9 by .1) label=(a=90);
   symbol1 i=join v=circle   c=black;
   symbol2 i=join v=square   c=red;
   symbol3 i=join v=triangle c=blue;
   label _pred_='Fitted probability';
```

The fitted probabilities are shown in Figure 37. When there are no young children in the home, a woman's probability of not working rises sharply with husband's income, while her probability of working full time declines sharply, and her probability of working part time increases modestly. With children present, the direction of these relations with husband's income are the same, however, the levels of not working and full time work are reversed.

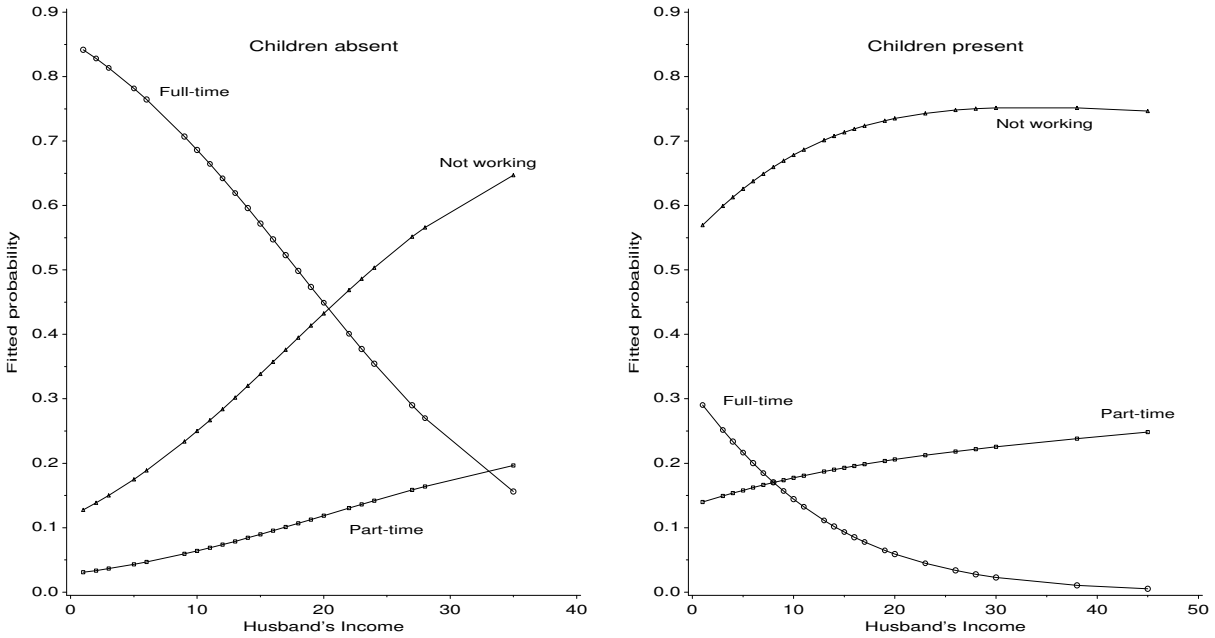


Figure 37: Fitted probabilities for the generalized logit model

8.7 Influence statistics and diagnostic plots

In ordinary least squares regression, measures of influence (leverage, Cook's D, DFBETAs, etc.) help you to determine whether individual cases have undue impact on the fitted regression model and the coefficients of individual predictors. Analogs of most of these measures have been suggested for logistic regression. (Some additional problems occur in logistic regression because the response is discrete, and because the leave-one-out diagnostics are more difficult to compute.)

8.7.1 Influence measures & diagnostics

leverage Measures the potential impact of an individual case on the results, which is directly proportional to how far an individual case is from the centroid in the space of the predictors. Leverage is computed as the diagonal elements, h_{ii}, of the "Hat" matrix, $\mathbf{H}$,

$$\mathbf{H} = \mathbf{X}^{\star}(\mathbf{X}^{\star\prime}\mathbf{X}^{\star})^{-1}\mathbf{X}^{\star\prime}$$

where $\mathbf{X}^{\star} = \mathbf{V}^{1/2}\mathbf{X}$, and $\mathbf{V} = \mathrm{diag}\hat{\mathbf{P}}(1-\hat{\mathbf{P}})$. As in OLS, leverage values are between 0 and 1, and a leverage value, $h_{ii} > 2k/n$ is considered "large"; k = number of predictors, n = number of cases.

residuals Pearson and deviance residuals are useful in identifying observations that are not explained well by the model. They are the (signed) square roots of the contribution that each case makes to the overall Pearson and deviance goodness-of-fit statistics.

influence measures Measure the effect that deleting an observation has on the regression parameters or the goodness-of-fit statistics. (The terminology for these measures is not well-established, so I use the terms from the *SAS/STAT User's Guide*.)

C

CBAR These are analogs of the Cook's D statistic in OLS. They are two standardized measures of the approximate change in all regression coefficients when the i-th case is deleted.

DIFCHISQ

DIFDEV Approximate measures of the amount $\Delta\chi^2$ by which the Pearson χ^2 (DIFCHISQ) or the likelihood-ratio deviance χ^2 (DIFDEV) would decrease if the i-th case were deleted. Values > 4 for either indicate "significant" change (since these are 1 df χ^2 or squared Normal values).

DFBETAij Measure the change in the logistic regression coefficient b_j for predictor x_j when observation i is deleted.

8.7.2 Printed output

All the influence statistics are printed when the `influence` option is used on the `MODEL` statement. For example,

```
proc logistic data=arthrit ;
   model better = _sex_ _treat_ _age_ / influence;
```

This produces a great deal of output, including the following:

```
                    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  |         *       |
  28    0.619  |        | *      |    0.053  |      *          |
  29    1.120  |        |   *    |    0.141  |                *|
  30    1.846  | *      |        |    0.052  |      *          |
  31    1.309  |        |    *   |    0.092  |          *      |
  32    0.647  |        | *      |    0.050  |      *          |
  33    0.955  |    *   |        |    0.070  |        *        |
  34    1.803  | *      |        |    0.049  |      *          |
```

8.7.3 Diagnostic plots of influence measures

Plots of the change in χ^2 (DIFCHISQ or DIFDEV) against either leverage or predicted probability are useful for detecting unduly influential cases. The actual influence of each case on the estimated coefficients (C or CBAR) can be shown in a bubble plot where the plotting symbols are circles proportional to C or CBAR.

Such plots are produced by the INFLOGIS macro. For example, these statements produce plots of DIFCHISQ against both the leverage (HAT) (Figure 38) and predicted probability (PRED) (Figure 39) using bubbles whose area is proportional to C:

```
title 'Arthritis treatment data';
title2 'Bubble size: Influence on Coefficients (C)';
goptions htext=1.6;
%include inflogis;
%include arthrit;
%inflogis(data=arthrit,
     y=better,               /* response   */
     x=_sex_ _treat_ age,    /* predictors */
     id=id,
     gy=DIFCHISQ,            /* graph ordinate  */
     gx=PRED HAT,            /* graph abscissas */
     lcolor=RED, bsize=14
     );
run;
```

The printed output from the INFLOGIS program includes a table identifying any observation of high leverage or influence. For example, case 29 is of high leverage, because she is unusual in terms of the predictors: a young woman given treatment; however, she is not influential in the fitted model. Case 77, is not of high leverage, but is poorly predicted by the model and has large contributions to χ^2.

CASE	BETTER	_SEX_	_TREAT_	AGE	HAT	DIFCHISQ	DIFDEV	C
1	1	0	1	27	.09	4.5781	3.6953	0.4510
22	1	0	0	63	.06	4.4603	3.5649	0.2898
29	0	1	1	23	.14	1.0183	1.4005	0.1679
30	1	1	0	31	.05	4.7485	3.6573	0.2611
34	1	1	0	33	.05	4.2955	3.4644	0.2236
55	0	1	1	58	.03	4.9697	3.6759	0.1602
77	0	1	1	69	.03	8.4977	4.7122	0.2758

9 Plots for logit models

A contingency table gives the joint distribution of two or more discrete, categorical variables. In a two-way table, one typically uses the χ^2 test of association to determine if the row and column variables can be considered independent. Loglinear and logit models generalize this test of association to three- and higher-way tables.

A log-linear model expresses the relationship among all variables as a model for the log of the expected cell frequency. For example, for a three-way table, the hypothesis that of no three-way association can be expressed as the log-linear model,

$$\log m_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} \tag{31}$$

The log-linear model treats the variables symmetrically: none of the variables is distinguished as a response variable. However, the association parameters may be difficult to interpret, and the absence of a dependent variable makes it awkward to plot results in terms of the log-linear model. In this case, correspondence analysis and the mosaic display may provide a simpler way to display the patterns of association in a contingency table.

On the other hand, if one variable can be regarded as a response or dependent variable, and the others as independent variables, then the effects of the independent variables may be expressed as a logit model. For example, if variable C is a binary response, then model (31) can be expressed as an equivalent logit model,

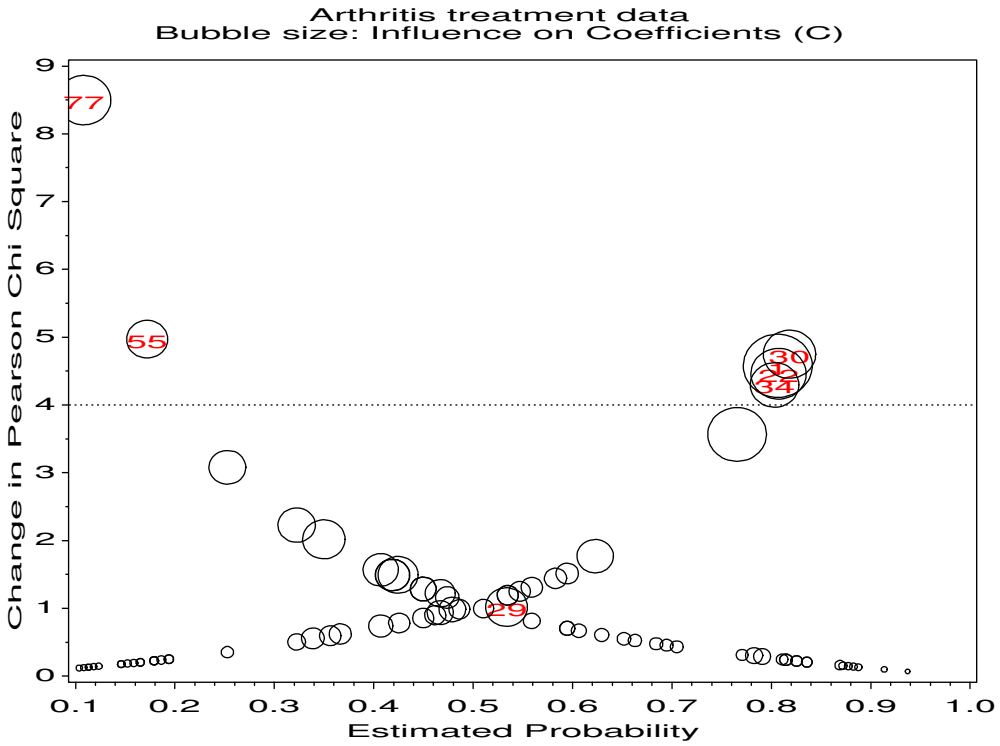


Figure 38: Influence plot for arthritis data. Cases with DIFCHISQ > 4 or leverage $> (2k)/n = 0.095$ are influential, as indicated by the size of the bubble symbol.

$$\begin{aligned} \text{logit}_{ij} = \log(m_{ij1}/m_{ij2}) &= (\lambda_1^C - \lambda_2^C) + (\lambda_{i1}^{AC} - \lambda_{i2}^{AC}) + (\lambda_{j1}^{BC} - \lambda_{j2}^{BC}) \\ &= \alpha + \beta_i^A + \beta_j^B \end{aligned} \tag{32}$$

where, because all λ terms sum to zero, $\alpha = 2\lambda_1^C$, $\beta_i^A = 2\lambda_{i1}^{AC}$, and $\beta_j^B = 2\lambda_{j1}^{BC}$.

Both log-linear and logit models can be fit using `PROC CATMOD` in SAS. For logit models, the steps for fitting a model and plotting the results are similar to those used for logistic models with PROC LOGISTIC. The main differences are:

- For `PROC CATMOD` the independent variables can be categorical or quantitative, character or numeric. With `PROC LOGISTIC` the independent variables must be numeric. To use a categorical variables, we had to construct dummy variables.
- The input data set is arranged differently.
- The output data set from `PROC CATMOD` contains similar information, but the variable names are different and the information is arranged differently.

9.1 Example

The table below shows data from the 1982 General Social Survey on votes in the 1980 US Presidential election for Reagan or for Carter or other in relation to race and political view (1=most liberal, 7=most conservative).

Political	---- White ----		--- Nonwhite ---	
View	Reagan	Carter	Reagan	Carter
1	1	12	0	6
2	13	57	0	16
3	44	71	2	23

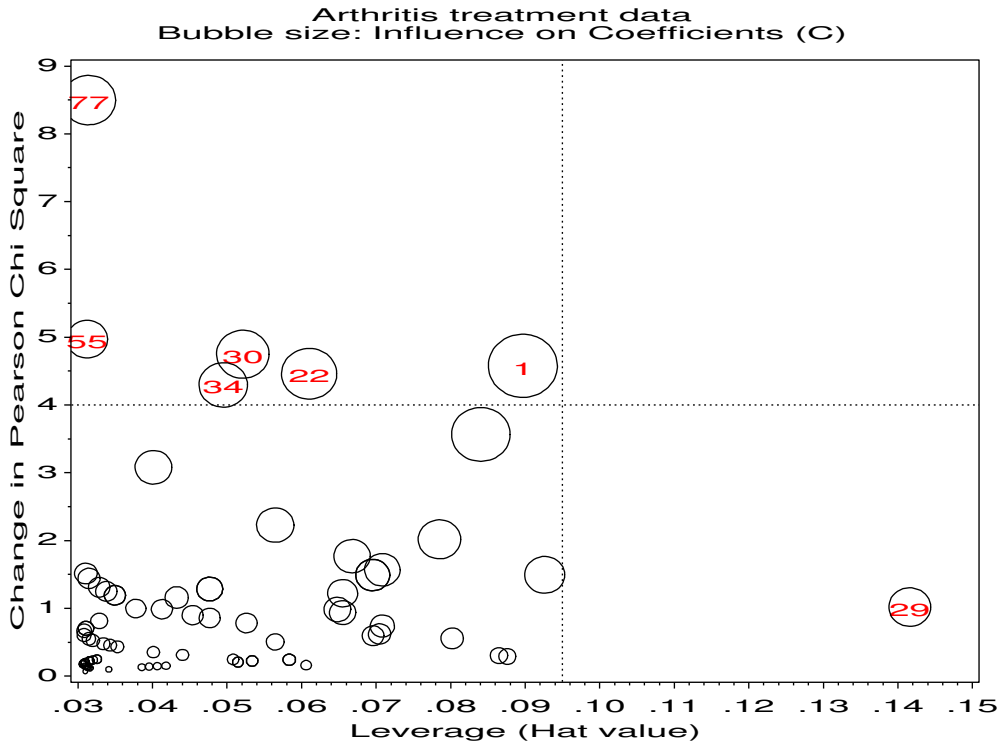


Figure 39: Changes in chi-square vs. predicted probability. The plot shows that most of the influential observations are those with very high or low predicted probabilities. The systematic pattern shown is inherent in the discrete nature of logistic regression.

```
4          155         146          1          31
5           92          61          0           8
6          100          41          2           7
7           18           8          0           4
```

Treating the vote for Reagan vs. Carter or other as the response, a logit model with nominal main effects for race and political view is

$$\text{logit (Reagan / Carter)} = \alpha + \beta_i^{RACE} + \beta_j^{VIEW} \tag{33}$$

This model does not use the ordinal nature of political view. A model which uses the value of political view as a direct, quantitative independent variable can be expressed as

$$\text{logit (Reagan / Carter)} = \alpha + \beta_i^{RACE} + \beta^{VIEW}\, view \tag{34}$$

9.2 Fitting the model with nominal main effects

The data are first read in to a data set `vote` in the frequency form that could be used as input to `PROC LOGISTIC`. `PROC CATMOD`, however, requires an explicit dependent variable and a separate observation for each value of the response. A second data step creates the response variable `votefor`.

```
proc format;
  value race 0='NonWhite'
             1='White';
data vote;
  input @10 race view reagan carter;
  format race race.;
  reagan= reagan + .5;        *-- allow for 0 values;
  carter= carter + .5;
```

```
  total = reagan + carter;
  preagan = reagan / total;
  logit = log ( reagan / carter);
datalines;
White    1 1    1   12
White    1 2   13   57
White    1 3   44   71
White    1 4  155  146
White    1 5   92   61
White    1 6  100   41
White    1 7   18    8
NonWhite 0 1    0    6
NonWhite 0 2    0   16
NonWhite 0 3    2   23
NonWhite 0 4    1   31
NonWhite 0 5    0    8
NonWhite 0 6    2    7
NonWhite 0 7    0    4
;
data votes;
   set vote;
   votes= reagan; votefor='REAGAN'; output;
   votes= carter; votefor='CARTER'; output;
```

Model (33) is fit using the statements below. The RESPONSE statement is used to produce an output data set, PREDICT, for plotting.

```
proc catmod data=votes order=data;
   weight votes;
   response / out=predict;
   model votefor = race view / noiter ;
   title2 f=duplex h=1.4
      'Nominal Main Effects of Race and Political View (90% CI)';
```

The results of the PROC CATMOD step include:

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	43.75	0.0000
RACE	1	41.37	0.0000
VIEW	6	67.84	0.0000
LIKELIHOOD RATIO	6	3.45	0.7501

ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	-1.4324	0.2166	43.75	0.0000
RACE	2	1.1960	0.1859	41.37	0.0000
VIEW	3	-1.6144	0.6551	6.07	0.0137
	4	-1.2000	0.2857	17.64	0.0000
	5	-0.1997	0.2083	0.92	0.3377
	6	0.2779	0.1672	2.76	0.0965
	7	0.6291	0.1941	10.51	0.0012
	8	1.1433	0.2052	31.03	0.0000

The data set PREDICT contains observed (_OBS_) and predicted (_PRED_) values, and estimated standard errors. There are 3 observations for each race-view group: logit values have _TYPE_= 'FUNCTION'; probabilities have _TYPE_ = 'PROB'.

RACE	VIEW	_TYPE_	_NUMBER_	_OBS_	_PRED_	_SEPRED_
White	1	FUNCTION	1	-2.120	-1.851	0.758
White	1	PROB	1	0.107	0.136	0.089
White	1	PROB	2	0.893	0.864	0.089
White	2	FUNCTION	1	-1.449	-1.437	0.297
White	2	PROB	1	0.190	0.192	0.046
White	2	PROB	2	0.810	0.808	0.046
...						

To plot the fitted logits, select the _TYPE_ = 'FUNCTION' observations in a data step.

```
data predict;
   set predict;
   if _type_ = 'FUNCTION';
```

A simple plot of predicted logits can then be obtained with the following PROC GPLOT step. (The plots displayed use the Annotate facility to add 90% confidence limits, calculated as _pred_ $\pm$ 1.645 _sepred_, and a probability scale at the right as illustrated earlier.)

```
proc gplot data=predict;
   plot _pred_ * view = race
        / haxis=axis1 hminor=0 vaxis=axis2;
   symbol1 i=none v=+      h=1.5 c=black;
   symbol2 i=none v=square h=1.5 c=red  ;
   axis1 label=(h=1.4 'Conservativism') offset=(2);
   axis2 order=(-5 to 2) offset=(0,3)
         label=(h=1.4 a=90 'LOGIT(Reagan / Carter)');
```

9.3 Other models

Again the predicted values in the output data set depend purely on the model specified in the PROC CATMOD step. The plotting steps remain the same.

For example, to test and plot results under the assumption that political view has a linear effect on the logit scale, as in model (34), we use the same MODEL statement, but specify VIEW as a direct (quantitative) predictor.

```
proc catmod data=votes order=data;
   direct view;
   weight votes;
   response / out=predict;
   model votefor = race view / noiter ;
   title2 'Linear Effect for Political View (90% CI)';
   run;
```

The results indicate that this model fits nearly as well as the nominal main effects model, and is preferred since it is more parsimonious.

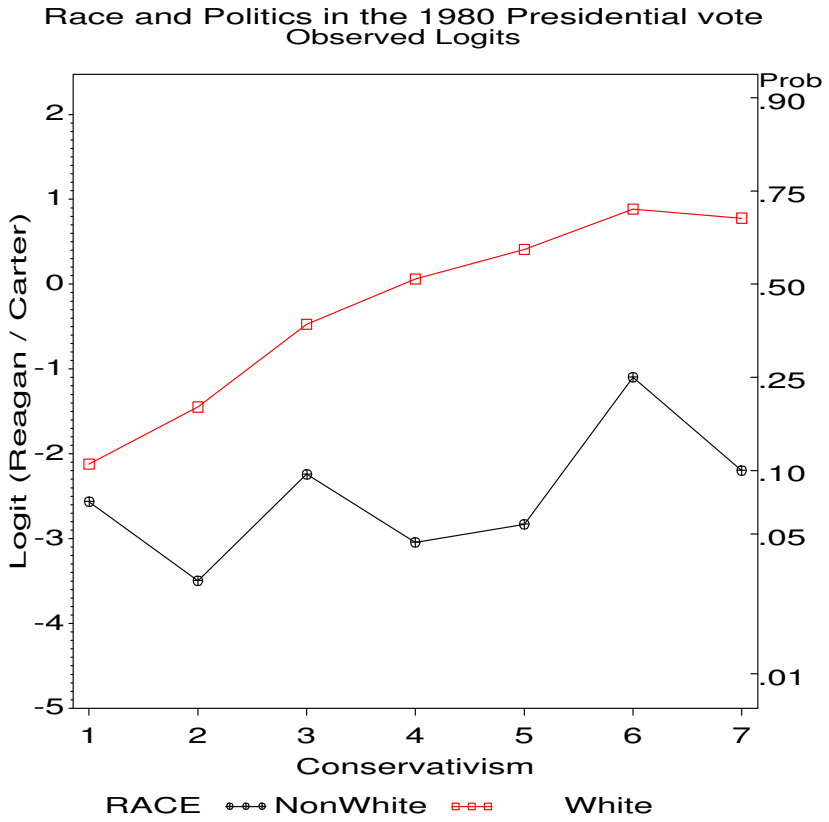


Figure 40: Observed log odds.

```
            MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

          Source                    DF   Chi-Square      Prob
          ------------------------------------------------------
          INTERCEPT                  1       101.39    0.0000
          RACE                       1        42.86    0.0000
          VIEW                       1        67.13    0.0000

          LIKELIHOOD RATIO          11         9.58    0.5688

               ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

                                            Standard    Chi-
    Effect            Parameter  Estimate    Error    Square   Prob
    ---------------------------------------------------------------
    INTERCEPT                 1   -3.1645    0.3143   101.39  0.0000
    RACE                      2    1.2213    0.1865    42.86  0.0000
    VIEW                      3    0.4719    0.0576    67.13  0.0000
```

To test whether a single slope for political view, β^{VIEW} is adequate for both races, fit a model which allows separate slopes (an interaction between race and view):

```
proc catmod data=votes order=data;
   direct view;
   weight votes;
   response / out=predict;
   model votefor = race view race*view;
```

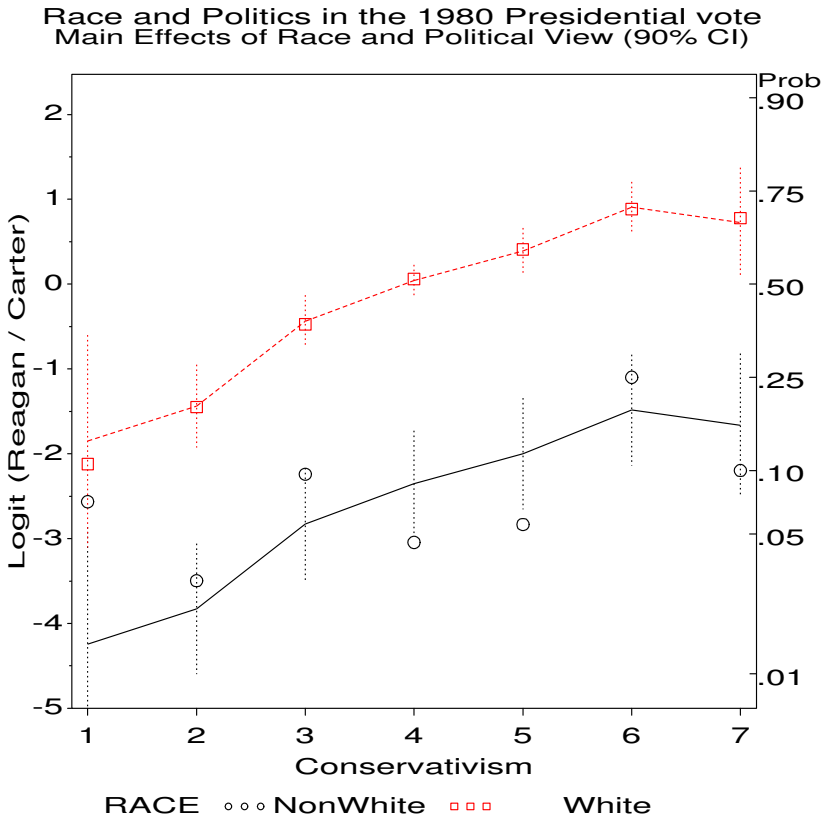


Figure 41: Fitted logits for main effects model. Points show observed logits; dotted lines show a 90% confidence interval around the predicted log odds.

```
title2 'Separate slopes for Political View (90% CI)';
```

The results show that this model does not offer a significant improvement in goodness of fit. The plot nevertheless indicates a slightly steeper slope for white voters than for non-white voters.

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	26.55	0.0000
RACE	1	2.02	0.1556
VIEW	1	9.91	0.0016
VIEW*RACE	1	0.78	0.3781
LIKELIHOOD RATIO	10	8.81	0.5504

ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	-2.7573	0.5351	26.55	0.0000
RACE	2	0.7599	0.5351	2.02	0.1556
VIEW	3	0.3787	0.1203	9.91	0.0016
VIEW*RACE	4	0.1060	0.1203	0.78	0.3781

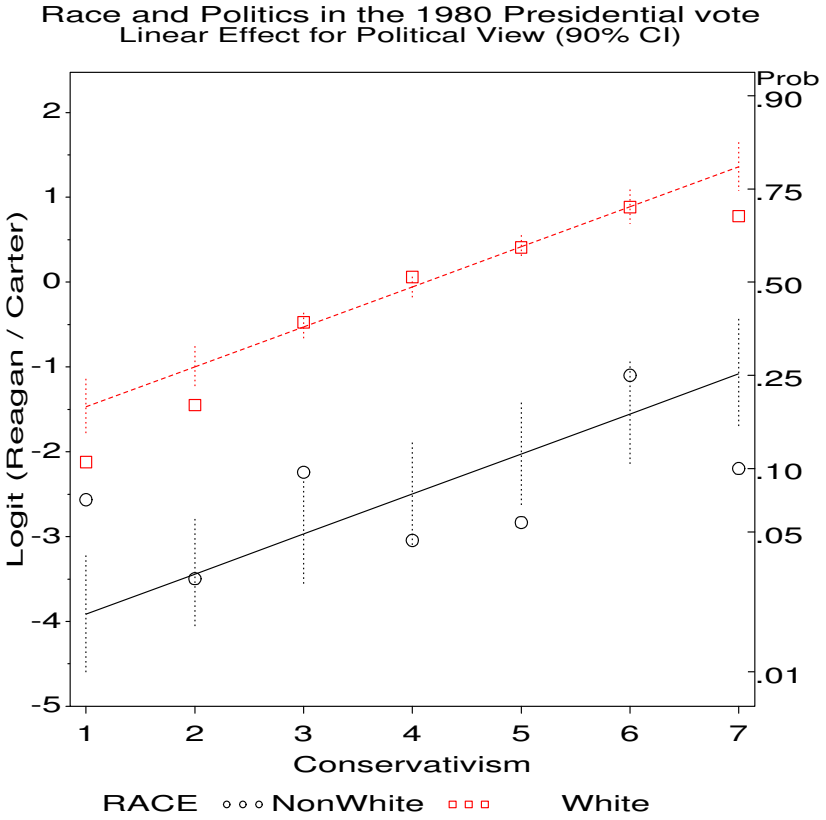


Figure 42: Fitted logits for linear effect of Conservativism.

10 Loglinear models

Whereas logit models focus on the prediction of one response factor, log-linear models treat all variables symmetrically and attempt to model all important associations among them. In this sense, log-linear models are analogous to correlation analysis of continuous variables where the goal is to determine the patterns of dependence and independence among a set of variables.

Formally, however, log-linear models are closely related to ANOVA models for quantitative data. For two variables, A and B, for example, the hypothesis of independence means that the expected frequencies, m_{ij} obey

$$m_{ij} = \frac{m_{i+}\, m_{+j}}{m_{++}}$$

This multiplicative model can be transformed to an additive (linear) model by taking logarithms of both sides:

$$\log(m_{ij}) = \log(m_{i+}) + \log(m_{+j}) - \log(m_{++})$$

which is usually expressed in an equivalent form in terms of model parameters,

$$\log(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B \tag{35}$$

where mu is a function of the total sample size, λ_i^A is the "main effect" for variable A, and λ_j^B is the "main effect" for variable B, and the same analysis of variance restrictions are applied to the parameters: $\sum_i \lambda_i^A = \sum_j \lambda_j^B = 0$. The main effects in log-linear models pertain to differences among the marginal probabilities of a variable. Except for differences in notation, the model (35) is formally identical to the ANOVA main-effects model for a two-factor design:

$$E(y_{ij}) = \mu + \alpha_i + \beta_j$$

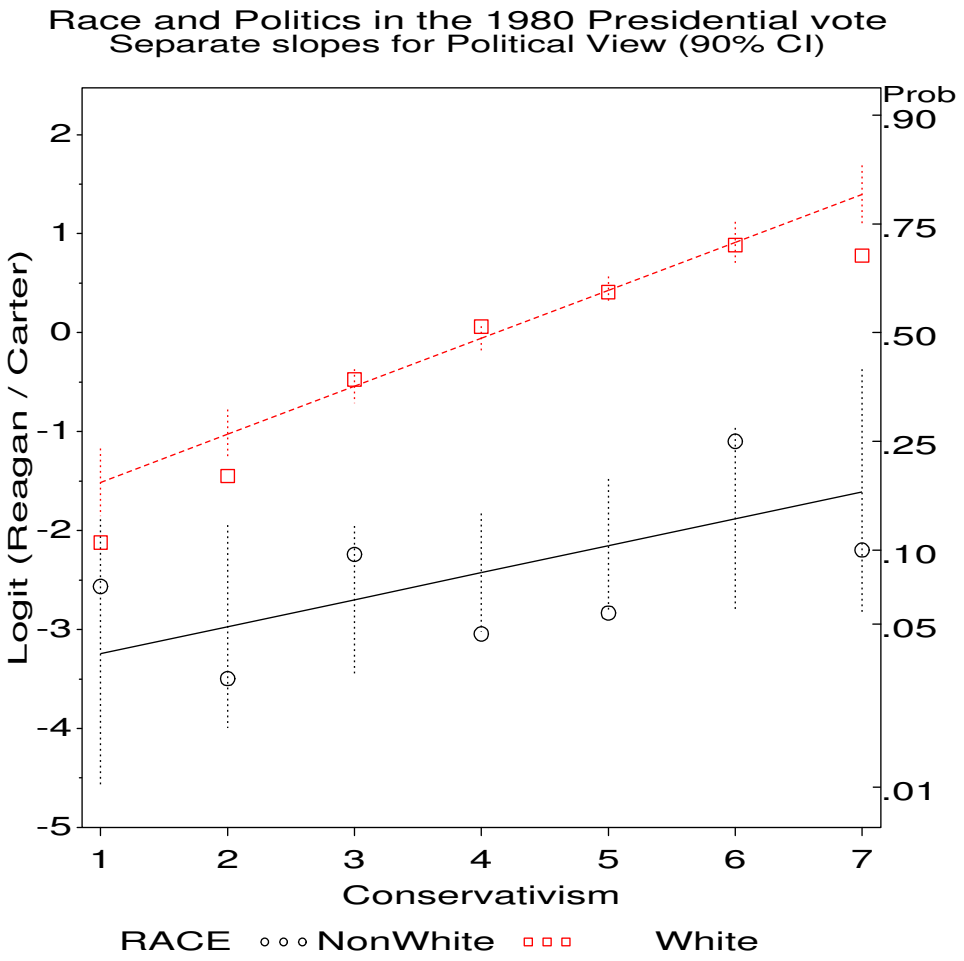


Figure 43: Fitting separate slopes.

For a two-way table, a model which allows an association between the variables is (the *saturated model*):

$$\log(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \tag{36}$$

which is similar to the two-factor ANOVA model with interaction:

$$E(y_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

Hence, associations between variables in log-linear models are analogous to interactions in ANOVA models. For example, in the Berkeley admissions data, the model

$$\log\, m_{ijk} = \mu + \lambda_i^A + \lambda_j^D + \lambda_k^G + \lambda_{ij}^{AD} + \lambda_{ik}^{AG} + \lambda_{jk}^{DG} \tag{37}$$

includes effects for associations between:

- Admission and Department (different admission rates across departments),
- Admission and Gender (which, if significant, would indicate gender-bias), and
- Department and Gender (males and females apply differentially across departments).

Such log-linear models are usually hierarchical: lower-order relatives (e.g., λ_i^A, λ_j^D) of the highest-order terms (λ_{ij}^{AD}) are (almost always) included in the model. Consequently, it is necessary only to specify the high-order terms in the model, and model (37) can be denoted as $[AD][AG][DG]$, or as $(AD,\ AG,\ DG)$. Since Admission is a binary response variable, however, model (37) is also equivalent to the logit model (38),

$$\log\left(\frac{m_{Admit(ij)}}{m_{Reject(ij)}}\right) = \alpha + \beta_i^{Dept} + \beta_j^{Gender} \tag{38}$$

10.1 Fitting Loglinear Models

10.1.1 Strategy

Fitting a log-linear model is a process of deciding which associations are significantly different from zero; these terms are included in the final model used to explain the observed frequencies. Terms which are excluded from the model go into the residual or error term, which reflects the overall badness-of-fit of the model. The usual goal of log-linear modelling is to find a small model (few terms) which nonetheless achieves a reasonable fit (small residual).

To illustrate the reasoning here, consider the log-linear model

$$\log\, m_{ijk} = \mu + \lambda_i^{Sex} + \lambda_j^{Treat} + \lambda_k^{Outcome} + \lambda_{ij}^{SexTreat} + \lambda_{ik}^{SexOutcome} + \lambda_{jk}^{TreatOutcome}. \tag{39}$$

(or [SexTreat] [SexOutcome] [TreatOutcome], or [ST] [SO] [TO] in abbreviated notation) which was fit to test whether the association between treatment and outcome in the arthritis data was homogeneous across sex. This model includes the main effect of each variable and all two-way associations, but not the three-way association of sex * treat * outcome. The three-way term [STO] therefore forms the residual for this model, which is tested by the likelihood ratio G^2 below.

```
                No 3-way association
   MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source                    DF   Chi-Square      Prob
-------------------------------------------------
SEX                        1        14.13    0.0002
TREAT                      1         1.32    0.2512
SEX*TREAT                  1         2.93    0.0871
IMPROVE                    2        13.61    0.0011
SEX*IMPROVE                2         6.51    0.0386
TREAT*IMPROVE              2        13.36    0.0013

LIKELIHOOD RATIO           2         1.70    0.4267
```

Because the likelihood ratio $G^2 = 1.70$ is not significant I can conclude this is an acceptable model. However, it may not be necessary to include all the remaining terms in the model, and the small χ^2 values for the terms [ST] and [SO] suggested that I might refit a reduced model excluding them.

10.1.2 Test statistics

The **overall** goodness-of-fit of a log-linear model is assessed by comparing the observed cell frequencies ($n_{ij...}$) to the estimated expected frequencies under the model ($\hat{m}_{ij...}$). The usual Pearson chi-square statistic, $\chi^2 = \sum(n - \hat{m})^2/\hat{m}$ can be used for this test of fit, but the likelihood ratio G^2, defined as

$$G^2 = 2 \sum n \log_e \frac{n}{\hat{m}}$$

is more commonly used, since it is the statistic that is minimized in maximum likelihood estimation.

The chi-square statistics printed for the *individual terms* in a log-linear model are based on ***Wald tests***, a general methodology for testing whether a subset of the parameters in a model are zero. Wald tests are analogous to the Type III (partial) tests used in ANOVA and regression models. They test they hypothesis that the given terms can be deleted from the model, given that all other terms are retained.

10.1.3 Software

SPSS offers the LOGLINEAR procedure and the GENLOG and HILOGLINEAR procedures (for hierarchical models). The HILOGLINEAR procedure provides several model selection methods (forward or backward selection); the LOGLINEAR procedure allows variables to be treated as quantitative ones.

With SAS, log-linear models can be fit using `PROC CATMOD`, a very general procedure for categorical modelling. Loglinear models can also be fit with `PROC GENMOD` (as of SAS 6.08), a new procedure for generalized linear models. In SAS/INSIGHT, the Fit (Y X) menu also fits generalized linear models. When one variable is a binary response, you can also fit the equivalent logit model with `PROC LOGISTIC`.

10.2 Using PROC CATMOD

The log-linear model (37) can be fit to the Berkeley data using `PROC CATMOD`. The cell variables and frequency are read in as follows:

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

For log-linear models, the `MODEL` statement should specify `_RESPONSE_` to the right of the = sign; use the `LOGLIN` statement to specify the model to be fit. When the data are in frequency form, as here, use a `WEIGHT` statement statement to specify the frequency variable.

```
proc catmod order=data
            data=berkeley;
   format dept dept. admit admit.;
   weight freq;
   model dept*gender*admit=_response_ /
         ml noiter noresponse nodesign noprofile  pred=freq ;
   loglin admit|dept|gender @2 / title='Model (AD,AG,DG)';
 run;
```

On the LOGLIN statement, the "bar" notation `admit|dept|gender @2` means all terms up to two-way associations. The printed output includes the following, which indicates that only the two-way terms DEPT*ADMIT and DEPT*GENDER are significant. In particular, there is no association between Gender and Admission.

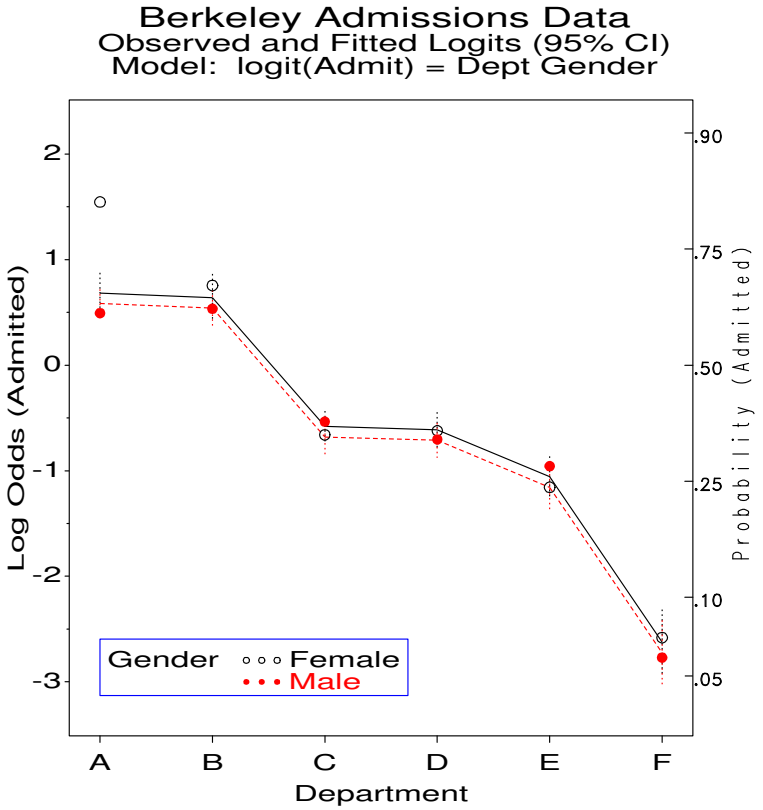


Figure 44: Observed and Fitted Log odds of admission in the model [AD][AG][DG]

Model (AD,AG,DG)
MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
ADMIT	1	262.45	0.0000
DEPT	5	276.37	0.0000
DEPT*ADMIT	5	534.71	0.0000
GENDER	1	197.99	0.0000
GENDER*ADMIT	1	1.53	0.2167
DEPT*GENDER	5	731.62	0.0000
LIKELIHOOD RATIO	5	20.20	0.0011

As usual, these effects are easier to discern in a plot of fitted and observed logits, shown for this model in Figure 44. It is apparent that the effect of gender is negligible.

Consequently, we drop the [AG] term and fit the reduced model $(AD,\ DG)$. The RESPONSE statement produces an output dataset containing observed and predicted frequencies, used in the following section.

```
   response / out=predict;
   loglin admit|dept dept|gender / title='Model (AD,DG)';
 run;
```

The overall fit of the model, shown by the likelihood-ratio $G^2 = 21.7$ on 6 df is not good. This, however, turns out to be due to the data for one department, as we shall see.

For comparison, the lines below show how the model [AD][AG][DG] is fit as the equivalent logit model, and the use of the CATPLOT macro to produce Figure 44.

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

legend1 position=(bottom inside left)  offset=(4,3)
        mode=share cborder=blue across=1
        shape=symbol(6,1.5)
        label=('Gender')
        value=(c=black 'Female'
               c=red   'Male');
%catplot(data=predict, class=gender, xc=dept, z=1.96, anno=pscale,
         legend=legend1);
```

10.3 Influence and diagnostic plots for log-linear models

As in logit models, there are analogs of leverage, Cook's D, and the leave-one-out $\Delta\chi^2$ statistics for log-linear models. Most of these diagnostic quantities are calculated by PROC GENMOD The INFLGLIM macro is provided for calculating additional diagnostics (hat values and Cook's D) supplied by PROC GENMOD and for producing useful plots of these measures. Models fit using PROC CATMOD require a bi more work for diagnostic displays.

10.3.1 Model diagnostics with PROC GENMOD and the INFLGLIM macro

Most of the model diagnostics are calculated by PROC GENMOD and made available for plotting by use of the statement MAKE 'OBSTATS' OUT= and the options OBSTATS RESIDUALS on the MODEL statement. A loglinear model is fit with PROC GENMOD as a linear model for the cell frequency, with a Poisson distribution for errors.

These diagnostic plots may be illustrated with the Berkeley admissions data using the loglinear model $[AD][GD]$, or the equivalent logit model, $\text{logit}(\text{Admit}) = \alpha + \beta_i^D$. Recall that we found this model fit well, except in department A. To give useful labels for influential cells, we first combine the factor variables into a charcter identifier, CELL.

```
data berkeley;
   set berkeley;
   cell = trim(put(dept,dept.)) ||
          gender ||
          trim(put(admit,yn.));
```

We would fit this model using PROC GENMOD and obtain observation statistics as follows. Note that the response variable is FREQ, and the CLASS statement handles the categorical variables.

```
proc genmod data=berkeley;
   class dept gender admit;
   model freq = dept|gender dept|admit / dist=poisson obstats residuals;
   make 'obstats' out=obstats noprint;
```

The same model is fit with the INFLGLIM macro. We ask for an influence plot of adjusted Pearson residuals against hat values (showing Cook's D by bubble size, by default):

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

The plot (Figure 45) clearly indicate that the only cells which do not fit ($|r_i| > 2$) are for department A. Notice also, that the cells for males applying to this department (with high expected frequencies) have have large leverage, and therefore large influence (Cook's D) on this model.

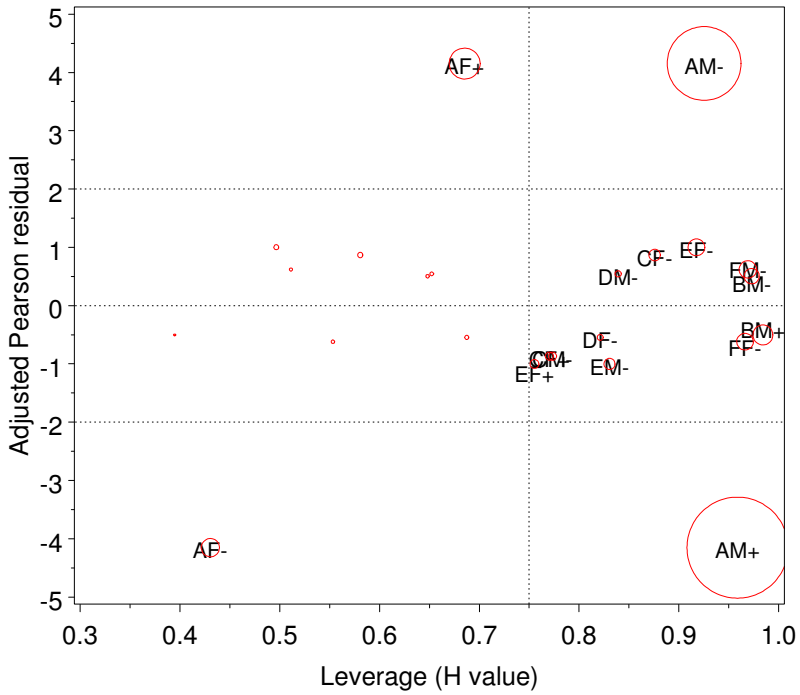


Figure 45: Influence plot for Berkeley admissions data, Model $[AD][GD]$. Bubble areas are proportional to Cook's D. Department A has a substantial impact on the fitted model

10.3.2 Model diagnostics with PROC CATMOD

These diagnostic quantities are not computed by PROC CATMOD; however, with some effort, they may be obtained from the results of PROC CATMOD. The technique described here is based on the idea that a log-linear model is essentially an ANOVA-type model for log frequency, which can be fit by weighted least squares regression. The steps are:

1. Fit the log-linear model, obtaining the cell frequencies n_i and estimated expected frequencies $\hat{m}_i$ in a dataset (see the statement RESPONSE / OUT=PREDICT above).

2. Use a regression program (PROC REG) which allows case weights and computes regression diagnostics. Fit the regression with:

   independent variables: Dummy variables for all marginals and association terms in the log-linear model. (PROC REG does not provide a CLASS statement.)

   dependent variable: The "working" response, $y_i = \log(\hat{m}_i) + (n_i - \hat{m}_i)/\hat{m}_i$

   weights: $\hat{m}_i$

3. Leverages will be correctly reported in the output. The standardized residuals, however, will have been divided by $\sqrt{MSE}$, and the Cook's D values will have been divided by MSE. For a model with k parameters and N cells the average leverage will be k/N, so a value $h_{ii} > 2k/N$ would be considered "large".

We continue with the results from model (AD, DG). Fitting the regression model for the working response using PROC REG is conceptually simple, though tedious because PROC REG (like PROC LOGISTIC) cannot generate the dummy variables itself. This must be done in a data step. In the PREDICT dataset, $\hat{m}_i$ is named _pred_, n_i is _obs_, and $e_i = n_i - \hat{m}_i$ is _resid_.

```
data regdat;
   set predict;
   drop _sample_ _type_ _number_ i;
   where (_type_='FREQ');
   cell = trim(put(dept,dept.)) ||
          gender ||
          trim(put(admit,yn.));

   *-- Working response;
   y = log(_pred_) + _resid_/_pred_;

   *-- Construct dummy variables for the regression;
   array d5  d1-d5;
   array ad5 ad1-ad5;
   array gd5 gd1-gd5;
   g = (gender='M') - (gender='F');
   a = (admit=1) - (admit=0);
   do i=1 to 5;
      d(i)=(dept=i) - (dept=6);
      ad(i) = d(i) * a;
      gd(i) = d(i) * g;
      end;
```

The weighted regression of the working response is then performed using the dummy variables as shown below. The output dataset is then processed to multiply the Cook's D and studentized residual by the appropriate factors of the MSE.

```
title2 'Diagnostics by weighted regression';
proc reg data=regdat outest=est;
   id cell;
   weight _pred_;
   model y = a g d1-d5 ad1-ad5 gd1-gd5;
   output out=regdiag
          h=hat cookd=cookd student=studres;

data regdiag;
   set regdiag;
   retain _rmse_;
   if _n_=1 then set est(keep=_rmse_ );
   adjres = studres * _rmse_;
   cookd = cookd * _rmse_**2;
```

The REGDIAG dataset is shown below. Note that all residuals are small in magnitude, except for those associated with Dept. A.

CELL	_OBS_	_PRED_	_RESID_	COOKD	HAT	ADJRES
AM+	512	531.431	-19.431	22.304	0.9588	-4.153
AM-	313	293.569	19.431	11.892	0.9254	4.153
AF+	89	69.569	19.431	2.087	0.6853	4.153
AF-	19	38.431	-19.431	0.724	0.4304	-4.153

```
BM+     353    354.188    -1.188     0.883    0.9842   -0.504
BM-     207    205.812     1.188     0.507    0.9729    0.504
BF+      17     15.812     1.188     0.026    0.6481    0.504
BF-       8      9.188    -1.188     0.009    0.3945   -0.504
CM+     120    113.998     6.002     0.058    0.5806    0.868
CM-     205    211.002    -6.002     0.142    0.7734   -0.868
CF+     202    208.002    -6.002     0.140    0.7701   -0.868
CF-     391    384.998     6.002     0.295    0.8758    0.868
DM+     138    141.633    -3.633     0.036    0.6873   -0.546
DM-     279    275.367     3.633     0.086    0.8391    0.546
DF+     131    127.367     3.633     0.031    0.6523    0.546
DF-     244    247.633    -3.633     0.076    0.8211   -0.546
EM+      53     48.077     4.923     0.054    0.4964    1.000
EM-     138    142.923    -4.923     0.272    0.8306   -1.000
EF+      94     98.923    -4.923     0.171    0.7552   -1.000
EF-     299    294.077     4.923     0.619    0.9176    1.000
FM+      22     24.031    -2.031     0.026    0.5531   -0.620
FM-     351    348.969     2.031     0.672    0.9692    0.620
FF+      24     21.969     2.031     0.022    0.5112    0.620
FF-     317    319.031    -2.031     0.612    0.9663   -0.620
```

An influence plot can now be constructed by plotting the adjusted residual (`ADJRES`) against leverage (`HAT`), using the value of Cook's D as the size of the bubble symbol at each point, giving the same plot as in Figure 45. Figure 46 shows a plot of the standard errors of residuals against expected frequency. Ordinary residuals do not adjust for the smaller standard errors associated with large expected frequencies.

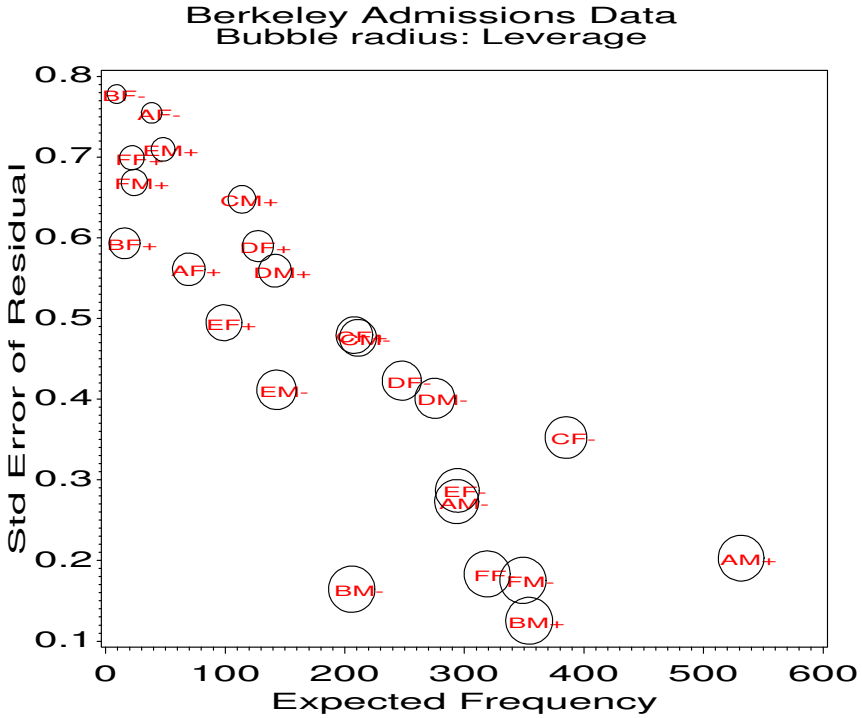


Figure 46: Standard errors of residuals decrease with expected frequency. This plot shows why ordinary Pearson and deviance residuals may be misleading.

11 Sequential analysis

Behavioral interactions between between mothers and children, husbands and wives, therapists and their clients, or baboons in the wild are often observed and recorded for the purpose of studying the dynamic processes of social interaction. The following illustrate the kinds of questions typically addressed:

- Does an infant's vocalization of distress lead to a nurturing response by the mother?
- Does positive or negative affect expressed by by a husband provoke a response of similar valence from the spouse?
- Is a probing question by the therapist more likely to lead to insightful self-reflection in the client?
- Is a threat gesture by a dominant baboon toward a baboon of lower dominance more likely to be met by a vocal response or an action response?

In such studies the behavior itself is typically recorded (video/audio tape) and transcribed, and each "event" by the participants is categorized according to a coding scheme developed by the researchers Bakeman and Gottman (1997). For example, in the dataset to be examined here Gottman and Roy (1990), husbands and their wives were observed discussing aspects of their marriage and audio-taped. Each talk-turn was classified by the speaker (H or W) and the affect (positive, neutral, or negative) expressed. Using the codes 1-6 representing the combinations of speaker and affect assigned as follows,

```
1='H+'  2='H0' 3='H-'
4='W+'  5='W0' 6='W-'
```

the first 100 of 392 speaker-events become data like this:

```
2 5 2 4 2  5 2 4 2 4  2 2 5 2 4  2 5 3 4 3  4 1 4 4 1
2 5 2 1 5  4 2 5 1 4  4 2 3 6 6  2 2 6 5 5  5 6 3 6 3
3 6 3 2 5  3 5 2 6 2  2 2 5 2 4  4 5 2 5 2  5 5 4 2 5
5 5 2 5 2  5 2 5 2 2  5 2 5 2 2  5 5 2 2 2  5 2 5 2 5
...
```

The development of a coding scheme for behavioral interaction sequences and the entry of data involves many choices: What are the events? Are codes mutually exclusive? Is duration information important? See Bakeman and Gottman (1997) for discussion of these issues. Bakeman and Quera (1995) describe a flexible system (SDIS/GSEQ, available on Hebb) for entering sequential data in a variety of formats, pooling events, calculating frequencies at various lags, etc. For further information, see the SDIS/GSEQ web site, `http://www.gsu.edu/~psyrab/sg.htm`. Here we consider some simple forms of analysis, with emphasis on visualizing patterns.

11.1 Analyzing time sequences

One simple form of analysis of such data considers the time course of the salient aspect of behavior as it changes over the duration of a session. One way to do this is to consider each event as having a value (e.g., +1, 0, -1), accumulating these over events, and graphing the result as a time series, which provides an analysis at the global or macro level.

The following statements read the numeric `code`s and use SAS programming statements to determine the speaker `who` and affect `value` for each event. The sequence number (`seq`) and cumulative `affect` are also calculated.

```
proc format;
  value code 1='H+'  2='H0' 3='H-'
             4='W+'  5='W0' 6='W-';
   value $who 'H'='Husband' 'W'='Wife';

data gottman;
   input code @@;
   who = substr(put(code, code.),1,1);
   value = 1 - mod(code-1,3);
```

```
   seq+1;
   affect+value;
   label affect = 'Cumulative Affect'
     seq = 'Sequential Event';
   lines;
2 5 2 4 2  5 2 4 2 4  2 2 5 2 4  2 5 3 4 3  4 1 4 4 1
2 5 2 1 5  4 2 5 1 4  4 2 3 6 6  2 2 6 5 5  5 6 3 6 3
3 6 3 2 5  3 5 2 6 2  2 2 5 2 4  4 5 2 5 2  5 5 4 2 5
5 5 2 5 2  5 2 5 2 2  5 2 5 2 2  5 5 2 2 2  5 2 5 2 5
...
```

The first 15 observations look like this:

SEQ	CODE	WHO	VALUE	AFFECT
1	2	H	0	0
2	5	W	0	0
3	2	H	0	0
4	4	W	1	1
5	2	H	0	1
6	5	W	0	1
7	2	H	0	1
8	4	W	1	2
9	2	H	0	2
10	4	W	1	3
11	2	H	0	3
12	2	H	0	3
13	5	W	0	3
14	2	H	0	3
15	4	W	1	4

The simplest useful display is just a plot of the affect `value` against sequence number (`seq`), as shown in Figure 47, either for both partners together, or with separate plotting symbols for each spouse. In such plots, the interpretation of overall trends is made easier by the addition of smoothed curves.

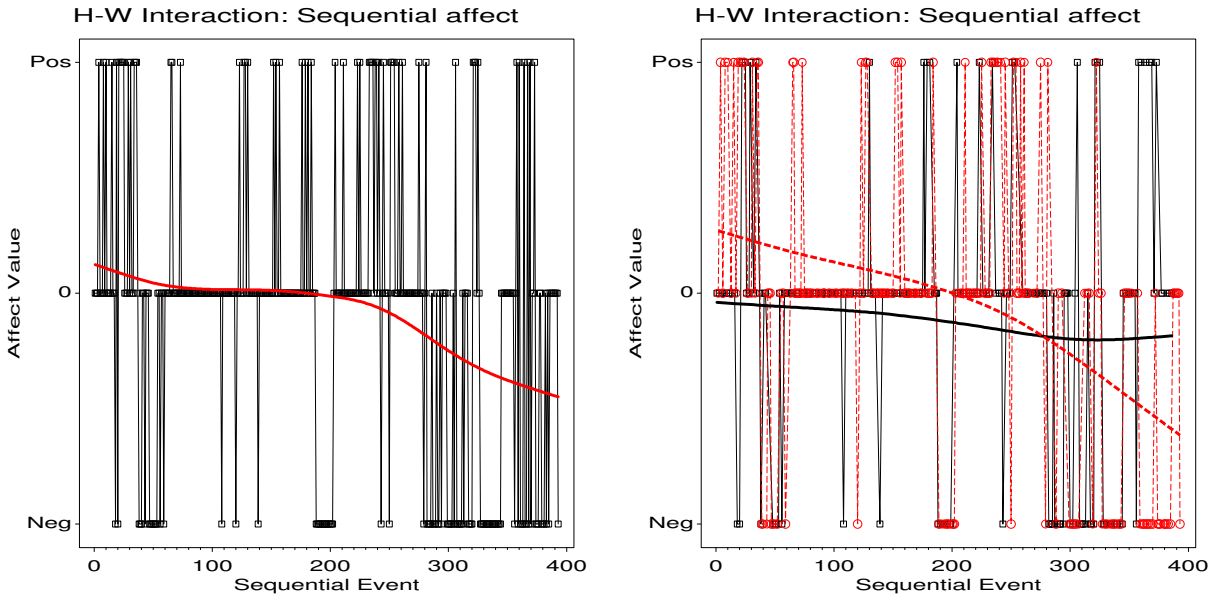


Figure 47: Time course of affect for one married couple. Left: Both partners; right: overlaid plots for each spouse (H: ——, W: – – –)

In the combined plot, we can see an initial period of largely positive or neutral affect interrupted by occasional negative exchanges, but ending on a largely negative series, as highlighted by the steady decline of the smoothed

curve. In the overlaid plot for each spouse, it appears that the husband's affect is, on average, slightly negative throughout, getting somewhat more negative as the session continues; the wife appears to start out with more positive affect, which declines steadily until the middle of the session, where it takes a shaper downward bend.

These simple plots are drawn by plotting the VALUE against sequence number *twice* on each graph – once showing the points joined by lines, and again showing only the smoothed curve. In the statements below (producing the left panel of Figure 47), the SYMBOL2 statement uses the option INTERPOL=SMnn to request a smoothed curve, where the number reflects a balance between closeness to the points ($nn = 0$) and smoothness ($nn = 99$). [It often takes several tries to find the right balance.]

```
proc gplot data=gottman;
   plot value * seq=1
        value * seq=2
      / overlay frame vaxis=axis1 vm=0  haxis=axis2 ;
   axis1 label=(a=90 'Affect Value')  offset=(3);
   axis2 offset=(2,1);
   symbol1 interpol=join v=square c=black l=1 h=1.5;
   symbol2 interpol=sm60 v=none  ci=red w=3;
```

A more satisfactory approach to smoothing, for discrete or continuous responses, is provided by a number of non-parametric regression methods, of which the LOWESS method (Cleveland, 1979, Cleveland and Devlin, 1988) is among the most popular. A plot similar to the left panel of Figure 47 is obtained with the LOWESS macro. The parameter f=0.5 specifies the smoothing parameter.

```
%lowess(data=gottman, x=seq, y=value, interp=join, f=0.5);
```

However, the discrete nature of the affect VALUE makes it somewhat difficult to see the details, and it is often more useful to plot the cumulative affect values, particularly when, as here, the individual values are coded $-1, 0, +1$. Then, increases in the plot reflect local changes toward more positive affect, while decreases reflect local changes toward negativity. Moreover, the cumulative affect at any point reflects the total valence of the session up to that time, postive or negative.

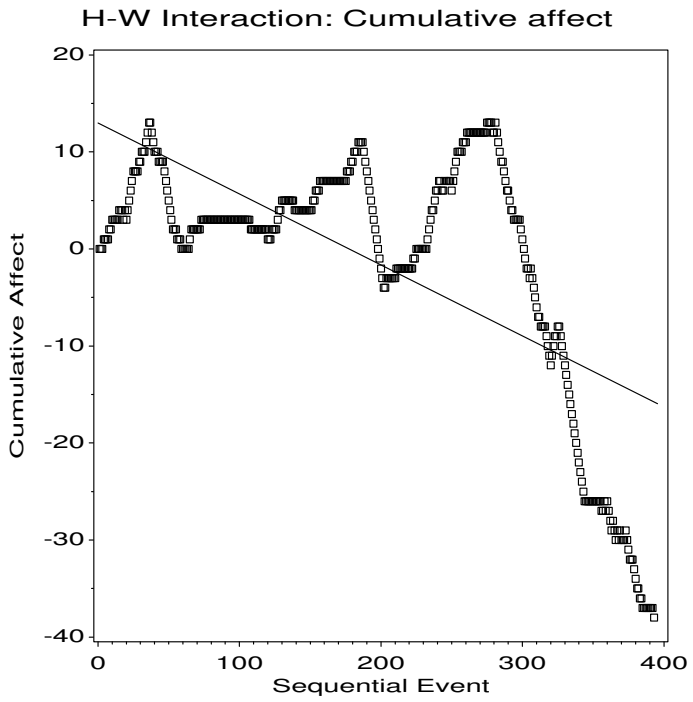


Figure 48: Cumulative time course of affect for one married couple

Thus, an overall plot of the time course of affect may be seen in a plot of the variable AFFECT against sequence number (SEQ), as shown in Figure 48. The graph rises and falls for the first two-thirds of the session, before beginning a precipitous decline toward increasingly negative affect. The line shows the linear regression of cumulative

affect on event number, which is useful only to a first, rough approximation (a smoothed curve might be preferable). This graph is produced by the following statements:

```
proc gplot data=gottman;
   plot affect * seq
      / frame vaxis=axis1 vm=1;
   axis1 label=(a=90);
   symbol1 interpol=rl v=square c=black l=1 h=1.4;
```

To study the changes over time in more detail, it is helpful to look at the separate trend in affect for the husband and wife. This can be done by calculating the running total affect separately for the husband and wife and plotting separate curves on one graph. The result, shown in Figure 49, indicates that overall the husband's affect is decreasing over time, while the wife's affect is relatively constant until the discordant decline at the end. More interestingly, the rise and fall in cumulative affect is closely matched between these spouses. It appears, however, that the wife begins the down turns slightly before the husband; she also begins sequences of positive affect somewhat earlier and with more consistency than the husband.

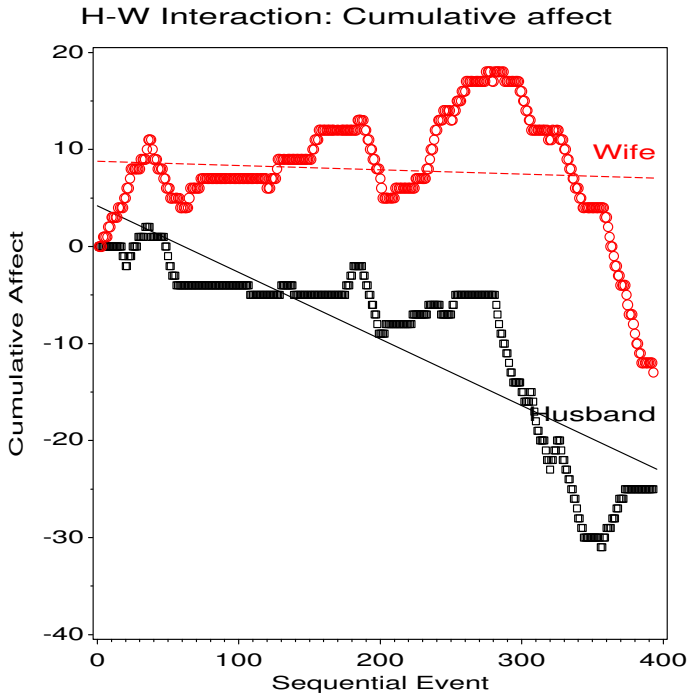


Figure 49: Separate affect trends for Husband and Wife. The lines show separate linear regressions for each spouse.

11.2 Lag-sequential analysis

A more fine-grained, micro-level analysis is provided by tabulating the pairs (or larger sets) of events which occur sequentially in time, or separated by varying numbers of intervening events. Lag-sequential analysis is explicitly designed to answer the kinds of questions posed at the outset of this section, e.g., "Does positive or negative affect expressed by by a husband provoke a response of similar valence from the spouse?" The key idea is to imagine a moving window sliding across the record of events, and tallying the occurrence of pairs (or larger sets) of events in time sequence.

In our running example, the observational record is a sequence of codes for discrete events, "event sequence data" (ESD) in the language of Bakeman and Quera (1995). In other observational contexts, the duration of individual states may be recorded (giving "state sequential data" (SSD)) and be of greater interest than their mere frequency. In the descriptions below, we can accommodate duration of events by weighting co-occurrences by their duration, so the tallies reflect number of time units, rather than number of events.

Similarly, other forms of sequential data allow for multiple, parallel streams of observations (infant, mother, grandmother), with different sets of codes for each (infant: attends actively, attends passively, does not attend; mother/grandmother: vocalizes, re-directs, maintains attention, etc.). In such cases it is often of greatest interest to restrict attention to sequential dependencies between the separate sets of codes.

To explain and illustrate the basic approach, we revert to the basic form of event sequential data.

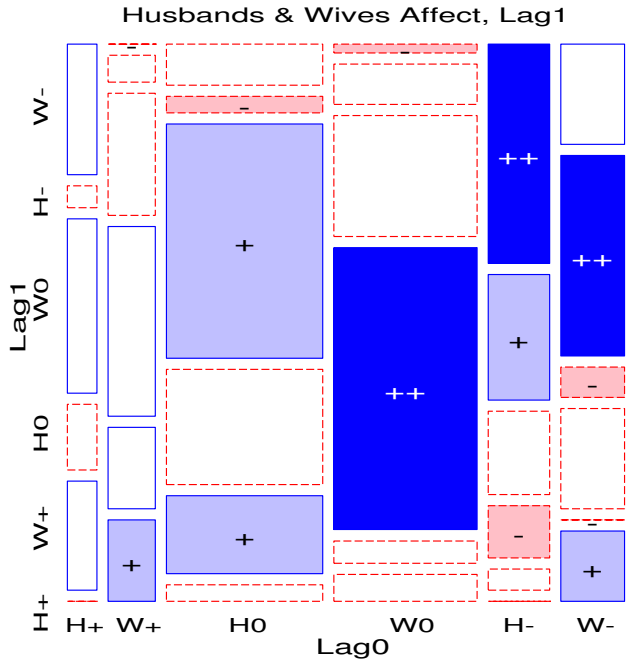


Figure 50: Lag-sequential mosaic for Husband and Wife

For example, in a "lag0 - lag1" table, we can tally sequential pairs of events in a 6×6 table. Conventionally, the first event (at lag 0) is often called the "given" event, and the second (at lag 1) is called the "target" event. For the Gottman/Roy data, the first five codes would give rise to the following four sequential pairs:

$$2\,5\,2\,4\,2 \quad \rightarrow \quad (2\,5), (5\,2), (2\,4), (4\,2)$$

For all 393 events, there are 392 sequential pairs, distributed as shown below. The "63", for (W0, H0) means than there were 63 pairs where the wife's neutral statement was immediately followed by a neutral affect statement of the husband.

LAG1	LAG0 (Given)						
(Target)	H+	H0	H-	W+	W0	W-	Total
H+	0	4	0	6	6	7	23
H0	3	28	5	14	63	10	123
H-	1	4	12	2	9	20	48
W+	5	19	2	6	5	0	37
W0	8	57	8	9	27	3	112
W-	6	10	21	0	2	10	49
Total	23	122	48	37	112	50	392

For any sequential data, lagged frequencies can be calculated not only for sequential pairs, but for events separated by one intervening event (lag0, lag2), two intervening events (lag0, lag3), etc. Or, a three-way or larger table can be constructed by considering sequential sets of three or more events (e.g., lag0, lag1, lag2). When duration is of interest, the joint occurrence can be weighted by the duration of the events.

A contingency table for any set of lagged events may be tallied from the original sequence by the LAGS macro. For example, the table above was computed by this statement:

```
%lags(data=gottman, outlag=lagdat, outfreq=freq,
      complete=all,var=code, varfmt=code., nlag=1);
```

The same statement, but using NLAG=2 gives a three-way frequency table FREQ of lag0 by lag1 by lag2 codes. In addition, the program produces an output dataset containing the original input variables plus the lagged event variables.

For such data, the mosaic display provides a visual depiction of the associations between the "given" events (lag0) and "target" events (lag1), or for relations among more than two events, whether weighted by duration or not.

The two-way mosaic for the lag0-lag1 data is shown in Figure 50. In this figure:

- The column widths show marginal frequencies of the given Lag0 event. These are greatest for neutral, least for positive events. Overall, the wife has a greater frequency of positive events than the husband
- The two-way mosaic reveals substantial association between husband and wife for the same affect value (the cells $(H+, W+)$. $(W+, H+)$; $(H0, W0)$, $(W0, H0)$; $(H-, W-)$, $(W-, H-)$ occur with greater than expected frequencies).
- The category codes were in Figure 50 were rearranged to group husband wife's codes by valence ($H+$, $W+$, $H0$, $W0$, $H-$, $W-$). This shows that there is greater tendency for negative affect of H and W to follow each other than other combinations. In this arrangement, one oddball cell also stands out as having greater than expected frequencies: $W- \rightarrow H+$!

11.2.1 Greater lags

Patterns of behavioural interaction often have more complex structure than one can see from just looking at adjacent items (lag0 and lag1). For example, a husband's negative affect can affect the pattern of interactions for several talk-turns. Hence, it is common to look at greater lags, and sometimes to posit and test models which relate them.

For larger lags we can calculate frequencies in a similar way, but the table becomes larger. For example, for lag0, lag1, lag2 in the current example, the table is of size $6 \times 6 \times 6$. Because the total number of observations is the same, the cell frequencies become smaller on average and we may expect to find more zero cells. As a result, unless the number of observations is quite large, it is often necessary to collapse the table in some way.

One way to do this is to collapse over intermediate lags. Thus, we could calculate a table of frequencies of lag0 by lag2 by summing the three-way (lag0, lag1, lag2) table over lag1 codes, as shown below. In the same way, one can calculate a two-way table of events separated by any number of intervening events, to study the association among non-adjacent event codes. Such analyses, over a number of intervening events, would indicate for how long the actions of one participant continued to influence the actions of the other.

```
%lags(data=gottman, outlag=lagdat, outfreq=freq,
      complete=all, var=code, varfmt=code., prefix=lag, nlag=2);

   /* collapse over lag1 */
proc summary data=freq nway;
   class lag2 lag0;
   var count;
   output out=lag02 sum=count;
```

This gives the two-way table shown below:

```
                          LAG0
LAG2     |  H+ |  H0 |  H- |  W+ |  W0 |  W- |  Total
---------+-----+-----+-----+-----+-----+-----+
H+       |   6 |   1 |   1 |   5 |   6 |   4 |  23
H0       |   6 |  62 |   8 |   7 |  32 |   8 | 123
H-       |   2 |   7 |  26 |   2 |   4 |   7 |  48
---------+-----+-----+-----+-----+-----+-----+
W+       |   2 |  13 |   3 |   8 |  11 |   0 |  37
W0       |   3 |  31 |   5 |  13 |  52 |   7 | 111
W-       |   4 |   8 |   5 |   2 |   6 |  24 |  49
---------+-----+-----+-----+-----+-----+-----+
Total       23   122    48    37   111    50   391
```

The mosaic display for this table is shown in Figure 51. We see again that the husband's and wife's statements continue to be associated by valence, perhaps more so than at lag1.

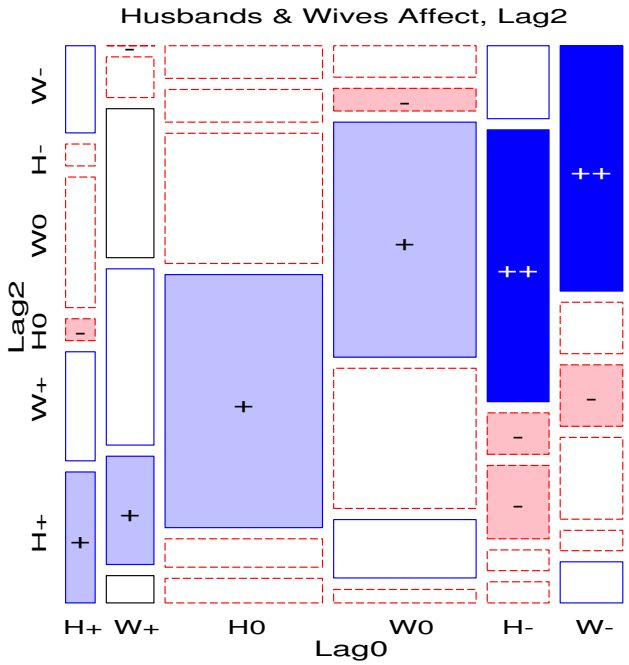


Figure 51: Lag0, Lag2 mosaic for Husband and Wife

Another way to look at larger lags, when the table is sparse, is to collapse some of the codes themselves in each table dimension. For example, we could look at just the valence of the affect by adding frequencies for the husband and wife to give a $3 \times 3 \times 3$ table of the events at lag0, lag1, and lag2.

This can be done by calculating the lagged frequencies based on the `value` of the code rather than the `code` itself.

```
proc format;
    value value -1 = 'Neg'  0='0' 1='Pos';

%lags(data=gottman, outlag=lagdat, outfreq=freq,
        complete=all, var=value, varfmt=value., prefix=lag, nlag=2);
```

Figure 52 shows the two-way mosaic between affect value at lag0 and lag 1. We see a strong positive association, with the largest positive residual for a negative response ("hit me, I'll hit you back"). We also see that a negative response is very unlikely to be preceded or followed by a neutral one.

In Figure 53 we have fit a one-step Markov chain model, [Lag0 Lag1] [Lag1 Lag2]. This model says that there may be associations between adjacent pairs of events, but none between events one-step removed. This model does not fit particularly well, G^2 (12) = 63.7, and the mosaic shows where the frequencies of events at Lag2 differ from expectation under this model. For example, the largest positive residual is for a sequence of three negative events, which accounts for over 25% of the overall lack of fit.

11.3 Extensions

11.3.1 Partial associations

Instead of collapsing over some lags merely to simplify the display, it is easier, and often more comprehensible, to stratify the analysis. That is, we fit separate models of (partial) independence, (with associated mosaics) for each level of one of the lag variables.

For loglinear models, there is a simple relation which connects an overall model of conditional independence with the models of simple independence within each of the layers of one (or more) conditioning vaariables.

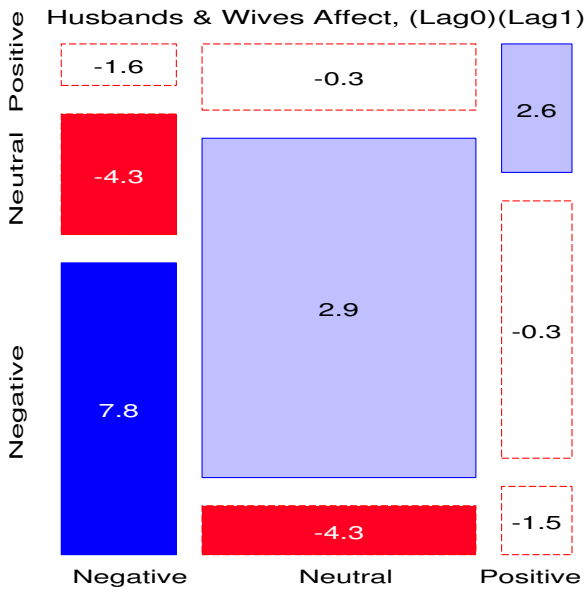


Figure 52: Lag0, Lag1 mosaic for Husband and Wife

Consider, for example, the model of conditional independence, $A \perp B \,|\, C$ for a three-way table. This model asserts that A and B are independent within *each* level of C. Denote the hypothesis that A and B are independent at level $C(k)$ by $A \perp B \,|\, C(k)$. Then one can show (Anderson, 1991) that

$$G^2_{A \perp B \,|\, C} = \sum_k^K G^2_{A \perp B \,|\, C(k)} \tag{40}$$

That is, the overall G^2 for the conditional independence model with $(I-1)(J-1)K$ degrees of freedom is the sum of the values for the ordinary association between A and B over the levels of C (each with $(I-1)(J-1)$ degrees of freedom). Thus,

- the overall G^2 may be decomposed into portions attributable to the AB association in the layers of C, and
- the collection of mosaic displays for the dependence of A and B for each of the levels of C provides a natural visualization of this decomposition. These provide an analog, for categorical data, of the conditioning plot, or ***coplot***, that Cleveland (1993) has shown to be an effective display for quantitative data.

See Friendly (1999, 2000b) for further details.

For example, the Markov model, [Lag0 Lag1] [Lag1 Lag2] is a model of conditional independence, Lag0 $\perp$ Lag2 | Lag1. By (40), we could partition this into separate model of independence of (Lag0, Lag2) for each level of the intervening Lag1 value. To conserve space we omit this analysis here, but note that the interpretation would be similar to that given below.

Alternatively, we could focus on the event at Lag0, and examine the partial associations (and mosaics) of the (Lag1, Lag2) events, depending on whether the Lag0 event was positive, neutral or negative The overall model is one of conditional independence, given the inital Lag0 event, [Lag0 Lag1] [Lag0 Lag2]. The partial mosaics for each level of the Lag0 event are shown in Figure 54.

Such mosaics are produced with the `mosaic` macro by including the parameter `BY=LAG0`, as shown below. In this application, we wish to show the formatted (text) values corresponding to value codes of -1, 0, +1, and to abbreviate the Lag2 values to avoid overlapping labels in the displays. The `table` macro provides this recoding, using the user-formats `VALUE.` and `VAL.`.

```
proc format;
  value value -1='Negative' 0='Neutral'  1='Positive';
  value val   -1='Neg' 0='Neut'  1='Pos';

%lags(data=gottman, outlag=lagdat, outfreq=freq,
```

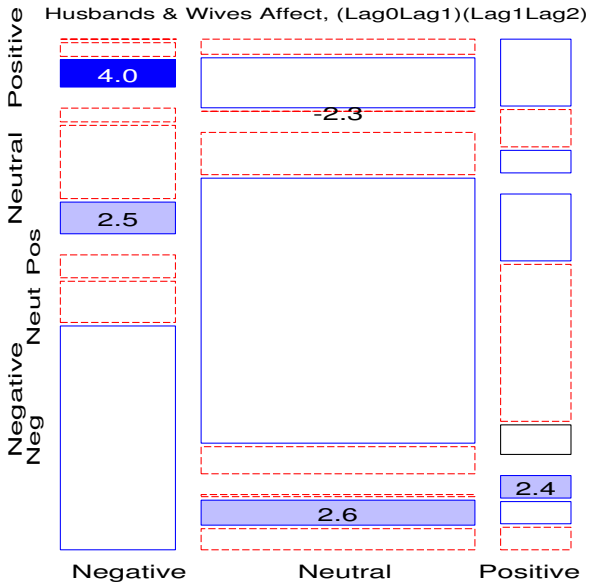


Figure 53: Lag0, Lag1, Lag2 mosaic for Husband and Wife

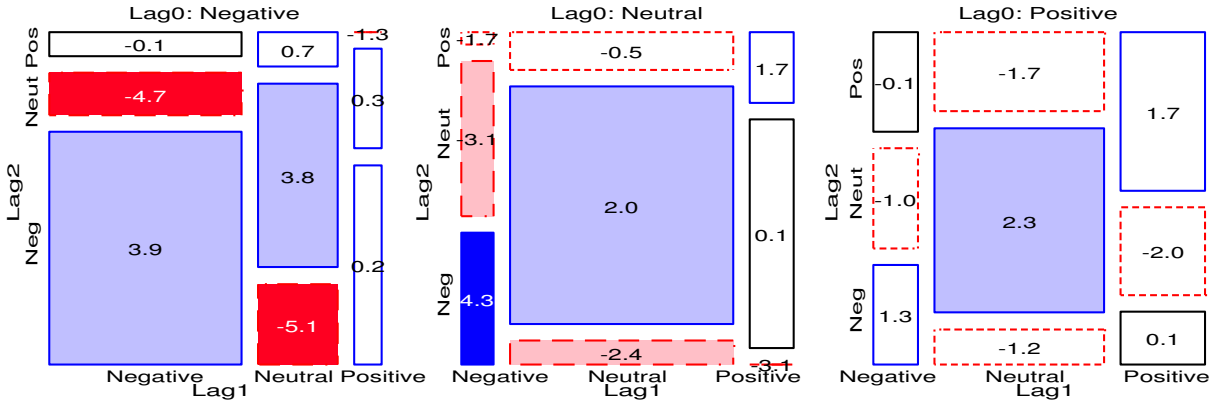


Figure 54: Lag1, Lag2 partial mosaics, by Lag0 value. Left: Lag0 negative; middle: Lag0 neutral, right: Lag0 positive

```
        complete=all,
        var=value, varfmt=value., prefix=lag,
        nlag=2);

*-- convert Lag values to formatted values;
%table(data=freq, var=Lag0 Lag1 Lag2, weight=count,
    format=Lag0 Lag1 value. Lag2    val.,
    out=freq2, char=Y);

*-- partial mosaics, by Lag0;
%gdispla(OFF);
%mosaic(data=freq2, vorder=Lag0 Lag1 Lag2, by=Lag0, count=count,
   devtype=lr adj, vlabels=2, cellfill=dev, htext=2.6);
%gdispla(ON);

%panels(rows=1, cols=3);
```

In Figure 54 we can see that there is a positive association between the Lag1 and Lag2 events for each level of

the initial event at Lag0. That is, whatever the initial event in a sequence of three, negative, neutral or positive following events (at Lag1) tended to be followed by events of the same valence (at Lag2), giving support to the notion of persistence of affect over time.

Table 2: Partial associations of Lag1, Lag2 events given Lag 0 event

Lag0	df	χ^2	p
Negative	4	23.701	0.0001
Neutral	4	29.844	0.0000
Positive	4	6.946	0.1388
Lag1 $\perp$ Lag2 | Lag0	12	60.491	

However, we can also see more subtle sequential dependencies, and examine the partial association G^2 statistics shown in Table 2. For example, when the initial event in the sequence is negative, the most frequent subsequent events are negative at both Lag1 and Lag2; the least likely subsequent (Lag1, Lag2) combinations are a neutral event at Lag1 followed by a negative event at Lag2. When the initial Lag0 event is neutral, however, most of the subsequent Lag1, Lag2 events are also neutral, but there is a significant excess of negative events at both Lag1 and Lag2, compared to the model of independence, Lag1 $\perp$ Lag2 | Lag0=neutral. Unfortunately for this couple, when the initial Lag0 event is positive, most subsequent Lag1, Lag2 events revert to neutral valence, and the association between Lag1 and Lag2 is weaker. The χ^2 statistics in Table 2 provide justification for interpreting the cell contributions to the overall χ^2 in the model of conditional association, Lag1 $\perp$ Lag2 | Lag0, when the Lag0 event is negative or neutral, but not when it is positive.

11.3.2 Mosaic matrices

As we noted earlier, cells in the contingency table become more sparsely populated as we increase the span of sequential events considered because the observations are distributed into a multiplicatively increasing number of cells. One way to reduce this "curse of dimensionality" is to consider only the bivariate relations between each pair of lags, in much the same way as when we examine the correlation matrix between all pairs of quantitative variables.

Just as the scatterplot matrix provides a visualization of the relations summarized in a correlation matrix, so too the mosaic matrix shows the relations between each pair of a set of categorical variables; see (Friendly, 1999) for a more complete description of the relations among various forms of bivariate plots for quantitative and categorical data.

Figure 55 gives the mosaic matrix among all pairs of the events at Lag0–Lag3. These results are quite simply interpreted: from the initial event at Lag0 up to events three-steps removed, all pairwise relations are remarkably similar: codes of a given affect value tend to follow the same code at all previous lags, with the greatest positive dependence between negative affects, though the greatest marginal freqencies are always for neutral affects.

Taken together, we believe that the graphical analyses of these data present the overview, and also capture the details of the event sequence in this record, in ways far more comprehensible than the equivalent collection of contingency tables and associated statistics could possibly do.

The example treated above is the simplest form of sequential data— a single couple, observed in one session, with only event information of interest. Even for event sequential data, we could have several sessions for each observational unit (couple), and we would usually have data on several such observational units.

The event-sequence plots presented earlier are specific to a particular couple-session. However, the lag sequential analysis and the varieties of mosaic plots could still be done, either pooling over sessions and/or couples (if the relations are believed to be homogeneous) or stratifying by these variables, as in the partial plots and mosaic matrices otherwise.

11.4 Multiple sessions, streams and durations

As noted earlier, other forms of sequential data allow for multiple, parallel streams of observations (e.g., infant, mother, grandmother), and often record the onset and offset of each behavior or action within its stream. The simplest data of this form is called *state sequential data*, and consists of one or more streams of codes for events or

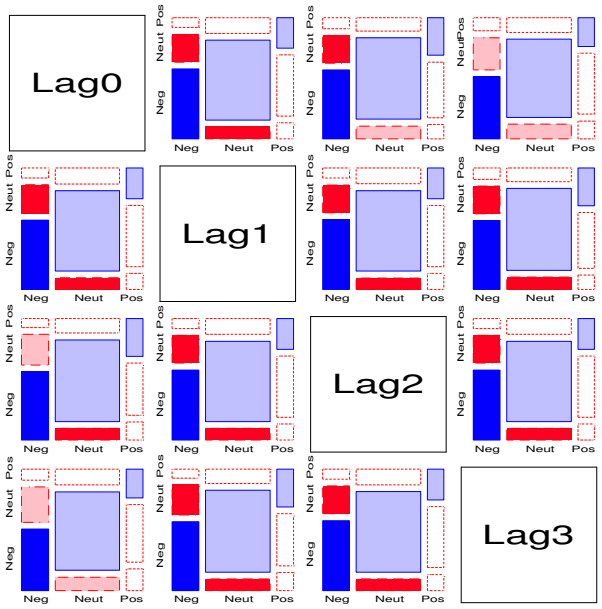


Figure 55: Mosaic matrix for events at Lag0–Lag3

states within an observational session, recorded with the time that each state began or ended. Typically, interest is focused on the duration of the various states within each stream, and perhaps the co-occurrence of states in different streams (Bakeman and Quera, 1995). A characteristic of state sequential data is that codes within a given stream are mutully exclusive and exhaustive, so that the beginning of one state implies the end of the previous state.

We illustrate the analysis of state sequences with data from the Ph.D. thesis of Sonya Rowland (Rowland, 1995, Bakeman, 2000). In this study, a sample of African American mothers were video recorded while interacting with their infants and the infant's maternal grandmother during free play. Included in the sample were 12 female infants and 24 male infants, each observed on two occasions: at age 6 and 12 months.

The initial coding included four streams: infant, mother, grandmother, and social-environmental events, with a variety of codes for each. Following (Bakeman, 2000), we simplify the data by ignoring the social-environmental stream and considering the following codes:

- Infant— **IA**: active attending to one or both adults (facial expressions, vocalizations, motor responses); **IP**: passive attending to one or both adults (looking); or **IN**: not attending to either adult.
- Mother and grandmother— **OR**: forcibly redirecting; **OM**: actively maintaining (verbal comments, facial expressions); **OA**: passive acknowledging (i.e., looking at the infant, but little else); and **ON**: not attending to the infant's current activity

For example, the listing below shows the codes and onset times for one female infant recorded at 6 months, in SDIS format. The mother and grandmother streams were represented initially by codes `Mx` and `Gx`. For ease of analysis, these were re-mapped into the codes `Ox` for "other" given above, with an additional variable, `who = M, G` distinguishing mother and grandmother.

```
<61406IA78>,0:30 IA,0:30 IP,0:42 IA,0:47 IN,0:52 OC,0:59 IA,1:01 IN,1:05
IP,1:08 IN,1:14 IP,1:31 IN,1:34 IP,1:45 IP,1:48 IP,1:52 IN,1:55 IP,2:02
IP,2:08 IN,2:12 IP,2:17 IA,2:23 IN,2:32 OC,3:24 IN,3:29 IP,3:33 IN,3:36
IP,3:57 IN,4:12 IP,4:42 IA,4:51 IP,5:02 IA,5:10 IN,5:24 IP,5:49 IA,5:52
IN,6:01 IP,6:04 IA,6:08 IP,6:13 IA,6:20 IN,6:25 &
%<61406IS78>%GN,0:30 GR,0:42 GM,0:45 GN,0:53 GA,0:59 GM,1:02 GA,1:05
GR,1:11 GA,1:14 GM,1:19 GA,1:22 GM,1:45 GN,1:48 GA,2:31 GR,2:34 GN,2:39
GR,2:51 GM,2:56 GR,3:02 GN,3:11 GR,3:32 GA,3:42 GM,3:51 GA,3:56 GN,4:12
GR,4:15 GA,4:31 GR,4:42 GA,4:49 GM,4:52 GA,5:00 GM,5:42 GA,5:48 GN,5:51
GA,6:04 GR,6:11 GA,6:20 &
%<61406IS78>%MM,0:30 MR,0:38 MN,0:42 MR,0:53 MN,0:59 MR,1:07 MA,1:15
```

```
MR,1:30 MM,1:38 MA,1:43 MR,1:48 MA,1:55 MR,2:00 MM,2:19 MA,2:34 MN,2:39
MR,2:46 MN,2:51 MR,2:58 MN,3:02 MR,3:11 MM,3:18 MR,3:22 MN,3:28 MR,3:32
MN,3:36 MR,3:56 MM,4:00 MN,4:12 MR,4:34 MA,4:41 MR,4:47 MA,4:53 MR,5:01
MA,5:21 MM,5:27 MA,5:32 MR,5:48 MM,5:53 MN,6:01 MR,6:04 MA,6:12 (female,6mo)/
```

Such data presents some additional challenges beyond the simple sequence of event codes found in event sequential data. However, the increased complexity is easily accommodated by creating a file of events or states, together with onset (`TIME`) and duration information. For the purpose of analyzing co-occurrence or sequences of events, sorting this file by `TIME` arranges the events in chronological sequence within each session. For example, the first 5 events for each stream are shown below in time order, using our revised coding scheme. The complete file for the Rowland data contains 5278 such records which had associated codes.

SUBJECT	AGE	SEX	WHO	STREAM	EVENT	CODE	TIME	DURATION
614	6	F	G	2	1	ON	0:00	.
614	6	F	M	3	1	OM	0:00	.
614	6	F	I	1	2	IA	0:30	0:00
614	6	F	I	1	3	IP	0:30	0:12
614	6	F	G	2	2	OR	0:30	0:12
614	6	F	M	3	2	OR	0:30	0:08
614	6	F	M	3	3	ON	0:38	0:04
614	6	F	I	1	4	IA	0:42	0:05
614	6	F	G	2	3	OM	0:42	0:03
614	6	F	M	3	4	OR	0:42	0:11
614	6	F	G	2	4	ON	0:45	0:08
614	6	F	I	1	5	IN	0:47	0:05
614	6	F	G	2	5	OA	0:53	0:06
614	6	F	M	3	5	ON	0:53	0:06
...								

Then, we may regard the entire dataset as a contingency table of states, classified by any one or more of the factors given above. With the data in this form, a key insight is that it is relatively easy to switch from analysis of the occurrence or co-occurrence of discrete events (ignoring duration) to analyses which take duration into account, simply by weighting each state code by the corresponding duration.

For example, we can calculate the contingency table of Age x Sex x Who x Code (collapsing over the infant Subject, and ignoring duration), or another contingency table of Age x Who x Code weighted by duration (collapsing over Subject and Sex).

In general, weighting by the duration of states makes any analysis or display reflect the number of time units which occur or co-occur, while ignoring duration counts only the individual states. Similarly, it is apparent that omitting any classification variable(s) in an analysis or plot is the same as pooling (summing) over the levels of those variable(s).

This analysis places all variables, and duration information on an equal footing. We can ignore (and pool over) Sex if we believe that the patterns of interactions between infants and adults are the same for male and female infants, *and* there are no important associations between Sex and the remaining variables. We can also ignore duration information *if* we believe that state-time is unimportant, *and* duration is not associated with the explanatory variables. This would not be the case, for example, if following a mother's or grandmother's re-directing action, infants tended to show longer periods of active attention; or if following an infant's active attending, the mothers or grandmothers spent more time in some states (OA or OR) than in others.

Nonetheless, the essential idea is that we can regard state sequential data, with multiple subjects (cross-classified by demographic variables like Age and Sex), sessions, and multiple streams as just one large contingency table, where the counts are discrete events, if we ignore duration, or are the number of time units, when we take time into account.

11.4.1 State-duration plots

As was the case with event sequential data, the simplest useful graphical representations for state sequential data portray the raw data, in a way designed to make it more comprehensible than the listings of the raw data shown above.

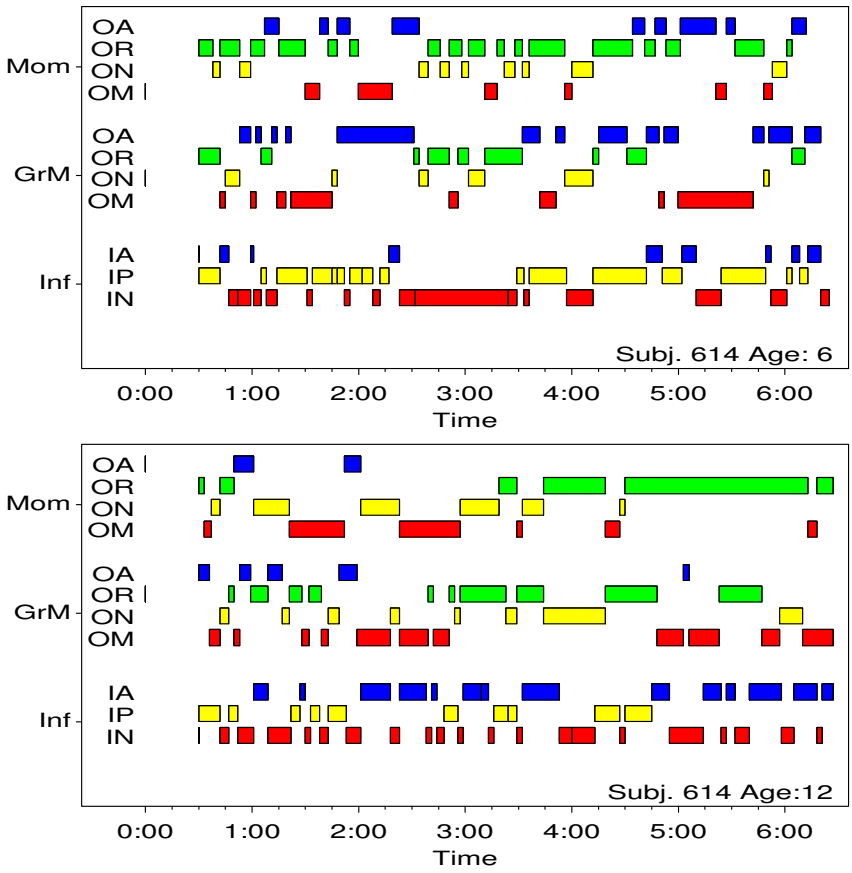


Figure 56: State-duration plots for one infant from Rowland

Figure 56 shows one example of such a display, called a "state-duration plot." Here, we display the three streams simultaneously over time for each session (6 and 12 months) of the infant whose data was listed above. Within each stream, the separate codes are mutually exclusive, so they cannot overlap. However, for ease of perception, they are offset vertically, and filled with different colors.

It is sometimes said that "sequential data are far too numerous to present in any direct way" (Bakeman, 2000), and researchers usually procced immediately to summary statistics and measures of association between different types of codes. However, graphs such as Figure 56 do present the raw state sequence data in a direct way, and yet, we believe, have the potential to highlight interesting patterns, suggest hypotheses, or reveal anomalies which might not be detected by other means.

For example, Figure 56 shows that this infant at 12 months spent had more and longer states of active attention than at 6 months, where passive attention and inattention were most dominant. As well, the mother was more likely to redirect the child's attention than the grandmother, at both sessions, and the durations of states tended to be longer for mothers and grandmothers at 12 months than at 6 months.

Of course, this is for just one infant, and any trends or associations observed should be tested across the entire sample. Nevertheless, plots such as these can play an important role in data screening and exploration in the same way that scatterplots, boxplots, and other graphic forms are used for quantitative data.

References

Anderson, E. B. *Statistical Analysis of Categorical Data*. Springer-Verlag, Berlin, 1991.

Andrews, D. F. and Herzberg, A. M. *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer-Verlag, New York, NY, 1985.

Bakeman, R. Behavioral observation and coding. In Reis, H. T. and Judd, C. K., editors, *Handbook of Research Methods in Social and Personality Psychology*. Cambridge University Press, New York, 2000.

Bakeman, R. and Gottman, J. M. *Observing Interaction*. Cambridge University Press, Cambridge, UK, 2nd edition, 1997.

Bakeman, R. and Quera, V. *Analyzing Interaction: Sequential Analysis with SDIS and GSEQ*. Cambridge University Press, New York, 1995.

Bangdiwala, K. Using SAS software graphical procedures for the observer agreement chart. *Proceedings of the SAS User's Group International Conference*, 12:1083–1088, 1987.

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

Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

Cleveland, W. S. *Visualizing Data*. Hobart Press, Summit, NJ, 1993.

Cleveland, W. S. and Devlin, S. J. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610, 1988.

Cohen, A. On the graphical display of the significant components in a two-way contingency table. *Communications in Statistics– Theory and Methods*, A9:1025–1041, 1980.

Fox, J. *Linear Statistical Models and Related Methods*. John Wiley and Sons, New York, 1984.

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

Fox, J. *Applied Regression Analysis, Linear Models, and Related Methods*. Sage Publications, Thousand Oaks, CA, 1997.

Friendly, M. *SAS System for Statistical Graphics*. SAS Institute, Cary, NC, 1st edition, 1991.

Friendly, M. A fourfold display for 2 by 2 by K tables. Technical Report 217, York University, Psychology Dept, 1994a.

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

Friendly, M. SAS/IML graphics for fourfold displays. *Observations*, 3(4):47–56, 1994c.

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. *Visualizing Categorical Data*. SAS Institute, Cary, NC, 2000a.

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

Gottman, J. M. and Roy, A. K. *Sequential Analysis: A Guide for Behavioral Research*. Cambridge University Press, Cambridge, UK, 1990.

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.

Hartigan, J. A. and Kleiner, B. A mosaic of television ratings. *The American Statistician*, 38:32–35, 1984.

Heuer, J. *Selbstmord Bei Kinder Und Jugendlichen*. Ernst Klett Verlag, Stuttgard, 1979. [Suicide by children and youth.].

Hoaglin, D. C. A poissonness plot. *The American Statistician*, 34:146–149, 1980.

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.

Hosmer, D. W. and Lemeshow, S. *Applied Logistic Regression*. John Wiley and Sons, New York, 1989.

Koch, G. and Edwards, S. Clinical efficiency trials with categorical data. In Peace, K. E., editor, *Biopharmaceutical Statistics for Drug Development*, pp. 403–451. Marcel Dekker, New York, 1988.

Koch, G. and Stokes, M. Workshop on categorical data analysis using the SAS system. *Proceedings of the SAS User's Group International Conference*, 16, 1991. ???–???

Mosteller, F. and Wallace, D. L. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*.

Springer-Verlag, New York, NY, 1984.

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

Peterson, B. and Harrell, F. Partial proportional odds models for ordinal response variables. *Applied Statistics*, 39: 205–217, 1990.

Riedwyl, H. and Schüpbach, M. Siebdiagramme: Graphische darstellung von kontingenztafeln. Technical Report 12, Institute for Mathematical Statistics, University of Bern, Bern, Switzerland., 1983.

Riedwyl, H. and Schüpbach, M. Parquet diagram to plot contingency tables. In Faulbaum, F., editor, *Softstat '93: Advances In Statistical Software*, pp. 293–299. Gustav Fischer, New York, 1994.

Rowland, S. B. *Effects of adolescent mother-infant and grandmother-infant interaction on infant attention: A normative study of an African American sample*. Unpublished doctoral dissertation, Georgia State University, Atlanta, 1995.

Shrout, P. E. and Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86: 420–428, 1979.

Snee, R. D. Graphical display of two-way contingency tables. *The American Statistician*, 28:9–12, 1974.

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

van der Heijden, P. G. M. and de Leeuw, J. Correspondence analysis used complementary to loglinear analysis. *Psychometrika*, 50:429–447, 1985.

Glossary

binary variable A binary **random variable** is a discrete random variable that has only two possible values, such as whether a subject dies (event) or lives (non-event). Such events are often described as success vs failure. *page 52*

covariate A covariate is a variable that may affect the relationship between two variables of interest, but may not be of intrinsic interest itself. As in **blocking** or **stratification**, a covariate is often used to control for variation that is not attributable to the variables under study. A covariate may be a discrete **factor**, like a block effect, or it may be a continuous variable, like the X variable in an analysis of covariance. Note that some people use the term *covariate* to include *all* the variables that may effect the response variable, including both the primary (predictor) variables, and the secondary variables we call covariates.

distribution function A distribution function (also known as the probability distribution function) of a continuous **random variable** X is a mathematical relation that gives for each number x, the probability that the value of X is less than or equal to x. For example, a distribution function of height gives, for each possible value of height, the probability that the height is less than or equal to that value. For discrete **random variables**, the distribution function is often given as the probability associated with each possible discrete value of the random variable; for instance, the distribution function for a fair coin is that the probability of heads is 0.5 and the probability of tails is 0.5.

expected cell frequencies For nominal (categorical) data in which the frequency of observations in each category has been tabulated, the *observed frequency* is the actual count, and the *expected frequency* is the count predicted by the theoretical **distribution** underlying the data, or under the null hypothesis. For example, if the hypothesis is that a certain plant has yellow flowers 3/4 of the time and white flowers 1/4 of the time, then for 100 plants, the expected frequencies will be 75 for yellow and 25 for white. The observed frequencies will be the actual counts for 100 plants (say, 73 and 27).

factors A factor is a single discrete classification scheme for data, such that each item classified belongs to exactly one class (***level***) for that classification scheme. For example, in a drug experiment involving rats, *sex* (with levels *male* and *female*) or *drug received* could be factors. A one-way analysis of variance involves a single factor classifying the subjects (e.g., *drug received*); multi-factor analysis of variance involves multiple factors classifying the subjects (e.g., *sex* and *drug received*).

goodness of fit Goodness-of-fit tests test the conformity of the observed data's empirical **distribution function** with a posited theoretical distribution function. The **chi-square goodness-of-fit test** does this by comparing observed and expected frequency counts. The *Kolmogorov-Smirnov test* does this by calculating the maximum vertical distance between the empirical and posited distribution functions. *page 42*

independent Two **random variables** are independent if their joint probability density is the product of their individual (marginal) probability densities. Less technically, if two random variables A and B are independent, then the probability of any given value of A is unchanged by knowledge of the value of B. A **sample** of mutually independent random variables is an independent sample.

measures of association For cross-tabulated data in a **contingency table**, a measure of association measures the *degree* or strength of association between the row and column classification variables. Measures of association include the *coefficient of contingency*, *Cramer's V*, *Kendall's tau-B*, *Kendall's tau-C*, *gamma*, and *Spearman's rho*. Strength of an association is not the same as statistical significance: a strong assocaition may not be significant in a small sample; conversely, a weak association may be significant in a large sample. *page 14*

population The population is the universe of all the objects from which a **sample** could be drawn for an experiment. If a representative random sample is chosen, the results of the experiment should be generalizable to the population from which the sample was drawn, but not necessarily to a larger population. For example, the results of medical studies on males may not be generalizable for females.

random sample A random sample of size N is a collection of N objects that are **independent** and identically **distributed**. In a random sample, each member of the **population** has an equal chance of becoming part of the sample.

random variable A random variable is a rule that assigns a value to each possible outcome of an experiment. For example, if an experiment involves measuring the height of people, then each person who could be a subject of the experiment has associated value, his or her height. A random variable may be *discrete* (the possible outcomes are finite, as in tossing a coin) or *continuous* (the values can take any possible value along a range, as in height measurements).

residuals A residual is the difference between the observed value of a response measurement and the value that is fitted under the hypothesized model. For example, in a two-sample unpaired t-test, the fitted value for a measurement is the mean of the sample from which it came, so the residual would be the observed value minus the sample mean. In a contingency table, the residual is the difference between the observed and expected frequency in each cell. *page 18*

stratification Stratification involves dividing a sample into homogeneous subsamples based on one or more characteristics of the population. For example, samples may be stratified by 10-year age groups, so that, for example, all subjects aged 20 to 29 are in the same age stratum in each group. Like **blocking** or the use of **covariates**, stratification is often used to control for variation that is not attributable to the variables under study. Stratification can be done on data that has already been collected, whereas blocking is usually done by matching subjects before the data are collected. Potential disadvantages to stratification are that the number of subjects in a given stratum may not be uniform across the groups being studied, and that there may be only a small number of subjects in a particular stratum for a particular group.

Index