December 1995

RECEIVED
JAN 16 1995
OSTI

Modeling Patterns in Count Data Using Loglinear and Related Models

Corwin L. Atwood

NOTICE

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for any third party's use, or the results of such use, of any information, apparatus, product or process disclosed in this report, or represents that its use by such third party would not infringe privately owned rights.

DISCLAIMER

**Portions of this document may be illegible
in electronic image products. Images are
produced from the best available original
document.**

Modeling Patterns in Count Data Using Loglinear and Related Models

Corwin L. Atwood

Published December 1995

Idaho National Engineering Laboratory
Idaho Falls, Idaho 83415

Prepared for the
U.S. Nuclear Regulatory Commission
Office for Analysis and Evaluation of Operational Data
Reliability and Risk Analysis Branch
Under DOE Idaho Operations Office
Contract DE-AC07-94ID13223

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ABSTRACT

This report explains the use of loglinear and logit models, for analyzing Poisson and binomial counts in the presence of explanatory variables. The explanatory variables may be unordered categorical variables or numerical variables, or both. The report shows how to construct models to fit data, and how to test whether a model is too simple or too complex. The appropriateness of the methods with small data sets is discussed. Several example analyses, using the SAS computer package, illustrate the methods.

SUMMARY

This report explains the use of loglinear and logit models for analyzing Poisson and binomial count data. It emphasizes methods that can be used with moderately sparse data sets.

The report motivates the analysis by giving several examples in which such analysis is appropriate. It then explains the loglinear model, when the response variable consists of Poisson counts, and the logit model, when the response consists of binomial counts. The relation between the two models is given. Either model has explanatory variables that may be unordered categorical variables or numerical variables, or both. Analysis with categorical explanatory variables is analogous to analysis of variance, and analysis with numerical explanatory variables is analogous to regression.

The report then considers construction of models to fit data. Two tests are given to compare nested models, the likelihood ratio test and a test of the Pearson chi-square type. When the alternative hypothesis puts no special restrictions on the model parameters, these tests reduce to the goodness-of-fit tests based on the likelihood ratio deviance and the usual Pearson chi-square statistic. Applicability of these tests with small data sets is discussed.

The examples given initially to motivate the report are analyzed. These analyses form a major part of the report body and illustrate the issues raised in the earlier sections. Final sections discuss several topics that were not fully treated earlier and summarize the report.

Appendix A gives programs, written with the SAS computer package, and outputs for all the loglinear and logit analyses in the body of the report. Appendix B explains the exponential family of distributions and the generalized linear model; these two topics are extremely useful in understanding loglinear and related models, but are put in the appendix because of their somewhat abstract nature. Appendix B then gives details and extensions of facts stated or used in the body of the report. Appendix C is an index to the entire report.

FOREWORD

The Office for Analysis and Evaluation of Operational Data (AEOD) of the U.S. Nuclear Regulatory Commission (NRC) is continually reviewing data from U.S. commercial nuclear power plants. Statistical analysis forms an important part of this work. Accordingly, the Reliability and Risk Analysis Branch of AEOD has asked the Idaho National Engineering Laboratory (INEL) to write a series of reports presenting the appropriate statistical tools for the kinds of data most commonly encountered. These reports are being written in parallel with a series of studies on the performance of safety systems in nuclear power plants, and they reflect the influence of those studies.

The reports are expected to provide tools and guidance to analysts of NRC data, although the statistical methods can, by their nature, be applied to data from many other fields. The reports are intended to quickly help a new user. The report on collecting operational data should be understandable by anyone with a technical background, although the examples are slanted towards an engineer with nuclear experience. The reports on statistical methods should be readable and immediately usable by a person with training in statistics but with no experience analyzing such data. Some of the reports, of necessity, are more advanced than others; for example, the report on loglinear modeling will be easier to understand if the reader has first assimilated the reports on binomial data, Poisson data, and linear models. Nevertheless, the reports are all intended to be introductory to the extent possible, suitable as brief self-study texts to help readers move quickly to the tasks of data collection and analysis. In addition, the reports should be usable as texts or references in short courses for persons with less training.

The first reports written or planned in this series are

Collecting Operational Event Data for Statistical Analysis, September 1994, EGG-RAAM-10086, by Corwin L. Atwood

Hits per Trial: Basic Analysis of Binomial Data, September 1994, EGG-RAAM-11041, by Corwin L. Atwood

Events in Time: Basic Analysis of Poisson Data, September 1994, EGG-RAAM-11088, by M. E. Engelhardt

Modeling Patterns in Continuous Data Using Linear and Related Models DRAFT, 1995, INEL-95/0120, by M. E. Engelhardt (Final version planned for 1996)

Modeling Patterns in Count Data Using Loglinear and Related Models, December 1995, INEL-95/0121, by Corwin L. Atwood

Statistical Analysis of Random Duration Times DRAFT, 1995, INEL-95/0206, by M. E. Engelhardt (Final version planned for 1996)

Practical Guidance for Statistical Analysis of Operational Event Data, October 1995, INEL-95/0234, by Corwin L. Atwood

ACKNOWLEDGMENTS

Mark DeHaan helped write the SAS programs. Comments by Larry Blackwood, Max Engelhardt, Cindy Gentillon, and Harry Martz improved both the content and the presentation.

CONTENTS

ABSTRACT	iii
SUMMARY	v
FOREWORD	vii
ACKNOWLEDGMENTS	ix
INTRODUCTION	1
Examples	1
Scope of This Report, and Related Reading	6
TERMINOLOGY AND NOTATION	8
MODELS	10
Loglinear Models	10
Poisson Response, Unordered Categorical Explanatory Variables	10
Poisson Responses with Unequal Exposure Times	12
Poisson Response, Numerical Explanatory Variables	13
Logit Models	13
Binomial Response, Unordered Categorical Explanatory Variables	13
Binomial Response, Numerical Explanatory Variables	14
Relationships between Distributions and between Models	15
CONSTRUCTING MODELS	16
Estimates of Model Parameters	16
Tests of Fitted Models	16
Dimension of a Model	16
Likelihood Ratio Test	19
Pearson Chi-Square Test	22
Relation between the Likelihood Ratio and Pearson Tests	22
Small Sample Sizes	23
EXAMPLE ANALYSES	25
RCIC/HPCI Example: Binomial Response, Categorical Explanatory Variables	25
Unplanned-Demands Example: Poisson Regression	29
MOV-Test Example: Logistic Regression	36
DISCUSSION	41
Interactions, and Simpson's Paradox	41
Nested and Crossed Effects	44
Treatment of Zero Counts	44
Random Effects	47
REVIEW	49
REFERENCES	52

Appendix A — SAS Programs and Full Outputs	A-1
A-1. PROGRAM OUTLINE FOR CALCULATING $X^2(H_0 H_1)$	A-3
A-2. RCIC/HPCI EXAMPLE	A-4
A-3. MACROS FOR SIMULTANEOUS CONFIDENCE BANDS	A-9
A-4. UNPLANNED-DEMANDS EXAMPLE	A-12
A-5. MOV-TEST EXAMPLE	A-16
Appendix B — Details and Extensions	B-1
B-1. EXPONENTIAL FAMILY	B-3
B-1.1 Definition	B-3
B-1.2 Examples	B-4
B-1.2.1 Normal(μ , σ^2) Distribution	B-4
B-1.2.2 Poisson(μ) Distribution	B-4
B-1.2.3 Binomial(n , π) Distribution	B-5
B-1.3 Properties	B-5
B-2. GENERALIZED LINEAR MODELS	B-7
B-2.1 Definition	B-7
B-2.2 Examples	B-7
B-2.2.1 Standard Linear Model	B-7
B-2.2.2 Loglinear Model	B-8
B-2.2.3 Logit Model	B-8
B-3. RELATION OF POISSON AND MULTINOMIAL DISTRIBUTIONS	B-10
B-3.1 Statement of Relation	B-10
B-3.2 Proof of the Relation	B-12
B-4. SOME PROPERTIES OF LOGLINEAR MODELS	B-13
B-4.1 Hierarchical Models and Marginal Constraints	B-13
B-4.2 Relation between Logit and Loglinear Models	B-14
B-5. DEVIANCE RESIDUALS	B-15
B-6. SMALL-SAMPLE BEHAVIOR OF LIKELIHOOD RATIO AND PEARSON TEST STATISTICS	B-17
B-7. POINTWISE AND SIMULTANEOUS CONFIDENCE STATEMENTS	B-19
B-7.1 Derivations	B-19
B-7.2 Discussion	B-21
B-8. REFERENCES	B-23

FIGURES

1. Examples of nested hypotheses.	17
2. Pointwise 90% confidence intervals and simultaneous 90% confidence band for Poisson event rate in the Unplanned-Demands Example.	34
3. Fitted Poisson event rate and simultaneous 90% confidence band, based on three ways of fitting the Unplanned-Demands data.	35
4. Fitted trend lines for the failure probability, and simultaneous confidence bands on the fitted lines at the two plants, for the MOV-Test Example.	39
5. Hypothetical plots of observed response.	41
6. Plot of observed responses, showing variation in number of demands.	42
7. Observed mortality versus prenatal care and clinic.	43

TABLES

1. Prenatal care versus infant mortality.	6
2. Portion of initial analysis of the RCIC/HPCI Example.	26
3. Analysis of the RCIC/HPCI Example, deleting plants with no failures.	28
4. Portion of analysis printout for the Unplanned-Demands Example.	31
5. Preliminary analysis of the MOV-Test Example.	38
6. Final analysis of the MOV-Test Example.	40
7. Four comparison analyses of a sample data set, with point estimates and 90% intervals.	46

EXAMPLES

The RCIC/HPCI Example. Failures on demand, for turbine-driven pump assemblies in two systems at many plants.	2
The Unplanned-Demands Example. Unplanned demands for the HPCI system per plant operating year, by calendar year.	3
The MOV-Test Example. Failures of motor-operated valves on demand, by calendar year, at two plants.	4
The SFL Example. Licensee Event Reports with safety function lost or available.	4

Modeling Patterns in Count Data Using Loglinear and Related Models

INTRODUCTION

This report presents ways to analyze counts of events, when other variables (*explanatory* variables) are present that can influence the event rate or the event probability. The mean of the count is assumed to depend on the known values of the explanatory variables and on estimable parameters. The nature of the dependence is described by the model, a loglinear model or a related model.

Because of the presence of explanatory variables, such a situation is more complicated than the simple settings considered for binomial data by Atwood (1994a) and for Poisson data by Engelhardt (1994). This report will be easier to read if you are already familiar with the basic analysis of binomial and Poisson data, for example as presented in those reports. Familiarity with the basic ideas of linear modeling, as presented in textbooks and by Engelhardt (1995), will also make this report easier to understand. The methods given here extend the methods given in the three earlier reports.

Some example data sets are now given to motivate this work. Following the examples, the remainder of this introduction describes the scope of the report and mentions books that go further.

Examples

Loglinear and related models are used in many ways, typically with count data such as binomial or Poisson data. Several examples are given in the next few pages. These examples will be used repeatedly throughout this report to illustrate such modeling and its applications. Therefore, they are given brief names, to make them easy to refer to.

The first example, the *RCIC/HPCI example*, involves only *categorical* data, that is, data that consist of counts of occurrences of categories. The categorical variables in this example are outcome (that is, success or failure), plant, and system. (Outcome is not shown directly in the example data as printed, but it can be found from the number of failures and the number of demands.) The possible values of a variable are sometimes called *levels*. In this example, outcome has two levels, failure and success, plant has 23 levels, and system has two levels, RCIC and HPCI. Of the three variables, outcome is naturally thought of as a *response* variable, and the other two as explanatory variables. The distribution of a response variable is assumed to depend on the levels of the explanatory variables. This is a convenient but artificial distinction, because it is possible to arrange the variables in a three-way table that shows the number of occurrences of each combination of variables, with no indication as to which

variables are explanatory and which are responses. Some computer programs use the explanatory/response distinction naturally, and others find the distinction most unnatural.

The RCIC/HPCI Example. Failures on demand, for turbine-driven pump assemblies in two systems at many plants.

Event summaries were examined for turbine-driven pump assemblies in the reactor core isolation cooling (RCIC) and high-pressure coolant injection (HPCI) systems in commercial nuclear power plants. For each assembly, the total number of failures on unplanned demands and periodic surveillance tests was tallied, and the corresponding number of demands was estimated. For the 23 plants with both a RCIC and a HPCI system, the data are given here. The data sources are partly proprietary and are not necessarily of the highest quality. Therefore, the plants are not identified by name here. The plant codes (A, B, ...) have no relation to plant codes used elsewhere. This data set comes from an internal NRC study.

Plant ID	System			
	RCIC		HPCI	
	Failures	Demands	Failures	Demands
A	1	24	2	24
B	0	72	0	72
C	0	24	0	25
D	1	25	0	75
E	3	26	3	26
F	1	24	0	24
G	0	73	2	73
H	0	72	7	72
I	2	74	2	72
J	0	24	0	72
K	0	74	1	74
L	1	77	4	79
M	0	74	2	79
N	2	75	1	75
O	1	24	6	74
P	0	72	2	73
Q	0	17	3	17
R	0	26	3	24
S	0	25	1	14
T	0	28	1	33
U	0	74	5	29
V	0	25	2	25
W	1	25	2	25

When analyzing the data of the RCIC/HPCI Example, we will try to estimate the probability of failure on demand. To do this, we will need to ask whether that probability depends on either or both of the explanatory variables. For example, if the probability is different for the two systems, two system-specific probabilities must be estimated. Investigating such questions will form the core of the analysis.

The second example, the *Unplanned-Demands Example*, differs from the RCIC/HPCI Example in two ways. First, the counts are Poisson distributed rather than binomially distributed. Associated with each Poisson count is a normalizing variable, *exposure time*, the time during which events could occur. In this example, it is natural to investigate whether the *rate* of events, that is, the expected number of events per plant operating year, depends on calendar year. This investigation treats the number of unplanned demands as a response, and the calendar year as an explanatory variable. The number of plant operating years is treated as a normalizing factor for the response, not as a possible explanatory variable. The reason for treating plant operating years in this way is that the explanatory relation between event count and plant operating years is presumed to be understood: the expected event count should be proportional to the operating time, at least when other things are held constant. Therefore, a model hypothesizing a relation between counts and operating time is of little interest. However, a model relating the count rate to calendar time is of definite interest.

The Unplanned-Demands Example. Unplanned demands for the HPCI system per plant operating year, by calendar year.

During the years 1987 through 1993, at U.S. commercial boiling water reactors, the HPCI system had 63 unplanned demands, as shown. An unplanned demand on the HPCI system can occur only when the plant is operating. The total number of plant operating years was 116.6, divided as shown. This data set comes from Grant et al. (1995).

	1987	1988	1989	1990	1991	1992	1993	Total
Unplanned demands	16	10	7	13	9	6	2	63
Plant operating years	14.98	14.27	15.89	18.16	17.76	17.63	17.91	116.60

The second difference between the Unplanned-Demands Example and first example is that the explanatory variable, calendar year, is numerical, with a natural chronological order. In fact, the data could be analyzed with methods that ignore this order and treat the different years as unordered categories, but it is of greater interest to treat year as a numerical value and search for a possible trend.

The *MOV-Test Example* has both an unordered explanatory variable, plant, and an ordered explanatory variable, year. Like the RCIC/HPCI Example, it has a binomial response variable. As in the RCIC/HPCI Example, an order can be assigned to the plants, but it is arbitrary, and if more plants were in the data set they would not have any order that is especially natural. And as with that earlier

example, the data will be analyzed to estimate the probability of failure on demand, when this probability may depend on the explanatory variables.

The MOV-Test Example. Failures of motor-operated valves on demand, by calendar year, at two plants.

In two sister plants, selected motor-operated valves (MOVs) are tested annually for leakage. Nine years of data for 52 valves at each plant are summarized by the failure counts here.

	Year of observation								
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>
Plant 1	4	2	3	1	4	3	4	9	6
Plant 2	1	1	1	1	2	3	1	4	3

This data set comes from Simola (1992), who comments that the leakage probability seems to be increasing, as could be expected because the sealing surfaces wear in use and no preventive maintenance is done.

The SFL Example. Licensee Event Reports with safety function lost or available.

195 Licensee Event Reports were submitted for HPCI system inoperabilities in 1987-92, as given by Atwood (1994a). Of these events, 123 involved loss of the safety function. The others were violations of various technical specifications or design requirements, but did not reflect an inability of the system to function when required. A few of these 195 events are listed here, with their dates converted to decimal numbers.

<u>Date</u>	<u>SFL</u>
1987.014	1
1987.071	1
...	
1992.773	0
1992.910	1
1992.997	0

The MOV-Test Example, with binomial counts, could be modified to use time as a continuous explanatory variable, if the dates of the individual tests were known. The final example, the *SFL Example*, is of this type. It has the values of 195 Bernoulli random variables and their dates. The Bernoulli random variables correspond to events with safety function lost (called hits, and coded as 1) and events with safety function available (called misses, and coded as 0); these codes are shown in the

safety function lost (SFL) column. The dates have been written as decimals; for example the first date was January 5, 1987, which is written as 1987.014 because 5/365 equals 0.014. Virtually every event occurs on a different date, so the data consist of 0's and 1's and their corresponding dates—the hits and misses are not accumulated in larger cells. This example is taken from Atwood (1994a), where it is analyzed as an ordered sequence of events, but without using the event dates. It will not be analyzed in this report, because time has absolutely no effect, and the analysis to show that fact is not interesting. The two reasons for mentioning the example are to show that time can be a continuous explanatory variable with binomial (here Bernoulli) counts, and that binomial events can represent dichotomies other than failure versus success on a demand.

An analogous example with Poisson data and time as a continuous explanatory variable is not within the scope of this report. Such a data set would consist simply of a list of event occurrence dates. This does not fit the framework of some explanatory variables and a response count. One could argue that a set of Poisson events is a limiting case of a set of Bernoulli events, with every very short time interval corresponding to a Bernoulli trial. This is mathematically true, but the number of Bernoulli trials must approach infinity for the asymptotic distribution to hold. The asymptotic results for loglinear and related models, as presented here, are not valid when the number of random variables increases indefinitely. Such data sets should be considered in the subject of density estimation, and may be treated in a report in this series dealing with that topic.

At some point in this report, the reader may ask whether the complex models to be given are really necessary. For example, is it necessary to consider the variables in the RCIC/HPCI or MOV-Test Examples simultaneously, or can the different explanatory variables be considered individually? Bishop, Feinberg, and Holland (1975, p. 41) and Everitt (1992, p. 64) give a widely used example showing that the simple analysis can give incorrect answers. The study was to relate infant mortality to prenatal care, based on data from two clinics. The analysis from the clinics individually and from the pooled data are given in Table 1.

Table 1. Prenatal care versus infant mortality.

	Clinic A		Clinic B		Combined	
	Care		Care		Care	
	Less	More	Less	More	Less	More
Died	3	4	17	2	20	6
Survived	176	293	197	23	373	316
X^2	0.084		0.000		5.255	
p -value	1.000		1.000		0.026	

Beneath each 2×2 table are two numbers: X^2 is the Pearson chi-square statistic for testing independence of the two factors, and the p -value is the attained significance level from Fisher's exact 2-tailed test. (These are explained in many statistics books and in Atwood 1994a.) The tables for the individual clinics show virtually no relation between prenatal care and infant mortality. The table with the combined data shows the opposite, a statistically significant relation. In fact, the combined data give a wrong conclusion. When other things (as summarized by the variable "Clinic") are held constant, prenatal care has no evident relation to infant mortality. The apparent conclusion from the combined data does not account for the differences between the two clinics, with their differences in clientele and perhaps also in equipment and procedures. This is an example of "Simpson's paradox," discussed in reference books and in a later section of this report.

Scope of This Report, and Related Reading

This report is a brief introduction to the topic of loglinear and related models, dealing only with selected aspects and applications. As mentioned above, the reader should already be familiar with basic analysis of binomial and Poisson data (as presented, for example, by Atwood 1994a and Engelhardt 1994). In addition, some knowledge of linear models (as summarized, for example, by Engelhardt 1995) is definitely helpful.

The body of this report guides the reader through some standard models and analyses. Some basic distributional facts are given first, followed by an introduction to the most common models. The models are illustrated using the above examples. Model construction is presented next, including tests of whether a model is complex enough (a goodness-of-fit test) and whether it is overly complex (that is, whether it contains unnecessary parameters). The most efficient way to get through that material is probably to skim it first, noting the topics covered, reading the parts that appear interesting, and saving the parts that appear too formidable for a later re-reading.

The first three examples given above are then analyzed, with the analyses illustrating the issues raised in the presentations of the methods. These example analyses form more than one third of the report body. They use the material presented earlier, and thus help encourage the reader to return to parts of the report that were skipped on the first pass.

A discussion section then elaborates on several topics that were either briefly mentioned or carefully avoided earlier. A final section summarizes the report and suggests guidance on use of the methods given. As a further aid to grasping the contents, an index to the entire report is given in Appendix C.

Computer programs, using the statistical computing package SAS*, are listed in Appendix A. Inquisitive readers can find more advanced material in Appendix B. That appendix contains sketched proofs of certain facts that are stated without proof in the body of the report and a brief outline of the unifying theory of generalized linear models. Some computer programs, such as the SAS procedure GENMOD used in the example analyses, use the language of generalized linear models introduced in Sections B-1 and B-2 of Appendix B.

For further reading, two relatively simple books on this topic are Everitt (1992) and Fienberg (1980). Two more advanced books are Agresti (1990) and McCullagh and Nelder (1989). Many of the fine points of categorical data analysis are presented, as clearly as possible, by Agresti. Bishop, Fienberg, and Holland (1975) was an impressive compilation on the subject when it first appeared, although it does not cover recent work in theory or computing. Christensen (1990) concentrates on large tables with large data sets.

* Mention of specific products and/or manufacturers in this document implies neither endorsement nor preference, nor disapproval by the U.S. Government or any of its agencies or contractors of the use of a specific product for any purpose.

TERMINOLOGY AND NOTATION

Explanatory variables can be *numerical* or *categorical*, depending on whether they describe quantitative or qualitative features. In the above examples, date and calendar year are numerical. Plant, on the other hand, is categorical. That is, a plant does not correspond naturally to a numerical value. It is true that numbers can be assigned as part of the plants' names, as in the MOV-Test Example, but the numbers are simply identifying symbols, not numerical quantities. The difference between two levels of a numerical variable, such as two dates or two years, is defined. Categorical variables, on the other hand, cannot be treated arithmetically; in particular, the difference between two categorical levels is not defined.

These distinctions between categorical and numerical explanatory variables are similar to those made with linear models—see standard statistics books. The cases with categorical explanatory variables (the RCIC/HPCI Example) correspond to analysis of variance, and the cases with numerical explanatory variables (the Unplanned-Demands and SFL Examples) correspond to regression analysis. The MOV-Test Example, with a categorical explanatory variable and a numerical explanatory variable, is sometimes called analysis of covariance in the literature of linear models.

Numerical variables are *ordered*; that is, the possible levels can be arranged in a unique natural order from smallest to largest. Categorical variables, on the other hand, may be *unordered*; for example plants, manufacturers, or vendors have no particular natural ordering. An intermediate case between numerical variables and unordered categorical variables is given by ordered categorical variables, such as a variable that takes values "very thin," "thin," "medium," and "thick." This report only gives methods for analyzing unordered categorical variables and numerical variables. For ordered categorical variables, either the order is ignored or numerical values are assigned to the levels.

In many situations, the observed data can be tabulated in *cells*, with a cell defined by the level(s) of one or more explanatory variables. This is the case for the RCIC/HPCI, Unplanned-Demands, and MOV-Test Examples, where several events typically occur for each combination of levels of the explanatory variables. It is not the case for the SFL Example, where each new event corresponds to a new level of the explanatory variable, date.

When cells can be defined with fewer cells than events, the following notation is useful. Let N_i be the number of events in the i th cell, where now i is a very generic notation standing for a subscript or set of subscripts. In some contexts, i is one-dimensional. For example, the Unplanned-Demands Example has N_i equal to the number of unplanned HPCI demands in the i th year, with i taking values from (say) 1 to 7. In other examples the index is two-dimensional, and we write ij instead of i . In the MOV-Test Example, we could ignore plant, and consider only outcome and year. The data could then be tabulated in a 2×9 table, with the first row representing failures and the second row representing

successes. Then N_{ij} is the number of outcomes of type i (failures when $i = 1$ and successes when $i = 2$) in year j . We could add to the complexity by letting k index plant. Then N_{ijk} is the number of outcomes of type i in year j at plant k , and these numbers would be the entries in a $2 \times 9 \times 2$ table. Clearly, the number of subscripts can be arbitrarily large in complex data sets.

Why not simply use linear models—analysis-of-variance or regression models—for the above examples? In a linear model the mean of the response is a function of the explanatory variables and the unknown parameters, and this function is linear in the unknown parameters:

$$E(Y_i) = \sum_j x_{ij} \beta_j \quad (1)$$

Here Y_i is the i th observation of the response variable, $E(Y_i)$ denotes its expected value, the x_{ij} 's are the explanatory variables, and the β_j 's are the unknown parameters. This model allows $E(Y_i)$ to take any value, positive or negative, large or small, because the β_j 's are unrestricted. If Y is a Poisson count, model (1) allows the expected count to be negative, which is impossible. If Y is a fraction (number of failures)/(number of trials), model (1) allows the expected fraction to be smaller than 0 or larger than 1, which is also impossible. Therefore, linear models do not obey the restrictions inherent in the situations. Loglinear and related models are used when the response variable can take values only in a restricted range. They are described next.

MODELS

Consider ways to model the relation between response variables and explanatory variables. This section gives the two most important types of models, the loglinear and logit models, with examples. Sections B-1 through B-4 of Appendix B go into more depth, showing the unifying theory behind these models and others. This can sometimes be useful for an analyst who must apply software expressed in terms of one model when the data are naturally modeled using the other model. The loglinear model is used most naturally with Poisson data, while the logit model can be used only with binomial data.

Loglinear Models

The *loglinear model* says that the logarithm of the expected value of the response is a function of the explanatory variables, and this function is linear in any unknown parameters. Here are several examples. In the terminology of generalized linear models (Section B-2 of Appendix B), the logarithm is the *link function*, which connects the expected value of the response to the function of the explanatory variables and unknown parameters. In this report, the term *logarithm* and the symbol $\log()$ always mean the natural logarithm, with base e .

Poisson Response, Unordered Categorical Explanatory Variables

Suppose the data set has two categorical explanatory variables, indexed by i and j , respectively. None of four examples given above has this form, but the explanatory variables might be year and plant, for example. Each explanatory variable is treated as unordered, even if it has a natural ordering, as year does. Let N_{ij} be the corresponding response, assumed to have a Poisson distribution with mean m_{ij} . This distribution is denoted $\text{Poisson}(m_{ij})$ in this report. This example does not decompose the parameter m_{ij} into a rate times an exposure time; that is deferred until the next example. One possible loglinear model says that

$$\log(m_{ij}) = \alpha + \beta_i + \gamma_j. \quad (2)$$

Equation (2) can be rewritten as

$$m_{ij} = ab_i g_j,$$

where $a = \exp(\alpha)$, $b_i = \exp(\beta_i)$, and $g_j = \exp(\gamma_j)$. Therefore, the additive model in $\log(m_{ij})$ is equivalent to a multiplicative model in m_{ij} . If such a model seems reasonable for the data, the statistical analysis may be based on it. Whichever way the model is written, it says that some years tend to have high counts (regardless of the plant), corresponding to large positive numbers b_i , and some years tend to have

low counts, corresponding to small positive numbers b_i . Note that small positive values of b_i correspond to large negative values of β_i —as b_i ranges from 0 to ∞ , β_i ranges from $-\infty$ to ∞ . Similarly, some plants tend to have high counts (large positive γ_j) and some have low counts. The loglinearity of the model means that the effect of year and the effect of plant are additive in the logarithm. Note that this satisfies the definition of a loglinear model: the logarithm of the expectation of the response is a function of year and plant, and is a linear function of the unknown parameters. In some contexts, fundamental consideration of the underlying process that produces the data may show that multiplicative effects are reasonable. Frequently, however, a loglinear model is just a simple model with no obvious inadequacies, and is used primarily for that reason.

As stated, the model is not unique: for example, the same m_{ij} values would be obtained if some constant were subtracted from α and added to every β_i . To make the parameters unique, some additional constraints are necessary. One approach uses the constraints $\Sigma \beta_i = 0$ and $\Sigma \gamma_j = 0$. Another approach (followed by SAS) arbitrarily picks one i , such as the last one, and sets that $\beta_i = 0$, and similarly sets one $\gamma_j = 0$. The discussion below follows the first approach, but either can be used.

The model suggests various ways to analyze data. One possible analysis would be to estimate the parameters. Another would be to test the hypothesis that some of the parameters are zero, for example, the hypothesis that $\beta_i = 0$ for all i . If β_i is identically zero, the response does not show a trend as year changes.

Equation (2) gives one possible model, containing two *main effects*, terms that each depend on one explanatory variable. If some of the parameters are treated as identically zero, the model simplifies. The model may also be made more complex, for example by adding an *interaction* term, an additive term that depends on both plant and year. If the model has many unordered categorical explanatory variables, the number of effects, including 2-way and higher-order interactions, can grow very large. Equation (2) could be generalized by using a different Greek letter for each modeled effect. However, to avoid running out of letters and to show clearly which interaction terms correspond to which explanatory variables, a notation with subscripts has become widely used. The generic version is

$$\log(m) = u + u_1 + u_2$$

and the more precise version rewrites the model just considered as

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)}, \quad \text{with } \Sigma u_{1(i)} = \Sigma u_{2(j)} = 0. \quad (3)$$

The effect of the first variable, year, is given by u_1 , and $u_{1(i)}$ is the effect of the i th year. Similarly, the effect of the second variable, plant, is given by u_2 .

An interaction term is denoted by u_{12} , and the value corresponding to the i th year and j th plant is $u_{12(ij)}$. The full model, with interactions, is

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}, \quad \text{with } \sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_i u_{12(ij)} = \sum_j u_{12(ij)} = 0.$$

Models with more unordered categorical explanatory variables are similar. The interaction terms can have a variety of forms, such as u_{12} , u_{23} , u_{14} , u_{234} , u_{1234} , etc. Such models are not emphasized in this report because complex models typically require large data sets, and such data sets are hardly ever available in the envisioned applications for the Nuclear Regulatory Commission (NRC).

Some of the facts stated in this report require the assumption that the model is *hierarchical*, that is, an interaction can be present in the model only if all the lower order terms are also present in the model. This assumption is discussed in Section B-4 of Appendix B.

Poisson Responses with Unequal Exposure Times

When the Poisson counts N_i have unequal exposure times t_i , as in the Unplanned-Demands Example, write the Poisson parameter m_i as $\lambda_i t_i$, and treat the λ_i 's as the fundamental parameters. Two ways are now possible to write the model. They are mathematically equivalent, but conceptually slightly different. Some computer programs are written to use one formulation and some to use the other.

One way is to treat N_i/t_i as the response, with expectation λ_i . The loglinear model then says

$$\log[E(\text{response}_i)] = \log[E(N_i/t_i)] = \log[E(N_i)/t_i] = \log[(\lambda_i t_i)/t_i] = \log(\lambda_i).$$

Here E denotes expected value. The last term, $\log(\lambda_i)$, can be modeled as a linear function of unknown parameters, for example

$$\log(\lambda_i) = u + u_{1(i)}.$$

When entering the data, one must enter N_i and t_i separately, not just the quotient N_i/t_i , because it is the numerator N_i that has the Poisson distribution.

The other model formulation is to treat N_i as the response. Then we have

$$\log[E(\text{response}_i)] = \log[E(N_i)] = \log(\lambda_i t_i) = \log(t_i) + \log(\lambda_i).$$

N_i has the same Poisson distribution as in the first formulation, and the term $\log(\lambda_i)$ can be modeled as a linear function of unknown parameters just as before. This is called a loglinear model with an *offset*,

$\log(t)$. The model to use, with or without the offset, depends on the analyst's taste and, more importantly, on the dictates of the available software. The two approaches yield completely equivalent results.

Poisson Response, Numerical Explanatory Variables

When the explanatory variables are numerical instead of unordered categories, *Poisson regression* is possible. (This report reserves the term "Poisson regression" for loglinear models with a Poisson response variable and numerical explanatory variables, although SAS uses the term for any loglinear model with a Poisson response.) For simplicity of presentation, suppose that the single explanatory variable corresponds to equally spaced times, such as successive years, indexed by i . The Unplanned-Demands Example is such a setting, because the difference between plants is ignored. A simple regression model would set

$$\log[E(\text{response}_i)] = \beta_0 + \beta_1 i .$$

This says that the logarithm of the expected response is linear in the unknown parameters, as required, and also linear in i . Taking exponentials on both sides shows that the expected response increases or decreases exponentially in i .

A more complicated model is of the form

$$\log[E(\text{response}_i)] = \beta_0 + \beta_1 i + \beta_2 i^2 .$$

This is linear in the parameters, as required, and is quadratic in i . These examples can be generalized according to the user's needs. Any function of i may be formed, not necessarily a polynomial function. As long as it is linear in the unknown parameters, the resulting model is loglinear.

Logit Models

Binomial Response, Unordered Categorical Explanatory Variables

The examples until now have all used Poisson counts. Now consider binomial counts, as in the RCIC/HPCI and MOV-Test Examples. Let N_{1j} and N_{2j} be, respectively, the number of failures and successes corresponding to the j th level of the explanatory variables. The index that is denoted by j here may be one-dimensional, corresponding to a single variable, or multidimensional, corresponding to a combination of explanatory variables. Let n_{+j} be the total number of successes and failures corresponding

to j ; n_{+j} is treated as a fixed number, not a random number. Either n_{+j} is truly fixed, or the results are conditional on the observed value. Assume that N_{ij} is binomial(n_{+j} , π_j), and that the N_{ij} 's are independent for different j ; this is a special case of the assumed independent multinomial distributions discussed in Section B-3.1 of Appendix B.

Let the response be N_{ij}/n_{+j} , which has expectation π_j . The function $\log[\pi/(1 - \pi)]$ is called the *logit* function of π , and $\pi/(1 - \pi)$ is sometimes called the *odds ratio*. The *logit model* says that $\log[\pi_j/(1 - \pi_j)]$ is a linear function of unknown parameters. This is very similar to the formulation of loglinear models, but now the response is a ratio of counts, and the logit is the link function. For example, suppose that two explanatory variables are present, system and plant, as in the RCIC/HPCI Example. Consider outcome as the first variable, system as the second, and plant as the third. The model

$$\logit[E(\text{response}_{jk})] = \log[\pi_{jk}/(1 - \pi_{jk})] = u + u_{2(j)} + u_{3(k)} \quad (4)$$

is the logit model with two main effects and no interaction. Other logit models could be constructed by dropping main effects or adding an interaction term.

Some books (such as Agresti 1990) express the logit model in terms of Bernoulli variables, the individual terms that are added to yield the binomial variable. If this is done, the response is simply the Bernoulli count, 0 or 1, and the divisor n_{+j} is unnecessary because it is identically 1.

Binomial Response, Numerical Explanatory Variables

A numerical explanatory variable can be discrete, such as calendar year, or continuous, such as calendar time. In either case, the levels of the variable can be plotted on a numerical scale. In the case when j indexes year, many demands could have occurred in the year, and the response is (number of failures)/(number of demands). The MOV-Test Example illustrates this, if only the data set from one plant is used. Then a logit model could be

$$\log[\pi_j/(1 - \pi_j)] = \beta_0 + \beta_j j. \quad (5a)$$

If, instead, t is the time of a demand (presumably a distinct time for each demand), the response could be the number of failures (either 0 or 1) on the demand. This is illustrated by the SFL Example. The logit model could be

$$\log[\pi_t/(1 - \pi_t)] = \beta_0 + \beta_1 t. \quad (5b)$$

More elaborate models can also be constructed by using powers of j or t , or other functions, just as in regression analysis. Such models, of the form (5a) or (5b) or their extensions, are called *logistic*

regression models. The terminology is standard, although its basis is obscure because the *inverse* of the logit function is a logistic function.

Relationships between Distributions and between Models

The Poisson and multinomial distributions are related as follows. Let N_i be $\text{Poisson}(m_i)$, and let the N_i 's be independent for i from 1 to k . Define $N_+ = \sum_i N_i$ and $m_+ = \sum_i m_i$. Let the observed value of N_+ be n_+ . Then, conditional on n_+ , the random variables $(N_1, \dots, N_k)$ have a multinomial(n_+ , $\pi_1, \dots, \pi_k$) distribution, where $\pi_i = m_i/m_+$. In particular, the conditional distribution of any N_i is binomial(n_+ , π_i), and the N_i 's satisfy the constraint $\sum_i N_i = n_+$. This relation is discussed and proved in Section B-3 of Appendix B. The relation is occasionally useful, because some software is written with only one distributional option, Poisson or multinomial. Knowing how to relate the distributions can help the analyst overcome apparent software limitations.

In principle, the logit model can be written as a loglinear model, as explained in Section B-4.2 of Appendix B. However, both models are useful components of the analyst's tool kit, because some data sets are analyzed much more naturally in terms of Poisson rates and loglinear models, and others in terms of binomial probabilities and logit models.

CONSTRUCTING MODELS

The previous section described two kinds of models, loglinear models for Poisson data and logit models for binomial data. To actually construct a useful model for data analysis, however, the parameters must be estimated, and the model must be tested to see if all the parameters are necessary (the concern about overfit) and if the model fits the data adequately (the concern about goodness of fit). For instance, in the RCIC/HPCI Example, we could postulate a model in which the probability of failure on demand depended on the plant, but was the same for the two systems. We could also postulate models in which the probability depended on system but not on plant, or depended on both, or on neither. We would want to decide which model was best, and to estimate the failure probability under the assumptions of that model. The example analyses to be given will work through this process for the RCIC/HPCI, Unplanned-Demands, and MOV-Test Examples.

In practice, the analyst can let the software provide the estimates. More understanding is required, however, when testing to see if the model is too simple or too complicated. Therefore, testing receives much more attention than estimation in the discussion below.

Estimates of Model Parameters

It is standard to use maximum likelihood estimators (MLEs) of the model parameters. The estimates are the same, whether the data counts are assumed to follow independent Poisson distributions or the corresponding multinomial distributions, as discussed in Section B-4.1. The MLEs generally do not have explicit algebraic forms, but must instead be found by numerical iteration. The numerical algorithms are of great interest to some theoreticians and programmers, but are not considered here. Bishop, Fienberg, and Holland (1975), Fienberg (1980), Agresti (1990), and Christensen (1990) discuss algorithms for finding the MLEs. The estimators will be denoted here by hats: $\hat{m}_{ij}$ is the MLE of m_{ij} , and so forth.

Tests of Fitted Models

The two tests considered here are the Pearson chi-square test and the likelihood ratio test. Both have the same asymptotic χ^2 distribution, with degrees of freedom equal to the difference in dimensions of the models specified by the null hypothesis and the alternate hypothesis.

Dimension of a Model

The random variables N_i have a joint distribution such as independent Poisson or independent binomial. Parameters are used to specify these distributions, and the values of some or all of these

parameters are typically unknown. In the discussion below, the term *model* will mean the set of possible values of the parameters. The distributional forms, such as Poisson or binomial, are actually part of the model, but they will be understood and not mentioned explicitly. Similarly, the usual restrictions—expected counts are nonnegative and probabilities sum to 1.0—will be assumed implicitly; if mentioned in the discussion below, they will simply be called the *necessary* restrictions. When more than one possible model is considered, each model may be hypothesized, that is, tentatively assumed to be true. We will use the terms *hypothesis* and model interchangeably, to specify the possible values of the parameters, with distributions and constraints being understood.

Consider examples that, like the Unplanned-Demands Example, have Poisson responses. If N_1 through N_3 are Poisson, the three parameters are the means m_1 through m_3 . (The Unplanned-Demands Example had seven Poisson counts. We consider a smaller example, with three counts, only to make the drawings easier.) The set of possible parameter values is shown in Figure 1a, a three-dimensional region. This space is defined by the hypothesis of no restrictions (other than the necessary ones) on the parameters. Denote that hypothesis by H_1 . The hypothesis that all the m_i 's are equal is shown as a line, and denoted H_0 . This line is a one-dimensional space, because one parameter completely determines the

a. Poisson data

b. Multinomial data

Figure 1. Examples of nested hypotheses.

possible means: if m_1 , say, is known, the other means are equal to it. H_0 can also be thought of as two constraints ($m_1 = m_2$ and $m_2 = m_3$) restricting the three parameters, with the dimension of H_0 equal to the dimension of H_1 minus the number of constraints, $3 - 2 = 1$.

If, instead, we had $m_i = \lambda_i t_i$, with unequal t_i 's, the hypothesis of equal λ_i 's would result in H_0 : $m_1/t_1 = m_2/t_2 = m_3/t_3$. The geometrical representation of this H_0 would also be a straight line, but with a different orientation.

When the above random variables are conditioned on $\Sigma N_i = n_+$, their joint distribution becomes multinomial(n_+ , π_1 , ...), with $\pi_i = m_i/n_+$. This fact was mentioned at the end of the above section on models. The three π_i 's satisfy the necessary constraint $\Sigma \pi_i = 1$, so that any two of the π_i s completely determine the distribution. This two-dimensional triangular region is shaded in Figure 1b, and labeled H_1 . It is the hypothesis of only the necessary restrictions on the π_i s. The hypothesis that the π_i s are all equal is a zero-dimensional region, the point labeled H_0 .

Note that in each case in Figure 1, the difference in dimensions between H_1 and H_0 is 2. In Figure 1a, the difference is $\text{dimension}(H_1) - \text{dimension}(H_0) = 3 - 1 = 2$. In Figure 1b, the difference is $\text{dimension}(H_1) - \text{dimension}(H_0) = 2 - 0 = 2$.

The examples above are extraordinarily simple. More realistic hypotheses specify $\log(m_i)$ or $\text{logit}(\pi_i)$ as a linear function of some parameters. These are harder to visualize, but the same idea applies: The dimension of the model is the number of parameters that completely specify the data distribution.

Hypotheses can be nested:

$$H_0 \subset H_1 \subset H_2 \subset \dots$$

when H_0 is a subspace of a higher dimensional H_1 , which in turn is a subspace of a higher dimensional H_2 , etc. Geometrically, H_{i-1} is a subspace of H_i ; in terms of equations, H_{i-1} is a special case, a restriction, on H_i , and H_i is more general than H_{i-1} . This terminology of *nested models*, or *nested hypotheses*, is standard (for example, Scheffé 1959 or Agresti 1990), but should not be confused with other terms such as "nested factors." To illustrate this with the RCIC/HPCI Example, we could have

H_0 is the model with no differences between system or plant: $\text{logit}(\pi_{ij}) = u$

H_1 is the model with a system effect: $\text{logit}(\pi_{ij}) = u + u_{1(i)}$

H_2 is the model with a system effect and a plant effect: $\text{logit}(\pi_{ij}) = u + u_{1(i)} + u_{2(j)}$

H_3 is the model with both main effects and an interaction: $\text{logit}(\pi_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$

Each model is built from the earlier model by adding new terms.

Sets of hypotheses are not always related in this way. For example, neither the model

$$\log(m_{ij}) = u + u_{1(i)}$$

nor

$$\log(m_{ij}) = u + u_{2(j)}$$

is a special case of the other, although both are special cases of the model

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)}$$

and of more general models.

At the end of any chain of nested hypotheses is the hypothesis that puts no restrictions on the parameters other than the necessary ones. This hypothesis at the end of the chain has no standard notation, but will be denoted H_∞ in this report:

$$H_0 \subset H_1 \subset H_2 \subset \dots \subset H_\infty$$

If the random variables are independent Poisson counts, the dimension of H_∞ is the number of random variables. If, instead, the random variables are multinomial counts, the dimension of H_∞ is the number of random variables minus 1, because the set of all the π_i 's but one completely specifies the data distribution.

In the above example of nested hypotheses for the RCIC/HPCI Example, H_4 states that each π_{ij} is unrestrained by any of the others, because the interaction term $u_{12(ij)}$ allows every $\text{logit}(\pi_{ij})$ to take a value independent of the others. Therefore, a more complicated model cannot be postulated (as long as we assume that the counts are binomial and independent). Therefore, H_∞ equals H_4 .

Likelihood Ratio Test

Test of a Nested Model against a More General Model. The likelihood ratio test is useful when the null hypothesis, H_0 , is a lower dimensional restriction of H_1 , as in Figure 1. The alternative hypothesis is all of H_1 except for H_0 , denoted $H_1 - H_0$. For example, in Figure 1a, H_1 is the three-dimensional region, and the alternative hypothesis is all of the three-dimensional region except for the line H_0 . Denote the (vector of) unknown parameters by θ , a multidimensional quantity. The *likelihood*

is the probability of the data, regarded as a function of θ ; denote the likelihood by $L(\theta)$. The likelihood ratio statistic is defined as

$$\Lambda = \max_{\theta \in H_0} L(\theta) / \max_{\theta \in H_1} L(\theta) .$$

Note that Λ is always less than 1, because H_0 is a restriction of H_1 , so the likelihood cannot be made as large when θ is restricted by H_0 as when restricted only by H_1 . If Λ is large, then H_0 is plausible, in the sense that it is consistent with the data, and should not be rejected. On the other hand, if Λ is small, then H_0 should be rejected in favor of H_1 .

Let us illustrate this with the above example of hypotheses for the RCIC/HPCI Example. To make the notation consistent with the notation just given for Λ , we will rename two of the hypotheses:

H_0 is the model with a system effect: $\text{logit}(\pi_{ij}) = u + u_{1(i)}$

H_1 is the model with a system effect and a plant effect: $\text{logit}(\pi_{ij}) = u + u_{1(i)} + u_{2(j)}$

Consider testing the hypothesis that we now call H_0 against the more general hypothesis H_1 . The two hypotheses have the same number of parameters u and u_1 . However, H_1 also has the parameters $u_{2(j)}$, where j indexes the plants; under H_0 , each $u_{2(j)}$ equals 0. The likelihood can be made larger if the $u_{2(j)}$ parameters are free to vary. Therefore, Λ must be between 0 and 1. We note also that the difference in dimensions between H_1 and H_0 is the number of independent parameters $u_{2(j)}$. This number is the number of plants minus 1, because each plant corresponds to one $u_{2(j)}$, and the parameters satisfy one constraint, $\sum_j u_{2(j)} = 1$.

The following very useful result is proved in advanced statistics texts: Under rather general conditions, when H_0 is true, the asymptotic distribution of $-2\log\Lambda$ is χ^2 , with degrees of freedom equal to the difference in dimensions of H_1 and H_0 . Because H_0 should be rejected when Λ is small, it follows that H_0 should be rejected when $-2\log\Lambda$ is large, with the cutoff being a percentile of the χ^2 distribution.

For the above example hypotheses for the RCIC/HPCI Example, when H_0 is true $-2\log\Lambda$ is approximately χ^2 , and the degrees of freedom is the number of plants minus 1.

A data analyst must consider whether the sample is large enough to use the χ^2 approximation. This issue is discussed in a subsection below. The analyst must also occasionally pay attention to one requirement of the theorem: The true value of θ is assumed to be in the *interior* of the spaces defined by H_0 and H_1 . The statistic $-2\log\Lambda$ does not have the asymptotic χ^2 distribution if the true θ is on a boundary. For example, in Figure 1b, if H_0 specified that $\pi_2 = 0$, H_0 would lie on one edge of the triangle H_1 . When H_0 is true, the true θ would be a boundary point of H_1 , and the assumptions of the

theorem would not hold. Even if H_0 is in the interior of H_1 , if an MLE of θ is a boundary point (for example, if $\hat{\pi}_i = 0$ for some i), the sample size is probably too small for the χ^2 distribution to be close to accurate.

When the N_i 's have independent Poisson(m_i) distributions, or when they have independent multinomial distributions with $E(N_i) = n_+ p_i = m_i$, it is not hard to show that

$$-2\log\Lambda = 2\sum n_i [\log(\hat{m}_{1i}) - \log(\hat{m}_{0i})] , \quad (6)$$

where $\hat{m}_{1i}$ and $\hat{m}_{0i}$ are the MLEs of m_i under H_1 and H_0 , respectively. Following a widely used notation, this quantity will be denoted by $G^2(H_0 | H_1)$ here.

Goodness-of-Fit Test, and Residuals. A special case of the likelihood ratio statistic arises when the less restrictive hypothesis is H_∞ , the hypothesis of no restrictions on the means. For Poisson counts, this hypothesis allows every observation (every cell) to have its own mean, unrelated in any way to the other means. For multinomial counts, H_∞ relates the cell means only by the fact that they sum to the observed total count n_+ . In either case, under H_∞ each cell mean is estimated by the observed cell count. Then $-2\log\Lambda$ is denoted G^2 in the goodness-of-fit literature. In the literature of generalized linear models it is called the *deviance*, and is denoted by D :

$$D \equiv G^2 = 2\sum n_i [\log(n_i) - \log(\hat{m}_i)] . \quad (7)$$

For simplicity, the estimated mean under H_0 is denoted $\hat{m}_i$, without the extra 0 subscript. G^2 , or D , has an asymptotic χ^2 distribution, with degrees of freedom equal to the dimension of H_∞ minus the dimension of H_0 . The asymptotic distribution holds as the cell counts become large, for a fixed number of cells; warnings concerning smaller sample sizes are given in a subsection below. The test based on the deviance is the likelihood ratio test against H_∞ . If this test rejects H_0 , it says that H_0 does not model the data adequately, but it does not say how to change the model to improve the fit.

The corresponding *residuals* are defined so that D is the sum of the squared residuals. This is not quite as simple as one might suppose, because some of the terms in Expression (7) can be negative. This is a result of having dropped terms that sum to zero in Equations (6) and (7). The algebraic formulas for the residuals are given in Section B-5 of Appendix B, but are not important for applied work. The important facts are that the residuals here are analogues of residuals in linear models, and the sum of their squares is the deviance. The i th residual is positive if $n_i > \hat{m}_i$ and negative otherwise. The residuals' numerical values are printed by standard software and can be used to show departures from the assumed model. To distinguish these from other residuals defined elsewhere, they are also called the *deviance residuals*.

Pearson Chi-Square Test

Goodness-of-Fit Test, and Residuals. As originally developed in the early 1900s, the Pearson chi-square test was a goodness-of-fit test. The statistic is defined as

$$X^2 = \sum (n_i - \hat{m}_i)^2 / \hat{m}_i ,$$

where n_i is the count in the i th cell, $m_i = E(N_i)$ when H_0 is true, and $\hat{m}_i$ is the MLE of m_i . Special cases of this statistic are discussed in many statistics books and in Atwood (1994a) and Engelhardt (1994). This is a test against H_∞ . The asymptotic distribution of X^2 is χ^2 , with degrees of freedom equal to the difference in dimensions between H_∞ and H_0 . This is the same as the asymptotic distribution of D .

The corresponding residuals are defined so that X^2 is the sum of the squared residuals: the i th residual is the square root of $(n_i - \hat{m}_i)^2 / \hat{m}_i$, with a positive sign if $n_i > \hat{m}_i$ and a negative sign otherwise. These are called the *Pearson residuals* to distinguish them from other residuals.

Test of a Nested Model against a More General Model. The above test can be generalized to the case when H_0 is a lower dimensional subspace of H_1 . The test statistic is

$$X^2(H_0 | H_1) = \sum_i (\hat{m}_{1i} - \hat{m}_{0i})^2 / \hat{m}_{0i} .$$

The notation parallels that for $G^2(H_0 | H_1)$. This statistic is not discussed in most textbooks but is presented by Agresti (1990, pp. 214 and 250). Under the same assumptions as used for the likelihood ratio test, this statistic has an asymptotic χ^2 distribution, with degrees of freedom equal to the difference in dimension between H_1 and H_0 . Note that when H_1 equals H_∞ , the hypothesis of no special restrictions on the means, the test statistic $X^2(H_0 | H_\infty)$ reduces to the goodness-of-fit statistic X^2 . For a general pair of nested hypotheses, $X^2(H_0 | H_1)$ is not computed by standard software. A program for performing the computation can be written in SAS, and an outline of such a program appears in Section A-1 of Appendix A. The details of the program depend on the models specified by H_0 and H_1 .

Relation between the Likelihood Ratio and Pearson Tests

The following discussion is based in part on Section 7.1.3 of Agresti (1990). It can be shown that $X^2(H_0 | H_1)$ and $G^2(H_0 | H_1)$ are asymptotically equal (see Section 12.3.4 of Agresti 1990); that is, for large samples, with high probability the test statistics have almost the same value. The two statistics therefore have the same asymptotic distribution. The subsection below considers the situation when the samples are small.

$G^2(H_0 \mid H_1)$ is popular among analysts who usually work with large data sets because it can be partitioned among nested hypotheses. That is, suppose H_0 is a restriction of a hypothesis H_1 , which in turn is a restriction of H_2 , etc. From Equation (6), it follows that

$$G^2(H_0 \mid H_2) = G^2(H_0 \mid H_1) + G^2(H_1 \mid H_2). \quad (8)$$

This beautiful result is analogous to the partitioning of sums of squares for nested linear models. It says that the chi-square statistics for testing between intermediate models add up to the chi-square statistic for testing the difference between the two ends of the chain. Suppose that the term on the left is large, so that H_0 is rejected in favor of H_2 . The breakdown on the right side shows whether this is primarily a difference between H_0 and H_1 , or primarily a difference between H_1 and H_2 , or a moderate difference at both stages. This result allows stepwise model building, analogous to stepwise regression in the context of linear models.

Unfortunately, X^2 does not behave in the same way. The analogue of Equation (8) is approximately true for large data sets, but it can be far from true for small data sets. For any actual problem, the analyst can compute the terms to see how good the approximation is.

Thus, the likelihood ratio statistic has the theoretical advantage of satisfying Equation (8). However, as will be seen below, the Pearson statistic has the practical advantage of approaching its asymptotic χ^2 distribution faster. Because each statistic has its advantages, data analysts often work with both statistics together.

Small Sample Sizes

A data analyst must consider whether the sample is large enough to use the χ^2 approximations. This issue is discussed in Section B-6 of Appendix B, and is sketched very briefly here. Some important details are given in the Appendix but omitted here!

First consider goodness-of-fit tests. In one-way tables (for Poisson counts) or $2 \times J$ contingency tables (for binomial counts), the χ^2 approximation for X^2 is typically adequate for testing at the $\alpha = 0.05$ level if the expected cell counts are 1.0 or greater. The statistic G^2 , however, converges to its χ^2 distribution more slowly than does X^2 ; expected cell sizes of 5.0 are considered sufficient for G^2 .

Now consider tests of nested hypotheses. The test statistics $G^2(H_0 \mid H_1)$ and $X^2(H_0 \mid H_1)$ approach the asymptotic distribution faster than do the corresponding goodness-of-fit statistics G^2 and X^2 , especially when the difference in dimensions between H_0 and H_1 is small. This helps explain why the SAS program GENMOD prints p -values corresponding to adding parameters to the model, although it refuses to print p -values corresponding to the calculated goodness-of-fit statistics.

For small samples, it can happen that the two asymptotic approximations for X^2 and G^2 give strongly contradictory results. If both statistics are printed and they do not give the same conclusion, the sample is probably too small to use the χ^2 distribution for G^2 . A commonly used rule of thumb is based on computing both test statistics; if the two tests agree, the sample size is probably large enough. I do not know if pathological examples exist in which this rule results in wrong conclusions.

A second approach, which will surely become more popular as software evolves, is to use exact distributions rather than asymptotic ones. Commercial computer packages that compute exact p -values are discussed in Appendix B-6.

EXAMPLE ANALYSES

These example analyses were performed using the SAS procedure GENMOD (for GENERALized linear MODEL), available in SAS version 6.09 (SAS 1993) and higher. The analyses could also be performed using the procedure NLIN (for NonLINEar least squares), available in SAS version 6.0 and higher. To use NLIN for loglinear or logit models, a macro GLIM, printed in the SAS (1989) user's guide, must be invoked. Loglinear models may also be analyzed using the SAS (1989) procedure CATMOD (for CATegorical MODELing). However, CATMOD is extremely difficult to understand. GENMOD is used here because it has simpler syntax and produces more outputs for testing goodness of fit than NLIN or CATMOD.

RCIC/HPCI Example:

Binomial Response, Categorical Explanatory Variables

This example study uses the data from the RCIC/HPCI Example, failures on demand of turbine-driven pump assemblies in two systems at 23 plants. First, we mention some simple analyses that are *not* adequate.

One might wish to pool the data from the systems, and see if an overall difference between plants exists. This test with pooled data could lead to erroneous conclusions, because one plant might have relatively more RCIC demands and another relatively more HPCI demands. If the two systems have different overall probabilities of failure, and the differences in demands at the plants were not accounted for, the analysis could mistakenly attribute differences from the systems to differences between plants. This would be an instance of Simpson's paradox, mentioned in connection with Table 1 of this report.

Similarly, one could analyze a 2×2 table to see if the systems differ for any one plant. However, one should not pool the data from the plants to see if there is an overall difference between systems. Ignoring the between-plant variability in this way can lead to erroneous conclusions.

To analyze all the data at once, we must construct a model that accounts for both factors, system and plant, simultaneously, as follows. Let j denote plant and k denote system, and let π_{jk} denote the probability of failure on demand in system k at plant j . Following Equation (4), we attempt to fit the model

$$\log[\pi_{jk}/(1 - \pi_{jk})] = u + u_{\text{plant}(j)} + u_{\text{sys}(k)} . \quad (9)$$

This is the simplest model that involves both plant and system as factors with unordered levels. The next step in complexity would be to add an interaction term $u_{\text{plant} \times \text{sys}(jk)}$. This model would say that every plant

system has its own $\pi_{j\cdot}$, unrelated to the others; that is, this model would be H_{10} . If the model without interactions turns out to give a good fit, we will conclude that some plants have a tendency to be bad for both systems and some tend to be good for both systems.

The SAS procedure GENMOD performs analysis using the formulation given in Section B-2 of Appendix B for a generalized linear model. The full SAS program is shown at the beginning of Section A-2 in Appendix A. (Actually, two analyses are performed, as discussed below; the programs are virtually identical, and only the second is shown.) The call to GENMOD declares that the data follow a binomial distribution and that the logit transformation is to be used. It states that the plant and system are "classification variables," that is, unordered categorical variables rather than numerical variables. And it states that the model is to include a main effect for plant and a main effect for system (with the overall mean understood). The only other specified options call for certain outputs—tests of various models and various statistics associated with the observations, such as fitted means and residuals. A portion of the output is shown in Table 2.

Table 2. Portion of initial analysis of the RCIC/HPCI Example.

Analysis Of Parameter Estimates						
Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-2.2577	0.6063	13.8676	0.0002
SYS	RCIC	1	-1.3931	0.3234	18.5508	0.0001
SYS	HPCI	0	0.0000	0.0000	.	.
PLANT	A	1	0.0445	0.8520	0.0027	0.9584
PLANT	B	1	-25.7661	128372.191	0.0000	0.9998
PLANT	C	1	-24.6961	128097.418	0.0000	0.9998
PLANT	D	1	-2.1276	1.1732	3.2891	0.0697
PLANT	E	1	0.7367	0.7474	0.9714	0.3243
PLANT	F	1	-1.1134	1.1796	0.8910	0.3452
PLANT	G	1	-1.5426	0.9337	2.7294	0.0985
PLANT	H	1	-0.2269	0.7177	0.0999	0.7519
PLANT	I	1	-0.8219	0.7887	1.0862	0.2973
PLANT	J	1	-25.3397	111369.559	0.0000	0.9998
PLANT	K	1	-2.2589	1.1712	3.7201	0.0538
PLANT	L	1	-0.6748	0.7559	0.7970	0.3720
PLANT	M	1	-1.6100	0.9335	2.9746	0.0846
PLANT	N	1	-1.1554	0.8396	1.8937	0.1688
PLANT	O	1	-0.0920	0.7211	0.0163	0.8985
PLANT	P	1	-1.5398	0.9338	2.7193	0.0991
PLANT	Q	1	0.4272	0.8600	0.2468	0.6194
PLANT	R	1	0.0254	0.8515	0.0009	0.9762
PLANT	S	1	-0.7099	1.1842	0.3593	0.5489
PLANT	T	1	-1.4080	1.1767	1.4318	0.2315
PLANT	U	1	0.0851	0.7630	0.0124	0.9112
PLANT	V	1	-0.4342	0.9438	0.2116	0.6455
PLANT	W	0	0.0000	0.0000	.	.

Note that some of the plant effects are estimated with enormous standard errors. These effects correspond to the plants with no failures at all. For each of those plants, and for each system, the MLE of the failure probability π_{ij} is 0. It follows that $\text{logit}(\pi_{ij}) = \log[0/(1-0)] = -\infty$, and therefore, the

estimated plant effect is also $-\infty$. Because $-\infty$ is not a number, the computer algorithm fails. Also, the estimated zeros for the failure probabilities put the MLE on the boundary of the possible space of parameters (on the edge of the triangle in Fig. 1b.) Therefore, the asymptotic theory is clearly invalid, and if distributional information, such as a list of the asymptotic variances, is requested, the SAS log contains warning messages complaining about an asymptotic covariance matrix that is not positive definite. The solution for the analyst is to recognize that, for plants with no failures in any system, the MLE of the failure probability π is zero, and the plant contributes no information about differences between systems.

Therefore, the three plants B, C, and J were dropped from the data set, and the remaining observations were again analyzed by the procedure GENMOD. An abbreviated and edited portion of the resulting output is shown in Table 3.

The section that SAS labels **Type 3 Analysis** shows G^2 for testing the model without the identified main effect, against the alternative model of all main effects. This corresponds to model selection by backward elimination of effects. As can be seen, dropping system from the model produces a chi-square statistic of 22.43 with 1 degree of freedom, and this is rejected with p -value 0.0001. The model that does not include plant effects is also rejected, but not as strongly. These results suggest building the model in a stepwise way, entering system effect into the model before plant effect.

This was done, and the portion of the output labeled **Type 1 Analysis** (SAS terminology for stepwise testing of nested models, as would be appropriate with forward selection of effects) shows the stepwise results. Let H_0 be the model with no main effects, let H_1 include only the system effect, let H_2 be the model with both main effects, and let H_∞ be the model with no special restrictions on the π_{jk} 's. The terms labeled **Deviance** are the goodness-of-fit statistics, $G^2(H_i | H_\infty)$, and the terms labeled **ChiSquare** are the test statistics for nested hypotheses, $G^2(H_{i-1} | H_i) = G^2(H_{i-1} | H_\infty) - G^2(H_i | H_\infty)$. Thus, adding a system effect to the model with no main effects gives a much better fit, with p -value 0.0001, and adding a plant effect to that model also improves the fit by a statistically significant amount, with p -value 0.0167. However, even this model is not good enough. The deviance for the full model, measuring goodness of fit, is 33.5254. If the model were correct, this deviance would have a χ^2 distribution with 19 degrees of freedom. SAS is unwilling to print the χ^2 tail probability, because often the asymptotic distribution is not very accurate, but in this case the deviance, G^2 , and the Pearson chi-square statistic, X^2 , are similar (as seen in the output portion labeled **Criteria For Assessing Goodness of Fit**.) This suggests that the sample size is large enough for the χ^2 approximation to be adequate. A glance at a χ^2 table in a statistics book shows that G^2 is beyond the 97.5th percentile of a χ^2 distribution with 19 degrees of freedom, and X^2 is beyond the 99th percentile. Therefore, we reject the fit of the overall model.

Table 3. Analysis of the RCIC/HPCI Example, deleting plants with no failures.

The GENMOD Procedure				
Model Information				
Description	Value			
Data Set	WORK.TDPFIDM			
Distribution	BINOMIAL			
Link Function	LOGIT			
Dependent Variable	NF			
Dependent Variable	ND			
Observations Used	40			
Number Of Events	62			
Number Of Trials	1921			
Class Level Information				
Class	Levels	Values		
SYS	2	RCIC HPCI		
PLANT	20	A D E F G H I K L M N O P Q R S T U V W		
Criteria For Assessing Goodness Of Fit				
Criterion	DF	Value	Value/DF	
Deviance	19	33.5254	1.7645	
Pearson Chi-Square	19	36.3049	1.9108	
Log Likelihood	.	-246.2075	.	
LR Statistics For Type 1 Analysis				
Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	88.8373	0	.	.
SYS	67.8717	1	20.9657	0.0001
PLANT	33.5254	19	34.3462	0.0167
LR Statistics For Type 3 Analysis				
Source	DF	ChiSquare	Pr>Chi	
SYS	1	22.4346	0.0001	
PLANT	19	34.3462	0.0167	

We conclude that the model with two main effects is too simple. The next terms that could be added to the loglinear model are interaction terms, $u_{\text{plant} \times \text{sys}}(\bar{x})$. As remarked above, the model with all the interaction terms is just H_∞ , the model with no restrictions on the π_{jk} 's. Under this model, the data from either system contribute negligible information about the other system, and the two systems might as well be analyzed separately. For such analyses, the loglinear model is not needed; see Atwood (1994a).

However, intermediate models exist between H_0 and the model of Table 3. Consider briefly whether any of these might be superior to the models discussed so far. The full output, including a list of the residuals, is shown in Section A-2 of Appendix A. The residuals are listed in the same order as the input data, so by counting down one can match residuals with plant-system combinations. The deviance residuals, labeled *Resdev*, and the Pearson residuals, labeled *Reschi*, do not rank the plant-system inputs in exactly the same way, but one large contributor is PlantD-RCIC (*Resdev* = 1.8205). Including the interaction term for Plant D would give perfect fit for PlantD-RCIC and PlantD-HPCI, thereby reducing those two residuals from 1.8205 and -1.3628 to 0 and 0. All the other residuals would change slightly, but the deviance would be reduced by *approximately* $1.8205^2 + (-1.3628)^2 = 5.17$. The resulting deviance would be approximately $33.5254 - 5.17$, with 18 degrees of freedom, corresponding to a *p*-value of approximately 0.05. These calculations are rough, and would need to be carried out on the computer to yield exact values. If one interaction were a dominant contributor to the deviance, the model with that one interaction might be justified. However, that does not appear to be the case here. For example, Plant U has a larger effect than Plant D in terms of the deviance residuals, though not in terms of the Pearson residuals. If no one interaction is the major contributor, as appears to be the case here, fitting the two systems separately is simple and justifiable.

Finally, when interpreting these results, recall the warning given with the RCIC/HPCI Example, that the data are not of the highest quality and should not be used to draw real conclusions about the HPCI and RCIC systems. One possible cause of a large deviance is a model with too few parameters. A second possible cause is overdispersion in the data—numbers with a larger variance than assumed by the binomial or Poisson model. This can result from occasional uncounted or misclassified events randomly scattered through the data. It is impossible to determine from the data whether the more extreme counts are systematic, reflecting interactions, or random, reflecting a large variance. However, in view of questions about the data quality, the second cause may apply to this example.

Unplanned-Demands Example: Poisson Regression

This analysis uses the data from the example of unplanned demands on the HPCI system. Because these events can occur only when a plant is operating, it is natural to treat the plant operating years as the *exposure time*. That is, we assume that in calendar year i , the number of events, N_i , has a $\text{Poisson}(\lambda_i t_i)$ distribution, where t_i is the exposure time in the year, and λ_i is the event rate, with units 1/year. A simple chi-square test, together with a graph of the estimated event rates by year, (for example, following Engelhardt 1994) shows that the event rate is not the same over all the years and may be decreasing. The nonconstancy was anticipated by the notation here, λ_i instead of simply λ . Now we try to construct a loglinear model that includes the trend, if in fact the variation can be modeled in that way.

The simplest nonconstant loglinear model is

$$\log[E(N/t_i)] = \beta_0 + i\beta_1 \quad (10)$$

This is analogous to a linear trend in time, except the trend is linear for the log of the mean, not for the mean. The index i can be the year (1987 through 1993), or a two-digit version of the year (87 through 93), or simply a number 1 through 7. If any computations are to be made based on the output, the third approach is definitely best, for the following reason. The term β_0 is $\log(\lambda_i)$ when i equals 0. If i is the calendar year, $i=0$ is the year 0, nearly 20 centuries ago; if i is the two-digit year, $i=0$ refers to the year 1900; if i ranges from 1 to 7, $i=0$ corresponds to 1986. The third case yields by far the least correlation between the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, and the least risk of losing all numerical accuracy in subsequent calculations because the printouts (or perhaps the computer's internally stored values) are not accurate enough. This is related to centering the data in regression models. To carry this centering to an extreme, we could let i run from -3 to 3 , with 0 corresponding to 1990. Letting i range from 1 to 7 gives enough accuracy, however.

The SAS procedure GENMOD was used to analyze this model. The full SAS program is shown in Section A-4 of Appendix A. The data set, constructed from the Unplanned-Demands Example, is as shown here. YEAR indexes the calendar year, with 1 corresponding to 1987, UDEMS is the count of unplanned demands, OPYRS is the total plant-operating-years during the calendar year, and LOPYRS is constructed to equal $\log(\text{OPYRS})$.

YEAR	UDEMS	OPYRS	LOPYRS
1	16	14.98	2.70672
2	10	14.27	2.65816
3	7	15.89	2.76569
4	13	18.16	2.89922
5	9	17.76	2.87695
6	6	17.63	2.86960
7	2	17.91	2.88536

The call to PROC GENMOD specifies that the model is linear in YEAR, that UDEMS has a Poisson distribution, that the model has a logarithmic link and an offset = LOPYRS. The other specifications call for certain outputs. For example, 90% confidence intervals, corresponding to $\alpha = 0.10$, are requested instead of the default 95% confidence intervals. The program and the output are shown in Appendix A. An abbreviated and rearranged version of the output is shown here in Table 4.

Table 4 first shows the model specifications. The estimates of the two parameters, β_0 and β_1 , are given next, with the estimated standard deviations (the *standard errors*) of the estimators. The approximate normality of $\hat{\beta}$ is used to treat $[\hat{\beta}_i/(\text{Std Err}_i)]^2$ as a χ^2 random variable. For β_1 , identified by the name YEAR in the printout, the estimate is much larger than its standard error, corresponding to a χ^2 value of 12.8, far in the tail of the $\chi^2(1)$ distribution. Thus, the parameter, corresponding to a slope in Equation (10), is clearly nonzero. The estimated intercept is approximately one standard

Table 4. Portion of analysis printout for the Unplanned-Demands Example.

The GENMOD Procedure

Model Information

Description	Value
Data Set	LIB.HPCIDEMS
Distribution	POISSON
Link Function	LOG
Dependent Variable	UDEMS
Offset Variable	LOPYRS
Observations Used	7

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	0.2623	0.2494	1.1061	0.2929
YEAR	1	-0.2381	0.0665	12.8291	0.0003

Estimated Covariance Matrix

0.06220	-0.01431
-0.01431	0.004418

Estimated Correlation Matrix

1.0000	-0.8630
-0.8630	1.0000

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	5	4.9803	0.9961
Pearson Chi-Square	5	4.7619	0.9524
Log Likelihood	.	80.7747	.

Observation Statistics

UDEMS	Pred	Lower	Upper	Resraw	Reschi	Resdev
16	15.3471	11.1369	21.1490	0.6529	0.1667	0.1655
10	11.5223	8.9958	14.7584	-1.5223	-0.4485	-0.4589
7	10.1120	8.2060	12.4608	-3.1120	-0.9786	-1.0367
13	9.1082	7.2850	11.3876	3.8918	1.2895	1.2110
9	7.0203	5.2903	9.3162	1.9797	0.7472	0.7156
6	5.4925	3.8081	7.9218	0.5075	0.2166	0.2133
2	4.3976	2.7745	6.9700	-2.3976	-1.1433	-1.2820

LR Statistics For Type 1 Analysis

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	18.4579	0	.	.
YEAR	4.9803	1	13.4775	0.0002

error from zero, but this is of no special interest; if β_0 were exactly zero, the Poisson rate would be exactly 1.0 [= $\exp(0)$] in 1986, but there is nothing special about either 1986 or the number 1.0. Therefore, we continue to leave β_0 in the model, and ignore the output that says that β_0 might equal zero. We note also that $\hat{\beta}_0$ and $\hat{\beta}_1$ have a fairly large correlation, -0.86 . The correlation is not so large that all numerical accuracy is lost, as would happen if, for example, i ranged from 1,987 to 1,993. The correlation could be reduced by letting i range from -3 to 3 , but this is not necessary for our purposes.

Table 4 also shows the goodness-of-fit statistics G^2 (or D), labeled Deviance in the output, and X^2 , labeled Pearson Chi-Square in the printout. The two statistics are numerically close to each other, suggesting that the sample size is large enough to perform the likelihood-ratio analysis. If the model is correct, G^2 and X^2 each have approximately a χ^2 distribution with 5 degrees of freedom (7 cells minus 2 parameters). The calculated values are close to 5, near the middle of the χ^2 distribution, giving no evidence of bad fit. The residuals can also be used to check the goodness of fit. Three sets of residuals are printed in the section labeled Observation Statistics. The raw residuals are $N_i - \hat{\lambda}_{t_i}$, and the Pearson and deviance residuals are as defined earlier. These are printed as Resraw, Reschi, and Resdev, respectively. The Pearson and deviance residuals should be approximately normal(0, 1) if the fitted model is correct. Quick examination of the values shows no evidence to the contrary—no extreme outliers, no strong skewness, no clumping or wide gaps. All three residuals are defined to have the same sign at any data point. These signs show no surprising pattern—the data points seem to fall above and below the fitted line more or less at random. We conclude that the assumed model fits the data adequately.

A slight refinement of this examination of the residuals is possible under version 6.10 of SAS (1994). In this release, PROC GENMOD allows the statement RESIDUALS, causing it to print not only the residuals shown in Table 4, but also the standardized residuals. SAS calls them Studentized residuals, but not all authors use this terminology (see Engelhardt 1995). These residuals are adjusted for the fact that ordinary residuals do not all have the same variance. The standardization (or in SAS terms, Studentization) results in residuals with mean 0 and variance 1, if the assumed model is correct. The individual standardized residuals are somewhat more relevant for comparison with a normal distribution. However, in this example the standardization results in little change, and so is not shown.

The section of the output labeled Type 1 Analysis shows the effect of adding parameters to the model sequentially. Let H_0 be the model that the Poisson rate λ is constant—in Equation (10), β_1 is zero. Let H_1 be the model specified by Equation (10). Let H_∞ be the model that each calendar year has its own event rate, unrelated to the others. Recall that the deviance is G^2 for testing an assumed model against H_∞ . The printout shows

$$G^2(H_0 \mid H_\infty) = 18.4579$$

$$G^2(H_1 \mid H_\infty) = 4.9803$$

$$G^2(H_0 \mid H_1) = 13.4775 (= 18.4579 - 4.9803) .$$

If H_0 is true, the test statistic $G^2(H_0 | H_1)$ is χ^2 distributed; the distribution has one degree of freedom, because H_1 has dimension 2 (2 unconstrained parameters), one more than H_0 . The calculated value is large, the $1 - 0.0002$ quantile of the distribution, leading us to reject H_0 in favor of H_1 .

Should we continue the sequence and try a more complex H_2 , for example, $\beta_0 + i\beta_1 + i^2\beta_2$? Probably not, because we have already determined that the model fits adequately. Furthermore, any $G^2(H_1 | H_2)$ must necessarily be less than $G^2(H_1 | H_\infty) = 4.98$, which is approximately the 97th percentile of the $\chi^2(1)$ distribution. The one extra parameter included in H_2 would have to account for virtually all of the residual variation, to make $G^2(H_1 | H_2) \approx 4.98$ and $G^2(H_2 | H_\infty) \approx 0$. It seems unlikely that this could happen, but if it did, the formal test procedure would reject H_1 in favor of H_2 . However, the resulting deviance would be extremely small, suggesting that H_2 is overfitted. The model would have to be assessed on engineering and physical grounds.

PROC GENMOD finds the fitted mean counts (λ_{t_i} , labeled as Pred in the output) with lower and upper confidence bounds on the fitted value. We are more interested in the fitted rate λ_i . This, and the confidence bounds, are all calculated from the printed values by dividing them by $t_i = \text{OPYRS}$. The resulting confidence bound is valid for any single value of i .

The bounds are not valid simultaneously for the entire curve $\lambda = \exp(\beta_0 + i\beta_1)$. The confidence statement at some year, is

$$P[\text{confidence interval for year}_i \text{ contains the true rate for year}_i] = 0.90 \quad (11)$$

A similar confidence statement applies to each year. The simultaneous statement would involve

$$P[\text{confidence interval for year}_i \text{ contains the true rate for year}_i \text{ AND} \\ \text{confidence interval for year}_j \text{ contains the true rate for year}_j \text{ AND} \\ \text{so forth}] \quad (12)$$

This probability is hard to quantify, because the intervals are all calculated from the same data set, and thus are correlated. However, Expression (12) is certainly smaller than 0.90, because the event in brackets in Expression (12) is more restrictive than the event in brackets in Equation (11). A simultaneous confidence band on the fitted curve for λ is not found by GENMOD, but was calculated by a program given in Section A-3 of Appendix A, using the formulas derived in Section B-6 of Appendix B. The two bands, one pointwise and one simultaneous, are shown in Figure 2. The band that is simultaneously valid for the entire line is slightly wider than the band that is valid only at any one point; the extra width is the price of the desired simultaneous coverage.

Figure 2. Pointwise 90% confidence intervals and simultaneous 90% confidence band for Poisson event rate in the Unplanned-Demands Example.

It is of interest to compare this result using a loglinear model with the corresponding result from a linear model. In Grant et al. (1995), the same data set was analyzed as follows. The values of $\log(N/t_i)$ were fitted by ordinary least squares to a line of the form $\beta_0 + i\beta_1$. A simultaneous 90% confidence band was found for the fitted line. The confidence intervals were based on treating the observed values $\log(N/t_i)$ as approximately normally distributed, and estimating the variance from the observed scatter around the fitted line. The two fitted lines and confidence bands are shown as the left and center portions of Figure 3. The observed values, plotted as dots, are the same in the two portions of the figure. As can be seen, the fitted lines are very similar, with each one included in the confidence band based on the other method. The most striking difference between the plots is the different widths of the confidence bands near 1987, the year corresponding to $i = 1$.

Why do the bands differ so much? One reason is that the variance of N/t is not constant. The variance of $\log(N/t)$ is approximately the relative variance of N/t , which is $1/E(N)$. Thus, the years with large counts provide more information (smaller variance) per year than do the years with small counts. The loglinear model accounts for this automatically. The resulting confidence band is pinched to its minimum (when expressed in terms of $\log\lambda$) somewhat to the left of the midpoint, 1990. When, instead,

Figure 3. Fitted Poisson event rate and simultaneous 90% confidence band, based on three ways of fitting the Unplanned-Demands data.

the method of unweighted least squares is used with $\log(N/t)$, the resulting confidence band has its minimum width (when expressed in terms of $\log\lambda$) at the exact midpoint. This difference in the locations of the minimum widths is one reason why the least-squares confidence band is so much wider than the loglinear band for 1987.

To improve the least-squares fits, one could use weighted least squares, basing the weights on the initial fitted values of $E(N)$. This would yield slightly different fitted values. Then one could adjust the weights based on the new fitted values. This iterative reweighting process was carried out, and answers became stable after several refitting steps. The resulting fitted line and simultaneous 90% confidence band are shown in the right portion of Figure 3. As can be seen, the method of iteratively reweighted least squares yields a fitted line that is very close to that from the loglinear model. However, the confidence band is not as narrow as that from the loglinear model. The reason is that the loglinear model assumes that the variance can be calculated from the fitted mean of the Poisson data, whereas the least squares method estimates the variance from the observed scatter around the line. The least squares method accounts for this extra estimated term by using a larger multiplier, $[2F_{0.90}(2,5)]^{1/2}$, than the corresponding multiplier, $[\chi^2_{0.90}(2)]^{1/2}$, for the loglinear model.

In summary, all three calculations are valid, and yield similar fitted lines. The method of unweighted least squares uses the data in a somewhat inefficient way, however, and therefore produces an unrealistically wide confidence band. The method of iteratively reweighted least squares provides a narrower confidence band, and the loglinear model provides the narrowest band of all. None of the calculations is exact: the least squares method treats $\log(\text{count}/\text{time})$ as approximately normal, and the loglinear model treats the parameter estimators as approximately normal. However, the analysis based on the loglinear model is preferable (if the Poisson assumption is correct!), because it gives the tightest confidence band, and reweighted least squares is second best. In cases where extra sources of variation might have entered the data, variation beyond that caused by the Poisson distribution (also called overdispersion), reweighted least squares would be best and the loglinear model would produce an unrealistically narrow band.

MOV-Test Example: Logistic Regression

The MOV-Test Example has two kinds of explanatory variables. Year is numerical and ordered, suggesting a regression analysis, analogous to the Poisson regression of the Unplanned-Demands Example. Plant is categorical and unordered, suggesting an analysis of the type performed in the RCIC/HPCI Example. The following model incorporates both explanatory variables. Let j index plant, for j from 1 to 2, and let k index the year, for k from 1 to 9. We reserve the index i for the outcome, with $i = 1$ corresponding to failures and $i = 2$ corresponding to successes. The full model is written most easily in terms of the variable k and a dummy variable x_{plant} , which is defined as 1 if the plant is Plant 1 and 0 if the plant is Plant 2. The full model then is:

$$\text{logit}(\pi_{ijk}) = \log[\pi_{ijk}/(1 - \pi_{ijk})] = \beta_0 + x_{\text{plant}}\beta_1 + k\beta_2 + x_{\text{plant}}k\beta_3 . \quad (13)$$

The presence of β_1 allows the plants to have different overall failure probabilities. The presence of β_2 allows a trend in each failure probability, with β_2 positive if the failure probability is increasing and negative if the failure probability is decreasing. From engineering considerations, if the valves are not being maintained or replaced, we expect β_2 to be positive. The presence of β_3 allows the trends to be different at the different plants—valves at one plant may wear out faster than at the other plant. We now experiment with the fits of various models. It is not clear from the data whether year or plant is more important, so we build the full model by entering year, plant, and the interaction in that order, using the SAS procedure GENMOD. In a separate call to GENMOD we build the model by entering first plant, then year and the interaction term. The full program and output are shown in Section A-5 of Appendix A. Portions of the output are shown in Table 5.

The goodness-of-fit portion of the output shows that the model seems to fit the data very well. $G^2=8.6776$ and $X^2=8.4932$ are both smaller than their degrees of freedom, 14. The portion of the printout labeled `Type 3 Analysis` shows the effect of leaving out each variable from the full model. Year, with a p -value of 0.0053, is the only one of the variables that may not be deleted. The interaction term, with a p -value of 0.8203, could be deleted from the full model with virtually no increase in lack of fit. Upon closer investigation, the conclusion for Year is slightly suspect, because of idiosyncrasies of SAS. If Plant is declared to be a categorical variable (a "classification" variable, in SAS terminology), when SAS enters the variables sequentially it does not allow Year*Plant to be entered before Year is entered. When deleting Year from the full model, as discussed here, SAS may also be deleting Year*Plant; it is difficult to identify G^2 terms that add up to reveal what SAS is doing. If instead Plant is numerical, a dummy variable that equals 0 or 1, then any term may be entered in any order.

The portions of the printout labeled `Type 1 Analysis` show the effect of adding variables to the model sequentially. When Year is added first, adding Plant next improves the fit substantially (p -value = 0.0064). When, instead, Plant is added first, adding Year also improves the fit substantially (p -value = 0.0039). In no case does the interaction improve the fit much, provided the other terms are already in the model.

Table 5 illustrates the fact that a p -value for a source depends on what else is in the model. If Plant and Year were highly correlated—for example, if some plants had data only for earlier years and others only for later years—adding Plant would improve the fit much more if Year were not yet in the model than if Year were already in the model. The improvement would be less in the latter case, because much of the information contained in Plant would already have been incorporated into the model through Year. A large improvement in the fit corresponds to a small p -value (strong statistical significance). As it is, Plant and Year seem to be nearly independent, because the p -value for Plant is 0.0066 when Plant is added first, and nearly as large, 0.0064, when Plant is added second. When Plant is added third, on

Table 5. Preliminary analysis of the MOV-Test Example.

Leaks of MOVs in 2 plants over 9 years
Entering year, plant, and year*plant, in that order

The GENMOD Procedure
Model Information

Description	Value
Data Set	WORK.VALVES
Distribution	BINOMIAL
Link Function	LOGIT
Dependent Variable	NF
Dependent Variable	ND
Observations Used	18
Number Of Events	53
Number Of Trials	936

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	14	8.6776	0.6198
Pearson Chi-Square	14	8.4932	0.6067

LR Statistics For Type 1 Analysis

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	24.4120	0	.	.
YEAR	16.1686	1	8.2434	0.0041
PLANT	8.7292	1	7.4394	0.0064
YEAR*PLANT	8.6776	1	0.0516	0.8203

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
YEAR	1	7.7670	0.0053
PLANT	1	1.5404	0.2146
YEAR*PLANT	1	0.0516	0.8203

Leaks of MOVs in 2 plants over 9 years
Same factors, entering plant before year

LR Statistics For Type 1 Analysis

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	24.4120	0	.	.
PLANT	17.0378	1	7.3741	0.0066
YEAR	8.7292	1	8.3086	0.0039
YEAR*PLANT	8.6776	1	0.0516	0.8203

the other hand, after both Year and the interaction Year*Plant, the p -value is much larger (0.2146), because some of the information in Plant is contained in the interaction.

In summary, this preliminary analysis indicates that the model (13) should be used with the interaction term deleted. A very abbreviated part of the output from the final analysis is shown in Table 6. This output shows no indication of any problem. As a final check on the fit, look at the residuals. The deviance residuals and the Pearson residuals, listed as Resdev and Reschi in the output, do not show any values that are surprisingly large compared to the others. The signs of the residuals do not appear to have any striking pattern. The only anomaly is that the residuals from Plant 1 tend to be somewhat larger in magnitude than those from Plant 2. The four largest residuals correspond to years 1, 4, and 8 at Plant 1, and year 7 at Plant 2. These are the years in the original data with the largest jumps in the failure counts. Perhaps someone at the plants knows why Plant 1 seemed to have somewhat more erratic behavior than Plant 2, and whether this is related to the overall higher failure counts at Plant 1. Without more details about the data and the plants, we can only note that the model fits the data acceptably, and that the reasons for the differences between plants are unknown.

Figure 4 shows the fitted lines for the failure probability, as a function of time. Each plant has its own line, which is shown together with a simultaneous 90% confidence band. The bands are simultaneous in the sense that they are simultaneously valid for all times and for both plants.

Figure 4. Fitted trend lines for the failure probability, and simultaneous confidence bands on the fitted lines at the two plants, for the MOV-Test Example.

Table 6. Final analysis of the MOV-Test Example.

Parameter Information					
Parameter		Effect	PLANT		
PRM1		INTERCEPT			
PRM2		YEAR			
PRM3		PLANT	Plant1		
PRM4		PLANT	Plant2		
Criteria For Assessing Goodness Of Fit					
Criterion	DF	Value	Value/DF		
Deviance	15	8.7292	0.5819		
Pearson Chi-Square	15	8.6372	0.5758		
Log Likelihood	.	-195.8091	.		
Analysis Of Parameter Estimates					
Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-4.1707	0.4247	96.4394	0.0001
YEAR	1	0.1624	0.0578	7.9112	0.0049
PLANT	Plant1	0.8003	0.3031	6.9700	0.0083
PLANT	Plant2	0.0000	0.0000	.	.
Estimated Correlation Matrix					
Parameter Number	PRM1	PRM2	PRM3		
PRM1	1.0000	-0.8121	-0.4898		
PRM2	-0.8121	1.0000	0.0158		
PRM3	-0.4898	0.0158	1.0000		
LR Statistics For Type 3 Analysis					
Source	DF	ChiSquare	Pr>Chi		
YEAR	1	8.3086	0.0039		
PLANT	1	7.4394	0.0064		
Observation Statistics					
NF	ND	Pred	Resraw	Reschi	Resdev
4	52	0.0389	1.9789	1.4199	1.2581
2	52	0.0454	-0.3613	-0.2407	-0.2469
3	52	0.0530	0.2442	0.1512	0.1492
1	52	0.0618	-2.2118	-1.2741	-1.4794
4	52	0.0719	0.2625	0.1409	0.1394
3	52	0.0835	-1.3417	-0.6726	-0.7091
4	52	0.0968	-1.0333	-0.4846	-0.5009
9	52	0.1120	3.1783	1.3978	1.3072
6	52	0.1292	-0.7159	-0.2960	-0.3007
1	52	0.0178	0.0723	0.0757	0.0748
1	52	0.0209	-0.0879	-0.0852	-0.0864
1	52	0.0245	-0.2751	-0.2467	-0.2562
1	52	0.0287	-0.4936	-0.4098	-0.4354
2	52	0.0336	0.2519	0.1938	0.1895
3	52	0.0393	0.9557	0.6819	0.6390
1	52	0.0459	-1.3883	-0.9197	-1.0363
4	52	0.0536	1.2130	0.7469	0.7034
3	52	0.0625	-0.2478	-0.1420	-0.1438

DISCUSSION

This section ties up some loose ends: topics that earlier report sections have touched briefly and topics that were carefully avoided in the earlier sections.

Interactions, and Simpson's Paradox

Graphs can reveal the effects of two explanatory variables, perhaps with interactions. On the vertical axis, plot the observed response. With Poisson data, the observed response could be the observed rate, events divided by exposure time, or it could be the logarithm of the observed rate. With binomial data, such as counts of failure on demand, the observed response could be the observed fraction of demands that resulted in failure, or it could be the logit of this fraction. The details do not matter. For simplicity of presentation, the discussion below speaks of failures on demand, but the analogous statements about events in some exposure time are also true. On the horizontal axis, plot one of the explanatory variables, and make a separate line for each level of the other explanatory variable. Figure 5 shows two such plots, with hypothetical data. To avoid complicating the discussion by consideration of the random noise present in any data, we assume that the data sets here are large, so that the observed responses are close to the theoretical mean responses.

a. Main effects, no interaction

b. Interaction present

Figure 5. Hypothetical plots of observed response.

Figure 5a shows a case with a plant effect and a year effect. Plant A has a higher response than does Plant B, regardless of year, and year 2 has a higher response than the other years, regardless of plant. The fact that the lines are nearly parallel shows that any interaction is very small. If the vertical

axis is logit(failure probability), the absence of an interaction term in the loglinear model corresponds to exactly parallel true lines and approximately parallel estimated lines. If the vertical axis is the untransformed probability, the lines are somewhat less parallel.

If the plants were pooled, the effect of year would still be evident. The dashed line shows the average for the two plants, the result that would be obtained from pooling the plants if the two plants have the same number of demands. Similarly, if the years were pooled, the effect of plant would still be evident.

Figure 5b shows a case with interactions—the lines are not parallel. Year has no overall effect, but for each separate plant it has an effect.

The above discussion has not considered the effect of different numbers of demands. Figure 6 repeats Figure 5a, except now the numbers of demands vary greatly. The large dots correspond to many demands and the small dots correspond to few demands. For example, Plant A had many more demands than Plant B in year 1, and the situation was reversed in year 2.

Figure 6. Plot of observed responses, showing variation in number of demands.

The dashed line shows the response if the plants were pooled. In each year, the plant with the most demands dominates the data, and the dashed line is not halfway between the lines for the individual plants. In this contrived example, pooling the plants nearly obliterates the evidence of a year effect.

The opposite situation, in which pooling creates an apparent effect that is not really present, is seen in the data from Table 1. Based on those data, Figure 7a plots the response (observed mortality = number of deaths divided by the number at risk) against level of prenatal care. Figure 7b uses the same data, but plots response against clinic. Figure 7a shows that Clinic B has a higher response than Clinic A, and that, within each clinic, the response does not depend on level of care. The number of infants with more prenatal care was much greater at Clinic A than at Clinic B (shown by the size of the dots); therefore, the dashed line corresponding to pooling clinics slopes downward. This downward slope appears to show that more care reduces the mortality. However, this conclusion is incorrect, as shown by the level lines for each clinic. The false conclusion results from pooling data sets with different failure probabilities and unbalanced numbers of demands. This is an instance of Simpson's paradox.

Figure 7. Observed mortality versus prenatal care and clinic.

Figure 7b plots the same data, but with a different horizontal axis. The lines have been offset slightly to prevent large dots from covering small dots. Like Figure 7a, this figure shows that care level does not change the mortality within a clinic, and that Clinic B has higher mortality than Clinic A. A dashed line is not shown, because it would lie on top of the other lines and make the picture too crowded. If shown, it would represent the effect of pooling care levels, not the effect of pooling clinics.

Agresti (1990, p. 139) shows a very similar figure to illustrate data involving application of the death penalty to convicted murderers. In that example, the effect of the defendant's race is reversed, depending on whether the race of the victim is considered or ignored. Agresti's diagram makes the area of the circles proportional to the number of data elements represented by the point, instead of merely making the circles "big" and "little" as was done here.

These methods work best when only two explanatory variables are considered. When more explanatory variables are present, one variable is plotted on the horizontal axis, and a separate line is drawn for each combination of levels of the other variables.

In summary, the levels of a variable can be pooled without producing invalid results if (a) the variable does not interact with other variables, and (b) the number of demands is approximately the same for all combinations of the variables. These conditions are more stringent than necessary, but certain violations of these conditions can cause pooling to yield erroneous conclusions.

Nested and Crossed Effects

An effect A is *nested* within another effect B if each level of A occurs at only one level of B . For example, a station may contain one or more plant units, and every plant unit occurs at only one station. Therefore, plant units are nested within stations. The levels for A may have the same name for different levels of B , yet they are understood to be different. For example, Unit 1 at one station is a completely different plant from Unit 1 at a second station.

A and B are *crossed* if neither is nested within the other. For example, plants and years are typically crossed, with data from one year at many plants, and data from one plant for many years.

Although this report has not discussed nested effects until now, the modeling described earlier in this report allows crossed effects and nested effects. The only difference is that crossed effects are estimated based on marginal totals, while nested effects generally are estimated based on smaller subsets of the data. Therefore, nested effects more commonly correspond to zero counts, causing difficulty in the loglinear and logit models.

Treatment of Zero Counts

Zero counts cause trouble if they result in a cell mean whose maximum likelihood estimate is zero. Setting the estimated mean m_{ij} to zero in Equation (3) or setting the estimated π_{ij} to zero in Equation (4) makes the logarithm undefined (or $-\infty$). Then at least one of the effects, such as u_{20} , must be $-\infty$. Computer software will try to achieve such a value, but will eventually give up, and report that the estimate did not converge. For example, the SAS procedure GENMOD stops its iterations and prints a warning. It reports a finite negative value for the estimated effect, with an enormous standard error.

The fitted values, m_{ij} or π_{ij} , will then be virtually correct, some value on the order of 10^{-12} instead of exactly zero. Therefore, for some applications the above procedure is adequate. For other

applications, however, this situation is unsatisfactory. For example, the reported confidence intervals for the nearly zero fitted values will be unreasonably large. The analyst then has three options:

- Give up, until more data can be obtained.
- Work with only a portion of the data.
- Alter the data.

The second option was followed in the analysis of the RCIC/HPCI example, when three plants with no failures were dropped from the analysis. The third option is discussed here.

One way to alter the data is to add some small number to the offending data counts. For example, the macro GLIM presented with the SAS (1989) procedure NLIN adds 0.5 to every zero Poisson count, and adds 0.1 to every zero binomial count. It does this without comment.

A less ad hoc alteration is to add some data, equivalent to a Bayes prior, to every count. Consider the Poisson case first. If a total of n_{tot} Poisson events were observed in total time t_{tot} , we might wish to use a prior for each cell that has the overall mean but that is otherwise noninformative. As shown by Atwood (1994b), a reasonable constrained noninformative prior on the Poisson rate is a gamma distribution with mean $(n_{tot} + 1/2)/t_{tot}$ and shape parameter 1/2. This is equivalent to a prior observation of half an event in time $t_{tot}/(2n_{tot} + 1)$. This suggests adding 1/2 to each Poisson count (not just the zero counts), and increasing each exposure time by $t_{tot}/(2n_{tot} + 1)$. Similarly, the constrained noninformative prior for binomial data is approximately a beta distribution with parameters tabulated by Atwood (1994b). This suggests adding a small failure count (slightly less than 1/2, as given by the tabulation) to each failure count, and a corresponding number of demands to each demand count. Adding this "data" to the true data results in fractional event counts, but most software can handle fractional counts.

The reason for using a prior with mean based on the total data, rather than a Jeffreys noninformative prior or some other prior, is this: Use of any prior pulls the estimates toward the prior mean, and the pull can be quite strong if the data set is small. If the estimated cell means are to be pulled in some direction, it seems preferable to pull them toward the mean of the overall data than toward some other value.

The resulting estimates of the loglinear effects are not MLEs, because they maximize not the likelihood but the posterior distribution obtained by updating the constrained noninformative prior with the original counts. They are sometimes called *generalized maximum likelihood estimates*, for example by DeGroot (1970). The analyst may decide that their use is justifiable.

Table 7 shows the effect of this on a sample data set having seven cells, listed as seven rows. NF and ND are the numbers of failures and demands. Note that cell 1 had no failures. The remaining rows are arranged in order of increasing NF/ND. The data were analyzed with a logit model containing a cell effect u_i : $\text{logit}(\pi_i) = u + u_{i(0)}$. A portion of the output from GENMOD, in the upper left of the table, shows the fitted π_i and an asymptotic 90% confidence interval. For comparison, each cell was analyzed individually, and the MLE and an exact 90% confidence interval were found; these are in the upper right of the table. Note that the two methods both find the same point estimates, the MLEs. The confidence intervals do not agree exactly, because the logit model uses asymptotic normality. The two methods produce similar intervals for all the cells except cell 1.

Table 7. Four comparison analyses of a sample data set, with point estimates and 90% intervals.

Use raw data and logit model					Use raw data and cell-specific analysis		
Raw data		Observation Statistics			MLE and Confidence Int.		
NF	ND	Lower	Pred	Upper	Lower	MLE	Upper
0	143	0	1.44E-12	1.0000	0.0000	0.0000	0.0207
6	1184	0.0026	0.005068	0.0099	0.0022	0.0051	0.0100
1	119	0.0016	0.008403	0.0423	0.0004	0.0084	0.0392
1	93	0.0021	0.0108	0.0538	0.0006	0.0108	0.0500
4	162	0.0109	0.0247	0.0550	0.0085	0.0247	0.0556
2	81	0.0077	0.0247	0.0760	0.0044	0.0247	0.0757
5	195	0.0123	0.0256	0.0525	0.0102	0.0256	0.0532

Use altered data and logit model					Beta(0.4927, 49.48) prior updated with raw data		
Altered data		Observation Statistics			Bayes mean and interval		
NF	ND	Lower	Pred	Upper	Lower	Mean	Upper
0.4927	192.97	0.0002	0.00255	0.0260	0.0000	0.0026	0.0098
6.4927	1233.97	0.0028	0.00526	0.0100	0.0024	0.0053	0.0090
1.4927	168.97	0.0023	0.00883	0.0333	0.0010	0.0088	0.0230
1.4927	142.97	0.0027	0.0104	0.0392	0.0012	0.0104	0.0271
4.4927	211.97	0.0098	0.0212	0.0453	0.0079	0.0212	0.0396
2.4927	130.97	0.0067	0.0190	0.0526	0.0044	0.0190	0.0418
5.4927	244.97	0.0112	0.0224	0.0446	0.0094	0.0224	0.0398

The lower half of the table uses altered data. The data set had a total of 19 failures and 1977 demands. Therefore, the beta approximation to a constrained noninformative prior was found having mean $(19 + 1/2)/(1977 + 1)$. The distribution was $\text{beta}(0.4927, 49.48)$, and is equivalent to prior data of 0.4927 failures in 49.97 demands. These numbers were added to the raw failure and demand counts to produce an altered data set. Two comparable analyses were then performed. The results shown in the lower left of Table 7 were based on the logit model with the altered data. The results in the lower right were based on the posterior distribution from the constrained noninformative prior and the original data. The point estimates from these two analyses agree with each other. The intervals do not agree,

because one analysis uses asymptotic normality of $\text{logit}(\hat{\pi}_i)$, while the other uses a beta distribution for π_i . In this example, the intervals from the logit model are consistently to the right of the Bayes intervals.

Now let us compare the upper and lower portions of the table. The point estimates are similar, though not exactly equal. The intervals in the lower portion are shorter than those in the upper portion, reflecting the additional "data" used with the logit model, and the fact that a constrained noninformative prior is somewhat informative.

Random Effects

Textbooks and commercial software are largely agreed in treating all categorical effects as fixed effects, not random. Nonlinear models with random effects are described only in occasional articles in research journals. Although a summary treatment of random effects for loglinear and logit models has apparently not been written at an elementary or intermediate level, the idea is quite simple, an exact analogue of random effects for linear models. The level of a categorical variable is assumed to be drawn randomly from a population, and the corresponding effect therefore has a probability distribution. Thus, the effects all resemble each other to some extent, being generated from the same probability distribution. This situation differs from the fixed-effects setting, where the levels of a categorical variable have effects that are completely unrelated, except for the possible convention that the effects must sum to zero.

In a model with random effects, the initial estimation problem is to estimate the parameters of the effects' distribution, while at the same time estimating any fixed effects that may be present. This extends empirical Bayes modeling (described, for example by Atwood 1994a and Engelhardt 1994) to models with more than one explanatory variable. The second estimation problem is to estimate the effects for each observed level; this corresponds to estimating the posterior means in empirical Bayes modeling.

In the applications envisioned for this report, random effects can arise naturally. For example, NRC operational event data typically come from many plants. The plants are not the same, and are not expected to have identical failure probabilities. Nevertheless, the plants are similar, and so the failure probabilities should also be somewhat similar. To model the plant effects as coming from a single population is often reasonable. Thus, analysts should plan to begin using software that allows random effects, as such software becomes available.

SAS has produced a macro, GLIMMIX, to analyze a generalized linear model (of which loglinear and logit models are special cases), with both fixed and random effects. It requires SAS/STAT and SAS/IML release 6.08 or later. The method is documented by Wolfinger and O'Connell (1993).

The Numerical Algorithms Group (NAG) software contains GENSTAT, which allows for random effects in a generalized linear model. The basis for that program is documented by Engel and Keen (1994).

The program FRAC (Failure Rate Analysis Code), by Martz et al. (1982), considered both fixed and random effects using loglinear models. The original version of FRAC considered normal approximations to exact Poisson distributions. Later versions of FRAC considered exact Poisson distributions, but only for fixed effects. A task is presently underway at Los Alamos National Laboratory to develop loglinear models for both fixed and random effects, building on the methods of FRAC and of current software, and using the SAS package. These models can be used by the NRC to estimate component failure rates from Poisson data.

REVIEW

This section reviews the tools that have been presented in this report.

Two kinds of data have been considered, Poisson counts and binomial counts. The Poisson count is assumed to have mean $m = \lambda t$, where λ is an unknown parameter (a rate) and t is a known exposure time. The binomial count is assumed to have mean πd , where π is an unknown parameter (a probability) and d is a known number of demands or trials. In the Poisson case, the unknown rate is transformed to $\eta = \log(\lambda)$, and $\eta + \log(t)$ is assumed to be a function of explanatory variables and unknown parameters. In this case $\log(t)$ is called the offset. In the binomial case, the unknown probability is transformed to $\eta = \text{logit}(\pi) = \log[\pi/(1-\pi)]$, and η is assumed to be a function of explanatory variables and unknown parameters. In each case, the function of the unknown parameters is assumed to be linear in the parameters. The transforming function, log or logit in this paragraph, is sometimes called the link function.

The explanatory variables can be numerical, such as time, or unordered categorical, such as plant, system, or manufacturer. Unordered categorical variables can correspond to fixed effects or random effects, although the literature and the software of loglinear models emphasize fixed effects much more than random effects. The effects of unordered categorical variables are not uniquely defined, because a constant could be subtracted from all of the effect terms $u_{1(i)}$ and added to the overall constant u without changing the value of η . Two methods for dealing with this fact are to impose constraints, such as $\sum u_{1(i)} = 0$, or to eliminate some effects and work with a reduced set of effects. These methods are equally valid, and mathematically equivalent.

The standard methods assume

- The model is hierarchical, that is, that the presence of an interaction term implies the presence of all lower order effects.
- The effects of unordered categorical variables are fixed, not random.
- The sample size is large enough for the MLEs of the unknown effects to be approximately normally distributed.

Under these assumptions, the approach of loglinear modeling can fit models, obtaining point estimates and confidence intervals for the individual effects and for the means of the observed counts. The methodology has tools to test whether a postulated model contains unnecessary parameters (parameters that could plausibly equal zero, based on the evidence of the data). The methodology also has tools for testing whether the model fits the data adequately, or instead needs to be augmented. All

these tests can be of two types, those based on the likelihood ratio test and using the statistic $G^2(H_0 | H_1)$, and those based on the Pearson statistic $X^2(H_0 | H_1)$. Both test statistics have an asymptotic χ^2 distribution when H_0 is true, and are asymptotically equivalent to each other. G^2 has a useful property: the test statistics for a chain of nested hypotheses add up to the test statistic for testing the hypothesis at one end of the chain against that at the other. However, X^2 approaches its asymptotic χ^2 distribution faster, at least in cases where empirical studies have been performed. Some guidelines have been given for deciding if the sample size is large enough to use the asymptotic distribution.

Three example data sets were analyzed. These data analyses all used the tools of model building presented earlier, and hypothesis tests and analysis of residuals to decide which model to use.

- The RCIC/HPCI Example had binomial counts and two unordered categorical explanatory variables. Several levels of one variable (several plants) had to be dropped from the analysis because they had no observed failures. It is invalid to pool one of the explanatory variables and test for differences between levels of the other variable. Instead, one must form a model containing both variables simultaneously, and test whether the effect of either explanatory variable is zero. In the example, both explanatory variables had statistically significant effects, but the resulting model (without an interaction term) still did not fit the data well.
- The Unplanned-Demands Example had Poisson counts, with each count corresponding to an exposure time. The explanatory variable was numerical, calendar year. A simple model was found to fit quite well. The fitted Poisson rate was plotted against calendar year, showing the trend, and a confidence band was found for the fitted line, simultaneously valid for the entire line. The result was similar to results obtainable using linear models, but the confidence band was tighter than the confidence band produced by linear modeling. The loglinear model band was narrower because it required only one estimated parameter, the Poisson rate, while the band from the linear model required two, a mean and a variance. If the Poisson assumption is believed, if no extraneous sources of variance influenced the data, the loglinear analysis is preferable.
- The MOV-Test Example had binomial data and two explanatory variables. One explanatory variable, year, was numerical and one, plant, was unordered categorical. It was found that including a main effect for plant and a linear term for year fitted the data well. No interaction term was necessary.

Several topics were then discussed. A kind of diagram was introduced to help show the kinds of terms (main effects, trends, interactions) that should be included in a model. These diagrams also help show the causes of Simpson's paradox. Possible treatments were suggested for data having too many

zero counts. Finally, random-effects models were mentioned, and some current and planned software was described.

REFERENCES

- Agresti, Alan, 1990, *Categorical Data Analysis*, New York: Wiley-Interscience, John Wiley & Sons.
- Atwood, Corwin L., 1994a, *Hits per Trial: Basic Analysis of Binomial Data*, EGG-RAAM-11041.
- Atwood, Corwin L., 1994b, "Constrained Noninformative Priors in Risk Assessment," INEL-94/00029, submitted for publication.
- Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul W. Holland, 1975, *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press.
- Christensen, Ronald, 1990, *Log-Linear Models*, New York: Springer-Verlag.
- DeGroot, Morris H., 1970, *Optimal Statistical Decisions*, New York: McGraw-Hill, p. 236.
- Engel, B., and A. Keen, 1994, "A Simple Approach for the Analysis of Generalized Linear Mixed Models," *Statistica Neerlandica*, 48, pp. 1-22.
- Engelhardt, Max, 1994, *Events in Time: Basic Analysis of Poisson Data*, EGG-RAAM-11088.
- Engelhardt, Max, 1995, *Modeling Patterns in Data Using Linear and Related Models*, DRAFT, INEL-95/0120.
- Everitt, B. S., 1992, *The Analysis of Contingency Tables, Second Edition*, London: Chapman & Hall.
- Fienberg, Stephen E., 1980, *The Analysis of Cross-Classified Categorical Data, Second Edition*, Cambridge, MA: MIT Press.
- Grant, G. M., W. S. Roesener, D. G. Hall, C. L. Atwood, C. D. Gentillon, and T. R. Wolf, 1995, *High-Pressure Coolant Injection (HPCI) System Performance, 1987-1993*, INEL-94/0158, February 1995.
- Martz, H. F., R. J. Beckman, and C. R. McInteer, *FRAC (Failure Rate Analysis Code): A Computer Program for Analysis of Variance of Failure Rates*, NUREG/CR-2434, LA-9116-MS, March 1982.
- McCullagh, P., and J. A. Nelder, 1989, *Generalized Linear Models, Second Edition*, London: Chapman and Hall.
- Scheffé, Henry, 1959, *The Analysis of Variance*, New York: John Wiley & Sons.

Simola, Kaisa, 1992, *Probabilistic Methods in Nuclear Power Plant Component Ageing Analysis*, Technical Research Centre of Finland, VTT Publications 94, pp. 54-55.

SAS/STAT Software: Changes and Enhancements, Release 6.10, 1994, Cary NC: SAS Institute, Inc.

SAS/STAT Software: The GENMOD Procedure, Release 6.09, 1993, SAS Technical Report P-243, Cary NC: SAS Institute, Inc.

SAS/STAT User's Guide, Version 6, Fourth Edition, 1989, Cary NC: SAS Institute, Inc.

Wolfinger, Russ, and Michael O'Connell, 1993, "Generalized Linear Mixed Models: A Pseudo-Likelihood Approach," *Journal of Statistical Computation and Simulation*, 48, pp. 233-243.

Appendix A

SAS Programs and Full Outputs

Appendix A

SAS Programs and Full Outputs

This appendix contains the outline of a program to calculate $X^2(H_0 \mid H_1)$, SAS programs to perform the example analyses, and the full outputs from those programs.

A-1. PROGRAM OUTLINE FOR CALCULATING $X^2(H_0 \mid H_1)$

```

* set the data set;
data base; set ...; run;

proc genmod data=base;
* set model to H0, the null hypothesis;
  model ... / ...
          obstats;
  make 'obstats' out=obst0;
  make 'modfit' out=modfit0;
  run;
proc genmod data=base;
* same data set, now set model to H1, the alternative hypothesis;
  model ... / ...
          obstats;
  make 'obstats' out=obst1;
  make 'modfit' out=modfit1;
  run;

* make data sets with the degrees of freedom and with fitted values;
data modfit0; set modfit0; df0=df; keep df0; if _N_=1 then output; run;
data modfit1; set modfit1; df1=df; keep df1; if _N_=1 then output; run;
data modfit; merge modfit0 modfit1; run;

data obst0; set obst0; fit0=pred; keep fit0; run;
data obst1; set obst1; fit1=pred; keep fit1; run;

* Make a single data set with the two sets of fitted values;
data pearson;
  merge base obst0 obst1;
  if _n_=1 then set modfit;
  run;

/*
  The next lines are for Poisson counts;
  If the counts are binomial, add (number of demands) to the keep statement,
  and define fit0 = fit0*(number of demands), fit1=fit1*(number of demands);
*/
data pearson; set pearson nobs=nobs;
  keep fit0 fit1 xsqi xsq pvalue df;
  df = df0 - df1;
  xsqi = ( (fit1 - fit0)**2 ) / fit0 ;
* xsq is the cumulative sum of the contributions xsqi;
  xsq + xsqi;
  if _N_=nobs then pvalue = 1 - probchi(xsq, df) ;
  run;
proc print; run;

```

A-2. RCIC/HPCI EXAMPLE

SAS Program for Analysis of RCIC/HPCI Example, Dropping Three Plants

```

data tdpfid;
  input plant $ sys $ Nf Nd ;
  cards;
A    RCIC    1    24
B    RCIC    0    72
C    RCIC    0    24
D    RCIC    1    25
E    RCIC    3    26
F    RCIC    1    24
G    RCIC    0    73
H    RCIC    0    72
I    RCIC    2    74
J    RCIC    0    24
K    RCIC    0    74
L    RCIC    1    77
M    RCIC    0    74
N    RCIC    2    75
O    RCIC    1    24
P    RCIC    0    72
Q    RCIC    0    17
R    RCIC    0    26
S    RCIC    0    25
T    RCIC    0    28
U    RCIC    0    74
V    RCIC    0    25
W    RCIC    1    25
A    HPCI    2    24
B    HPCI    0    72
C    HPCI    0    25
D    HPCI    0    75
E    HPCI    3    26
F    HPCI    0    24
G    HPCI    2    73
H    HPCI    7    72
I    HPCI    2    72
J    HPCI    0    72
K    HPCI    1    74
L    HPCI    4    79
M    HPCI    2    79
N    HPCI    1    75
O    HPCI    6    74
P    HPCI    2    73
Q    HPCI    3    17
R    HPCI    3    24
S    HPCI    1    14
T    HPCI    1    33
U    HPCI    5    29
V    HPCI    2    25
W    HPCI    2    25
;
data tdpfidm; set tdpfid;
  where plant^='B' and plant^='C' and plant^='J';
  output;
run;

```

```

proc genmod data=tdpfidm order=data;
  class sys plant;
  model Nf/Nd = sys plant / dist = bin
                        link = logit
                        type1
                        type3
                        obstats
                        alpha = 0.1;

run;

```

Output from Above Program

The SAS System

12:21 Friday, March 3, 1995³

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.TDPFIDM
Distribution	BINOMIAL
Link Function	LOGIT
Dependent Variable	NF
Dependent Variable	ND
Observations Used	40
Number Of Events	62
Number Of Trials	1921

Class Level Information

Class	Levels	Values
SYS	2	RCIC HPCI
PLANT	20	A D E F G H I K L M N O P Q R S T U V W

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	19	33.5254	1.7645
Scaled Deviance	19	33.5254	1.7645
Pearson Chi-Square	19	36.3049	1.9108
Scaled Pearson X2	19	36.3049	1.9108
Log Likelihood	.	-246.2075	.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-2.2577	0.6063	13.8676	0.0002
SYS	RCIC	1	-1.3931	0.3234	18.5508	0.0001
SYS	HPCI	0	0.0000	0.0000	.	.
PLANT	A	1	0.0445	0.8520	0.0027	0.9584

PLANT	D	1	-2.1276	1.1732	3.2891	0.0697
PLANT	E	1	0.7367	0.7474	0.9714	0.3243
PLANT	F	1	-1.1134	1.1796	0.8910	0.3452
PLANT	G	1	-1.5426	0.9337	2.7294	0.0985
PLANT	H	1	-0.2269	0.7177	0.0999	0.7519
PLANT	I	1	-0.8219	0.7887	1.0862	0.2973
PLANT	K	1	-2.2589	1.1712	3.7201	0.0538
PLANT	L	1	-0.6748	0.7559	0.7970	0.3720
PLANT	M	1	-1.6100	0.9335	2.9746	0.0846
PLANT	N	1	-1.1554	0.8396	1.8937	0.1688
PLANT	O	1	-0.0920	0.7211	0.0163	0.8985
PLANT	P	1	-1.5398	0.9338	2.7193	0.0991
PLANT	Q	1	0.4272	0.8600	0.2468	0.6194
PLANT	R	1	0.0254	0.8515	0.0009	0.9762
PLANT	S	1	-0.7099	1.1842	0.3593	0.5489
PLANT	T	1	-1.4080	1.1767	1.4318	0.2315
PLANT	U	1	0.0851	0.7630	0.0124	0.9112
PLANT	V	1	-0.4342	0.9438	0.2116	0.6455
PLANT	W	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

LR Statistics For Type 1 Analysis

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	88.8373	0	.	.
SYS	67.8717	1	20.9657	0.0001
PLANT	33.5254	19	34.3462	0.0167

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
SYS	1	22.4346	0.0001
PLANT	19	34.3462	0.0167

Observation Statistics

NF	ND	Pred	Xbeta	Std	HessWgt	Lower
1	24	0.0264	-3.6063	0.6531	0.6176	0.009188
1	25	0.003084	-5.7784	1.0491	0.0769	0.000551
3	26	0.0515	-2.9141	0.5052	1.2691	0.0231
1	24	0.008457	-4.7642	1.0463	0.2013	0.001523
0	73	0.005523	-5.1934	0.7588	0.4009	0.001592
0	72	0.0203	-3.8777	0.4655	1.4305	0.009533
2	74	0.0113	-4.4727	0.5694	0.8258	0.004455
0	74	0.002706	-5.9097	1.0373	0.1997	0.000492
1	77	0.0131	-4.3256	0.5241	0.9920	0.005554
0	74	0.005165	-5.2608	0.7596	0.3802	0.001486
2	75	0.008112	-4.8062	0.6391	0.6035	0.002850
1	24	0.0231	-3.7428	0.4933	0.5425	0.0104
0	72	0.005538	-5.1906	0.7590	0.3965	0.001595
0	17	0.0383	-3.2236	0.6621	0.6260	0.0132
0	26	0.0259	-3.6254	0.6507	0.6572	0.009053
0	25	0.0126	-4.3607	1.0425	0.3113	0.002293
0	28	0.006313	-5.0588	1.0453	0.1756	0.001137

0	74	0.0275	-3.5657	0.4999	1.9791	0.0123
0	25	0.0165	-4.0850	0.7700	0.4068	0.004719
1	25	0.0253	-3.6508	0.6522	0.6168	0.008805
2	24	0.0986	-2.2133	0.6074	2.1324	0.0387
0	75	0.0123	-4.3853	1.0062	0.9115	0.002375
3	26	0.1793	-1.5210	0.4503	3.8261	0.0943
0	24	0.0332	-3.3712	1.0166	0.7706	0.006411
2	73	0.0219	-3.8003	0.7168	1.5619	0.006831
7	72	0.0769	-2.4846	0.3972	5.1137	0.0416
2	72	0.0440	-3.0796	0.5142	3.0256	0.0194
1	74	0.0108	-4.5166	1.0067	0.7911	0.002082
4	79	0.0506	-2.9325	0.4620	3.7929	0.0243
2	79	0.0205	-3.8677	0.7161	1.5847	0.006396
1	75	0.0319	-3.4131	0.5891	2.3153	0.0123
6	74	0.0871	-2.3497	0.3954	5.8834	0.0474
2	73	0.0219	-3.7975	0.7168	1.5661	0.006851
3	17	0.1382	-1.8305	0.6190	2.0245	0.0548
3	24	0.0969	-2.2323	0.6071	2.1000	0.0380
1	14	0.0489	-2.9676	1.0246	0.6513	0.009443
1	33	0.0249	-3.6658	1.0127	0.8027	0.004814
5	29	0.1022	-2.1726	0.4843	2.6618	0.0488
2	25	0.0635	-2.6919	0.7302	1.4857	0.0200
2	25	0.0947	-2.2577	0.6063	2.1430	0.0371

Observation Statistics

Upper	Resraw	Reschi	Resdev
0.0736	0.3656	0.4652	0.4298
0.0171	0.9229	3.3288	1.8205
0.1108	1.6620	1.4753	1.2788
0.0455	0.7970	1.7766	1.2737
0.0190	-0.4031	-0.6367	-0.8992
0.0426	-1.4601	-1.2208	-1.7176
0.0283	1.1647	1.2817	1.0871
0.0147	-0.2002	-0.4481	-0.6332
0.0304	-0.005072	-0.005092	-0.005096
0.0178	-0.3822	-0.6198	-0.8754
0.0229	1.3916	1.7913	1.4153
0.0506	0.4446	0.6037	0.5436
0.0190	-0.3987	-0.6332	-0.8943
0.1058	-0.6509	-0.8227	-1.1521
0.0721	-0.6747	-0.8322	-1.1692
0.0662	-0.3152	-0.5650	-0.7965
0.0342	-0.1768	-0.4218	-0.5955
0.0605	-2.0350	-1.4466	-2.0315
0.0563	-0.4136	-0.6485	-0.9133
0.0706	0.3672	0.4675	0.4317
0.2290	-0.3656	-0.2504	-0.2565
0.0612	-0.9229	-0.9666	-1.3628
0.3142	-1.6620	-0.8497	-0.8974
0.1546	-0.7970	-0.9080	-1.2732
0.0678	0.4031	0.3226	0.3105
0.1381	1.4601	0.6457	0.6222
0.0968	-1.1647	-0.6696	-0.7165
0.0541	0.2002	0.2251	0.2167
0.1022	0.005072	0.002604	0.002604
0.0636	0.3822	0.3036	0.2929
0.0799	-1.3916	-0.9145	-1.0323
0.1546	-0.4446	-0.1833	-0.1853
0.0680	0.3987	0.3186	0.3069
0.3074	0.6509	0.4575	0.4416
0.2256	0.6747	0.4656	0.4476
0.2172	0.3152	0.3906	0.3666

0.1192	0.1768	0.1973	0.1910
0.2016	2.0350	1.2473	1.1485
0.1838	0.4136	0.3394	0.3269
0.2209	-0.3672	-0.2508	-0.2570

A-3. MACROS FOR SIMULTANEOUS CONFIDENCE BANDS

These macros perform the calculations to obtain simultaneous confidence bands. They are used in the analyses of the Unplanned-Demands and MOV-Test Examples. Macro `llpconf` is used for Poisson data with a loglinear model, and `llbconf` is used for binomial data with a logit model. Both macros call the macro `solve`, also listed here. These three macros were stored as files named `llpconf.mac`, `llbconf.mac`, and `solve.mac`, and were referenced that way, with `%include` statements, in the calling programs.

```
*-----;
%macro llpconf(indata, obstdat, covdat, outdat, expos, x1, x2, x3);
/* Finds pointwise and simultaneous confidence intervals for Poisson rate,
   in Poisson regression loglinear model
   Inputs are
       indata = input to genmod
       obstdat = output dataset from make obstats, input to macro
       covdat = output dataset from make cov, input to macro
       outdat = output dataset from macro = indata
                                   + genmod's obstat calculations
                                   + confidence limits
   x1, x2, x3 = explanatory variables in loglinear model
       If not all are used, set last ones to 0 or dummy.
   If xi corresponds to a classification variable, it must be converted
       to a numerical value (1 or 0) before the call to llpconf! For example,
       if PLANT is a variable, you must set x1=1 iff PLANT='Plant1',
                                   x2=1 iff PLANT='Plant2',
                                   etc., for all but the final plant.
       If there are more levels, modify the macro by adding more arguments in
       the obvious way.
   expos = exposure time for Poisson counts, a variable in indata.
       If indata does not have this, set it to 1 in the call.
   WARNING! This macro creates files with names beginning with xxx. Do not
       use files with such names in the calling program, or they may be
       overwritten.
*/
data xxxcombi; merge &indata &obstdat; run;

%global dim;
%solve(&covdat, xxxcombi, &outdat, nom, &x1, &x2, &x3);

data &outdat; set &outdat;
* drop nom ;
  fitrate = pred/&expos;
  fitlow = lower/&expos;
  fitup = upper/&expos;
  nom = exp( sqrt( nom * cinv(0.90, &dim) ) ) ;
  fitslow = fitrate / nom;
  fitsup = fitrate * nom;
run;

proc print; run;
%mend llpconf;
*-----;
```

```

*-----;
%macro llbconf(indata, obstdat, covdat, outdat, x1, x2, x3);
/* Finds pointwise and simultaneous confidence intervals for binomial prob.,
   in logistic regression model
   Inputs are
       indata = input to genmod
       obstdat = output dataset from make obstats, input to macro
       covdat  = output dataset from make cov, input to macro
       outdat  = output dataset from macro = indata
                           + genmod's obstat calculations
                           + confidence limits
       x1, x2, x3 = explanatory variables in loglinear model
       If not all are used, set last ones to 0 or dummy.
   If xi corresponds to a classification variable, it must be converted
   to a numerical value (1 or 0) before the call to llbconf! For example,
   if PLANT is a variable, you must set x1=1 iff PLANT='Plant1',
                                   x2=1 iff PLANT='Plant2',
                                   etc., for all but the final plant.
       If there are more levels, modify the macro by adding more arguments in
       the obvious way.
   WARNING! This macro creates files with names beginning with xxx. Do not
   use files with such names in the calling program, or they may be
   overwritten.
*/
data xxxcombi; merge &indata &obstdat; run;

%global dim;
%solve(&covdat, xxxcombi, &outdat, nom, &x1, &x2, &x3);

data &outdat; set &outdat;
* drop nom plusmin eetalow eetaup ;
  fitp   = pred;
  fitlow = lower;
  fitup  = upper;
  plusmin = sqrt( nom * cinv(0.90, &dim) );
  eetalow = exp(Xbeta - plusmin);
  eetaup  = exp(Xbeta + plusmin);
  fitslow = eetalow/(1 + eetalow);
  fitsup  = eetaup/(1 + eetaup);
run;

proc print; run;
%mend llbconf;
*-----;

*-----;
%macro solve(arraydat, indata, outdata, prod, x2, x3, x4);
/*
   Inputs arraydat, the SAS form of the covariance matrix
       indata, the "design matrix", with up to 3 columns (x2, x3, and x4)
           the numbering assumes an initial column of 1's
           For example, x2 might be age, x3 size, x4 ...
   Outputs prod = (1, x2, x3, ...) cov-matrix (1, x2, x3, ...)transpose,
       a single number
   This is put into outdata, which contains the original indata plus prod.
   If the input uses fewer than 3 variables, input dummy, or 0, for the
   last arguments.
*/
data _null_; set &arraydat nobs=dimen;
  call symput('dim',dimen); stop; run;

data xxxcalc;
  set &arraydat(in=covdat) &indata(in=rawdat);
  array cov(&dim,&dim);

```

```

retain cov;
keep &prod;

*read in covariance matrix;
if covdat then do; *if inputting cov data;
  %do i=1 %to &dim;
    cov[_n_,&i]=prm&i;
  %end;
end;

*calculate values from data;
if rawdat then do; *if inputting raw (design) data;
  if &dim >= 1 then &prod= cov[1,1];
  if &dim >= 2 then
    &prod=&prod + 2*&x2*cov[1,2] + &x2*&x2*cov[2,2];
  if &dim >= 3 then
    &prod=&prod + 2*&x3*cov[1,3] + 2*&x2*&x3*cov[2,3] + &x3*&x3*cov[3,3];
  if &dim >= 4 then
    &prod=&prod + 2*&x4*cov[1,4] + 2*&x2*&x4*cov[2,4] +
      2*&x3*&x4*cov[3,4] + &x4*&x4*cov[4,4];
  output;
end;
run;

data &outdata; merge &indata xxxcalc;
run;

%mend solve;
*-----;

```

A-4. UNPLANNED-DEMANDS EXAMPLE

SAS Program for Analysis of the Unplanned-Demands Example

```
option mprint;
%include 'llpconf.mac';
%include 'solve.mac';

libname lib 'k:\nrc\8205\loglin';

title 'Unplanned demands for HPCI';
data lib.hpcidems;
input year udems opyrs;
lopyrs = log(opyrs);
cards;
1 16 14.98
2 10 14.27
3 7 15.89
4 13 18.16
5 9 17.76
6 6 17.63
7 2 17.91
;
run;

proc print data=lib.hpcidems; run;

title2 'Model with year effect';
proc genmod data=lib.hpcidems;
model udems = year/ dist=poisson
      link=log
      offset=lopyrs
      a = 0.10
      LRCI
      covb
      corrb
      obstats
      type1;
make 'obstats' out=obstdat;
make 'cov' out=covdat;
run;

%llpconf(lib.hpcidems, obstdat, covdat, outdat, opyrs, year, 0, 0);
```

Output from Above Program

Unplanned demands for HPCI

17
11:43 Wednesday, March 8, 1995

OBS	YEAR	UDEMS	OPYRS	LOPYRS
1	1	16	14.98	2.70672
2	2	10	14.27	2.65816
3	3	7	15.89	2.76569
4	4	13	18.16	2.89922
5	5	9	17.76	2.87695
6	6	6	17.63	2.86960
7	7	2	17.91	2.88536

Unplanned demands for HPCI
Model with year effect

18

11:43 Wednesday, March 8, 1995

The GENMOD Procedure

Model Information

Description	Value
Data Set	LIB.HPCIDEMS
Distribution	POISSON
Link Function	LOG
Dependent Variable	UDEMS
Offset Variable	LOPYRS
Observations Used	7

Parameter Information

Parameter	Effect
PRM1	INTERCEPT
PRM2	YEAR

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	5	4.9803	0.9961
Scaled Deviance	5	4.9803	0.9961
Pearson Chi-Square	5	4.7619	0.9524
Scaled Pearson X2	5	4.7619	0.9524
Log Likelihood	.	80.7747	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	0.2623	0.2494	1.1061	0.2929
YEAR	1	-0.2381	0.0665	12.8291	0.0003
SCALE	0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

Estimated Covariance Matrix

Parameter Number	PRM1	PRM2
PRM1	0.06220	-0.01431
PRM2	-0.01431	0.004418

Unplanned demands for HPCI
Model with year effect

19

11:43 Wednesday, March 8, 1995

Estimated Correlation Matrix

Parameter Number	PRM1	PRM2
PRM1	1.0000	-0.8630
PRM2	-0.8630	1.0000

Likelihood Ratio Based Confidence Intervals For Parameters

Two-Sided Confidence Coefficient: 0.9000

Parameter	Confidence Limits	Parameter Values		
		PRM1		PRM2
PRM1	Lower	-0.1622	-0.1622	-0.1438
PRM1	Upper	0.6595	0.6595	-0.3329
PRM2	Lower	-0.3494	0.6013	-0.3494
PRM2	Upper	-0.1302	-0.1084	-0.1302

LR Statistics For Type 1 Analysis

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	18.4579	0	.	.
YEAR	4.9803	1	13.4775	0.0002

Observation Statistics

UDEMS	Pred	Xbeta	Std	HessWgt	Lower	Upper
16	15.3471	0.0242	0.1950	15.3471	11.1369	21.1490
10	11.5223	-0.2139	0.1505	11.5223	8.9958	14.7584
7	10.1120	-0.4520	0.1270	10.1120	8.2060	12.4608
13	9.1082	-0.6900	0.1358	9.1082	7.2850	11.3876
9	7.0203	-0.9281	0.1720	7.0203	5.2903	9.3162
6	5.4925	-1.1662	0.2227	5.4925	3.8081	7.9218
2	4.3976	-1.4043	0.2800	4.3976	2.7745	6.9700

Observation Statistics

Resraw	Reschi	Resdev
0.6529	0.1667	0.1655
-1.5223	-0.4485	-0.4589
-3.1120	-0.9786	-1.0367
3.8918	1.2895	1.2110
1.9797	0.7472	0.7156
0.5075	0.2166	0.2133
-2.3976	-1.1433	-1.2820

Unplanned demands for HPCI
Model with year effect

20

11:43 Wednesday, March 8, 1995

										H		
										E		
										S		
										L		
										U		
Y	U	O	L	Y	X					S	L	U
O	D	P	O	V	B					S	O	P
E	E	Y	P	A	E					W	W	P
B	A	R	R	R	T					G	E	P
S	R	S	S	1	D					T	R	R
1	1	16	14.98	2.70672	16	15.3471	0.02421	0.19495	15.3471	11.1369	21.1490	
2	2	10	14.27	2.65816	10	11.5223	-0.21388	0.15049	11.5223	8.9958	14.7584	
3	3	7	15.89	2.76569	7	10.1120	-0.45196	0.12698	10.1120	8.2060	12.4608	
4	4	13	18.16	2.89922	13	9.1082	-0.69005	0.13579	9.1082	7.2850	11.3876	
5	5	9	17.76	2.87695	9	7.0203	-0.92814	0.17202	7.0203	5.2903	9.3162	
6	6	6	17.63	2.86960	6	5.4925	-1.16622	0.22266	5.4925	3.8081	7.9218	
7	7	2	17.91	2.88536	2	4.3976	-1.40431	0.28001	4.3976	2.7745	6.9700	

O B S	R E S R A W	R E S C H I	R E S D E V	N O M	F I T R A T E	F I T L O W	F I T U P	F I T S L O W	F I T S U P
1	0.65289	0.16666	0.16550	1.51946	1.02451	0.74345	1.41181	0.67426	1.55670
2	-1.52229	-0.44847	-0.45893	1.38118	0.80745	0.63040	1.03422	0.58461	1.11523
3	-3.11205	-0.97865	-1.03665	1.31323	0.63638	0.51643	0.78419	0.48459	0.83571
4	3.89183	1.28955	1.21101	1.33829	0.50155	0.40116	0.62707	0.37477	0.67122
5	1.97965	0.74715	0.71562	1.44649	0.39529	0.29788	0.52456	0.27328	0.57178
6	0.50753	0.21656	0.21334	1.61256	0.31154	0.21600	0.44934	0.19320	0.50238
7	-2.39756	-1.14331	-1.28199	1.82373	0.24554	0.15492	0.38917	0.13463	0.44779

A-5. MOV-TEST EXAMPLE

SAS Program for Analysis of the MOV-Test Example

```
option mprint;
%include 'llbconf.mac';
%include 'solve.mac';

title 'Leaks of MOVs in 2 plants over 9 years';
data valves;
  input plant $ year Nf ;
  Nd = 52;
  cards;
Plant1 1 4
Plant1 2 2
Plant1 3 3
Plant1 4 1
Plant1 5 4
Plant1 6 3
Plant1 7 4
Plant1 8 9
Plant1 9 6
Plant2 1 1
Plant2 2 1
Plant2 3 1
Plant2 4 1
Plant2 5 2
Plant2 6 3
Plant2 7 1
Plant2 8 4
Plant2 9 3
;
title2 'Entering year, plant, and year*plant, in that order';
proc genmod data=valves order=data;
  class plant;
  model Nf/Nd = year plant year*plant / dist = bin
                                         link = logit
                                         type1
                                         type3
                                         ;
run;
title2 'Same factors, entering plant before year';
proc genmod data=valves order=data;
  class plant;
  model Nf/Nd = plant year year*plant / dist = bin
                                         link = logit
                                         type1
                                         ;
run;
title2 'Final model, with trend in year, plant main effect, no interaction';
proc genmod data=valves order=data;
  class plant;
  model Nf/Nd = year plant / dist = bin
                             link = logit
                             type1
                             type3
                             alpha = 0.10
                             covb
                             corrb
                             obstats
                             ;
  make 'obstats' out=obstdat;
  make 'cov' out=covdat;
run;
```

```

data valves; set valves;
  if (plant = 'Plant1') then pldum=1;
  else pldum=0;
run;
%llbconf(valves, obstdat, covdat, outdat, year, pldum, 0);

```

Output from Above Program

Leaks of MOVs in 2 plants over 9 years 9
 Entering year, plant, and year*plant, in that order
 11:43 Wednesday, March 8, 1995

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.VALVES
Distribution	BINOMIAL
Link Function	LOGIT
Dependent Variable	NF
Dependent Variable	ND
Observations Used	18
Number Of Events	53
Number Of Trials	936

Class Level Information

Class	Levels	Values
PLANT	2	Plant1 Plant2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	14	8.6776	0.6198
Scaled Deviance	14	8.6776	0.6198
Pearson Chi-Square	14	8.4932	0.6067
Scaled Pearson X2	14	8.4932	0.6067
Log Likelihood	.	-195.7833	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-4.2846	0.6668	41.2861	0.0001
YEAR	1	0.1813	0.1019	3.1700	0.0750
PLANT Plant1	1	0.9678	0.8019	1.4564	0.2275
PLANT Plant2	0	0.0000	0.0000	.	.
YEAR*PLANT Plant1	1	-0.0280	0.1237	0.0514	0.8207
YEAR*PLANT Plant2	0	0.0000	0.0000	.	.
SCALE	0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

LR Statistics For Type 1 Analysis

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	24.4120	0	.	.
YEAR	16.1686	1	8.2434	0.0041
PLANT	8.7292	1	7.4394	0.0064
YEAR*PLANT	8.6776	1	0.0516	0.8203

Leaks of MOVs in 2 plants over 9 years 10
 Entering year, plant, and year*plant, in that order
 11:43 Wednesday, March 8, 1995

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
YEAR	1	7.7670	0.0053
PLANT	1	1.5404	0.2146
YEAR*PLANT	1	0.0516	0.8203
Leaks of MOVs in 2 plants over 9 years			11
Same factors, entering plant before year			
			11:43 Wednesday, March 8, 1995

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.VALVES
Distribution	BINOMIAL
Link Function	LOGIT
Dependent Variable	NF
Dependent Variable	ND
Observations Used	18
Number Of Events	53
Number Of Trials	936

Class Level Information

Class	Levels	Values
PLANT	2	Plant1 Plant2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	14	8.6776	0.6198
Scaled Deviance	14	8.6776	0.6198
Pearson Chi-Square	14	8.4932	0.6067
Scaled Pearson X2	14	8.4932	0.6067
Log Likelihood	.	-195.7833	.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-4.2846	0.6668	41.2861	0.0001
PLANT	Plant1	1	0.9678	0.8019	1.4564	0.2275
PLANT	Plant2	0	0.0000	0.0000	.	.
YEAR		1	0.1813	0.1019	3.1700	0.0750
YEAR*PLANT	Plant1	1	-0.0280	0.1237	0.0514	0.8207
YEAR*PLANT	Plant2	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

LR Statistics For Type 1 Analysis

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	24.4120	0	.	.
PLANT	17.0378	1	7.3741	0.0066
YEAR	8.7292	1	8.3086	0.0039
YEAR*PLANT	8.6776	1	0.0516	0.8203

Leaks of MOVs in 2 plants over 9 years

12

Final model, with trend in year, plant main effect, no interaction

11:43 Wednesday, March 8, 1995

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.VALVES
Distribution	BINOMIAL
Link Function	LOGIT
Dependent Variable	NF
Dependent Variable	ND
Observations Used	18
Number Of Events	53
Number Of Trials	936

Class Level Information

Class	Levels	Values
PLANT	2	Plant1 Plant2

Parameter Information

Parameter	Effect	PLANT
PRM1	INTERCEPT	
PRM2	YEAR	
PRM3	PLANT	Plant1
PRM4	PLANT	Plant2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	15	8.7292	0.5819
Scaled Deviance	15	8.7292	0.5819
Pearson Chi-Square	15	8.6372	0.5758
Scaled Pearson X2	15	8.6372	0.5758
Log Likelihood	.	-195.8091	.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-4.1707	0.4247	96.4394	0.0001
YEAR		1	0.1624	0.0578	7.9112	0.0049
PLANT	Plant1	1	0.8003	0.3031	6.9700	0.0083
PLANT	Plant2	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

Leaks of MOVs in 2 plants over 9 years 13
 Final model, with trend in year, plant main effect, no interaction
 11:43 Wednesday, March 8, 1995

Estimated Covariance Matrix

Parameter Number	PRM1	PRM2	PRM3
PRM1	0.18037	-0.01992	-0.06305
PRM2	-0.01992	0.003335	0.0002758
PRM3	-0.06305	0.0002758	0.09189

Estimated Correlation Matrix

Parameter Number	PRM1	PRM2	PRM3
PRM1	1.0000	-0.8121	-0.4898
PRM2	-0.8121	1.0000	0.0158
PRM3	-0.4898	0.0158	1.0000

LR Statistics For Type 1 Analysis

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	24.4120	0	.	.
YEAR	16.1686	1	8.2434	0.0041
PLANT	8.7292	1	7.4394	0.0064

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
YEAR	1	8.3086	0.0039
PLANT	1	7.4394	0.0064

Observation Statistics

NF	ND	Pred	Xbeta	Std	HessWgt	Lower
4	52	0.0389	-3.2080	0.3320	1.9425	0.0229
2	52	0.0454	-3.0455	0.2845	2.2541	0.0289
3	52	0.0530	-2.8831	0.2415	2.6097	0.0363
1	52	0.0618	-2.7207	0.2058	3.0134	0.0448
4	52	0.0719	-2.5582	0.1819	3.4689	0.0543
3	52	0.0835	-2.3958	0.1746	3.9792	0.0640
4	52	0.0968	-2.2334	0.1859	4.5461	0.0732
9	52	0.1120	-2.0709	0.2129	5.1699	0.0816
6	52	0.1292	-1.9085	0.2504	5.8485	0.0894
1	52	0.0178	-4.0082	0.3793	0.9112	0.009640
1	52	0.0209	-3.8458	0.3377	1.0652	0.0121
1	52	0.0245	-3.6834	0.3014	1.2439	0.0151
1	52	0.0287	-3.5209	0.2727	1.4507	0.0185
2	52	0.0336	-3.3585	0.2541	1.6894	0.0224
3	52	0.0393	-3.1961	0.2478	1.9640	0.0265
1	52	0.0459	-3.0336	0.2548	2.2786	0.0307
4	52	0.0536	-2.8712	0.2741	2.6376	0.0348
3	52	0.0625	-2.7088	0.3033	3.0450	0.0389

Observation Statistics

Upper	Resraw	Reschi	Resdev
0.0653	1.9789	1.4199	1.2581
0.0706	-0.3613	-0.2407	-0.2469
0.0769	0.2442	0.1512	0.1492
0.0845	-2.2118	-1.2741	-1.4794
0.0946	0.2625	0.1409	0.1394
0.1083	-1.3417	-0.6726	-0.7091
0.1270	-1.0333	-0.4846	-0.5009
0.1518	3.1783	1.3978	1.3072
0.1829	-0.7159	-0.2960	-0.3007
0.0328	0.0723	0.0757	0.0748
0.0359	-0.0879	-0.0852	-0.0864
0.0396	-0.2751	-0.2467	-0.2562
0.0443	-0.4936	-0.4098	-0.4354
0.0502	0.2519	0.1938	0.1895
0.0580	0.9557	0.6819	0.6390
0.0682	-1.3883	-0.9197	-1.0363
0.0816	1.2130	0.7469	0.7034
0.0989	-0.2478	-0.1420	-0.1438

Leaks of MOVs in 2 plants over 9 years 15
Final model, with trend in year, plant main effect, no interaction
11:43 Wednesday, March 8, 1995

OBS	PLANT	YEAR	NF	ND	PLDUM	YVAR1	YVAR2	PRED	XBETA	STD	HESSWGT
1	Plant1	1	4	52	1	4	52	0.03887	-3.20798	0.33196	1.94252
2	Plant1	2	2	52	1	2	52	0.04541	-3.04554	0.28446	2.25411
3	Plant1	3	3	52	1	3	52	0.05300	-2.88310	0.24147	2.60971
4	Plant1	4	1	52	1	1	52	0.06176	-2.72067	0.20583	3.01340
5	Plant1	5	4	52	1	4	52	0.07188	-2.55823	0.18193	3.46889
6	Plant1	6	3	52	1	3	52	0.08349	-2.39579	0.17464	3.97918
7	Plant1	7	4	52	1	4	52	0.09679	-2.23336	0.18593	4.54612
8	Plant1	8	9	52	1	9	52	0.11196	-2.07092	0.21286	5.16991
9	Plant1	9	6	52	1	6	52	0.12915	-1.90849	0.25045	5.84850
10	Plant2	1	1	52	0	1	52	0.01784	-4.00825	0.37930	0.91119
11	Plant2	2	1	52	0	1	52	0.02092	-3.84581	0.33769	1.06518
12	Plant2	3	1	52	0	1	52	0.02452	-3.68337	0.30145	1.24385

13 Plant2	4	1 52	0	1	52	0.02872	-3.52094	0.27273	1.45066
14 Plant2	5	2 52	0	2	52	0.03362	-3.35850	0.25408	1.68936
15 Plant2	6	3 52	0	3	52	0.03931	-3.19607	0.24780	1.96396
16 Plant2	7	1 52	0	1	52	0.04593	-3.03363	0.25481	2.27864
17 Plant2	8	4 52	0	4	52	0.05360	-2.87119	0.27407	2.63762
18 Plant2	9	3 52	0	3	52	0.06246	-2.70876	0.30328	3.04499

OBS	LOWER	UPPER	RESRAW	RESCHI	RESDEV	NOM	FITP	FITLOW
1	0.022888	0.06526	1.97893	1.41987	1.25813	0.11020	0.03887	0.022888
2	0.028932	0.07059	-0.36134	-0.24068	-0.24693	0.08092	0.04541	0.028932
3	0.036254	0.07685	0.24425	0.15120	0.14916	0.05831	0.05300	0.036254
4	0.044821	0.08455	-2.21177	-1.27412	-1.47942	0.04237	0.06176	0.044821
5	0.054296	0.09458	0.26247	0.14093	0.13945	0.03310	0.07188	0.054296
6	0.063981	0.10827	-1.34169	-0.67260	-0.70915	0.03050	0.08349	0.063981
7	0.073156	0.12702	-1.03332	-0.48464	-0.50093	0.03457	0.09679	0.073156
8	0.081581	0.15177	3.17832	1.39784	1.30719	0.04531	0.11196	0.081581
9	0.089445	0.18294	-0.71586	-0.29601	-0.30071	0.06272	0.12915	0.089445
10	0.009640	0.03279	0.07226	0.07570	0.07477	0.14387	0.01784	0.009640
11	0.012113	0.03590	-0.08794	-0.08521	-0.08637	0.11403	0.02092	0.012113
12	0.015080	0.03964	-0.27512	-0.24668	-0.25622	0.09087	0.02452	0.015080
13	0.018532	0.04426	-0.49356	-0.40979	-0.43543	0.07438	0.02872	0.018532
14	0.022391	0.05018	0.25187	0.19378	0.18954	0.06456	0.03362	0.022391
15	0.026502	0.05795	0.95567	0.68193	0.63899	0.06141	0.03931	0.026502
16	0.030687	0.06821	-1.38833	-0.91972	-1.03633	0.06493	0.04593	0.030687
17	0.034824	0.08163	1.21300	0.74689	0.70341	0.07512	0.05360	0.034824
18	0.038881	0.09886	-0.24785	-0.14203	-0.14378	0.09198	0.06246	0.038881

OBS	FITUP	PLUSMIN	EETALOW	EETAUP	FITSLOW	FITSUP
1	0.06526	0.83000	0.017633	0.09274	0.017328	0.08487
2	0.07059	0.71123	0.023359	0.09688	0.022826	0.08832
3	0.07685	0.60374	0.030597	0.10235	0.029689	0.09285
4	0.08455	0.51464	0.039348	0.11014	0.037858	0.09921
5	0.09458	0.45487	0.049139	0.12205	0.046838	0.10877
6	0.10827	0.43664	0.058869	0.14098	0.055596	0.12356
7	0.12702	0.46487	0.067324	0.17059	0.063078	0.14573
8	0.15177	0.53222	0.074041	0.21466	0.068937	0.17672
9	0.18294	0.62618	0.079288	0.27740	0.073463	0.21716
10	0.03279	0.94835	0.007037	0.04689	0.006988	0.04479
11	0.03590	0.84431	0.009186	0.04971	0.009102	0.04736
12	0.03964	0.75370	0.011830	0.05341	0.011692	0.05071
13	0.04426	0.68189	0.014953	0.05848	0.014733	0.05525
14	0.05018	0.63528	0.018430	0.06566	0.018096	0.06162
15	0.05795	0.61958	0.022023	0.07604	0.021549	0.07067
16	0.06821	0.63709	0.025458	0.09103	0.024826	0.08344
17	0.08163	0.68526	0.028540	0.11237	0.027748	0.10102
18	0.09886	0.75828	0.031209	0.14221	0.030265	0.12450

Appendix B

Details and Extensions

Appendix B

Details and Extensions

The first two sections present a general context for loglinear and logit models, using the language of generalized linear models. The remaining sections give extensions and details for various specific topics mentioned in the body of this report.

B-1. EXPONENTIAL FAMILY

Sections B-1 and B-2 depart from the notation used elsewhere in this report. In the rest of this report, the random variable is N and its mean is m , as is standard in the literature of categorical data analysis. In Sections B-1 and B-2, the random variable is Y , with mean μ , as is standard in the literature of linear and generalized linear models.

B-1.1 Definition

The one-parameter exponential family of distributions, not to be confused with the exponential distribution, consists of distributions whose density or probability mass function have the form

$$f(y; \theta) = \exp(y\theta) \cdot \exp[-b(\theta)] \cdot c(y) . \quad (\text{B-1})$$

Note that the first term involves y and the parameter θ in an especially simple way. The second term does not involve y , and the final term does not involve θ . The parameterization that yields this form, with the product $y\theta$ in the first exponent, is the *natural*, or *canonical*, parameterization. This form is quite general, and includes one-parameter distributions such as the Poisson, the binomial, and the normal distribution with known variance.

To account for possible second parameters, or possible known variation such as different numbers of trials for binomial random variables or different exposure times for Poisson random variables, we follow McCullagh and Nelder (1989) and consider a generalization of this form, involving a second parameter ϕ :

$$f(y; \theta, \phi) = \exp[y\theta/a(\phi)] \cdot \exp[-b(\theta)/a(\phi)] \cdot c(y, \phi) . \quad (\text{B-2})$$

This form is also quite general. As will be seen below, the function $b(\theta)$ determines the mean of the distribution, while the parameter ϕ is typically related to the dispersion or shape of the distribution, and

is involved in the variance of θ . If ϕ is assumed known, it can be absorbed into the functions a and c , and the distribution belongs to the one-parameter exponential family with parameter θ .

The exponential family is discussed in most books on mathematical statistics. A summary of the selected important facts is given here.

B-1.2 Examples

B-1.2.1 Normal(μ, σ^2) Distribution

The normal(μ, σ^2) density can be rearranged into the form of Equation (B-2) as follows:

$$\begin{aligned} f(y) &= \sigma(2\pi)^{-1/2} \exp[-(y - \mu)^2/(2\sigma^2)] \\ &= \sigma(2\pi)^{-1/2} \exp[-y^2/(2\sigma^2)] \exp[y\mu/\sigma^2] \exp[-\mu^2/(2\sigma^2)] \\ &= \exp[y\mu/\sigma^2] \cdot \exp[-\mu^2/(2\sigma^2)] \cdot \sigma(2\pi)^{-1/2} \exp[-y^2/(2\sigma^2)] \\ &= \exp[y\mu/a(\sigma^2)] \cdot \exp[-b(\mu)/a(\sigma^2)] \cdot c(y, \sigma^2) \end{aligned}$$

which is of the form (B-2), with $\theta = \mu$, $\phi = \sigma^2$, $a(\sigma^2) = \sigma^2$, $b(\mu) = \mu^2/2$, and c an expression that does not depend on μ .

B-1.2.2 Poisson(μ) Distribution

The Poisson probability function is a one-parameter function. It can be written as

$$P(Y = y) = e^{-\mu} \mu^y / y! = \exp(y \log \mu) \cdot \exp(-\mu) \cdot (1/y!)$$

This is of the form (B-1), or (B-2) with $\phi = a(\phi) = 1$, $\theta = \log \mu$, $b(\theta) = \exp(\theta) = \mu$, and c a function that does not depend on $\log \mu$. Note that the canonical parameter is $\log \mu$, not μ .

If the Poisson mean μ equals λt , for some unknown rate λ and known exposure time t , let us denote the observable random variable by Yt , so that Y is the Poisson random variable divided by t . Then we have:

$$\begin{aligned}
P(Y = y) &= e^{-\lambda} (\lambda)^y / (y!) \\
&= \exp[y \log(\lambda)] \cdot \exp(-\lambda) \cdot [1/(y!)] \\
&= \exp(y \log \lambda) \cdot \exp(-\lambda) \cdot [\exp(y \log t) / (y!)]
\end{aligned}$$

which is of the form (B-2) with $\phi = 1$, $a(\phi) = 1/t$, $\theta = \log \lambda$, $b(\theta) = \lambda = \exp(\theta)$, and c a function that does not depend on $\log \lambda$.

B-1.2.3 Binomial(n, π) Distribution

To write the binomial(n, π) probability function in the form (B-2), denote the observable random variable by nY , so that Y is the binomial random variable divided by n :

$$\begin{aligned}
P(Y=y) &= \binom{n}{ny} \pi^{ny} (1-\pi)^{n-ny} \\
&= \binom{n}{ny} (1-\pi)^n [\pi/(1-\pi)]^{ny} \\
&= \exp\{ny \log[\pi/(1-\pi)]\} \cdot \exp[n \log(1-\pi)] \cdot \binom{n}{ny} .
\end{aligned}$$

This is of the form (B-2), with $\phi = 1$, $a(\phi) = 1/n$, $\theta = \log[\pi/(1-\pi)]$, and c a function that does not depend on θ . As for b , first invert the equation relating θ and π , to obtain $\pi = e^\theta / (1 + e^\theta)$. Then substitute this into $b(\pi) = -\log(1-\pi)$, to obtain $b(\theta) = \log(1 + e^\theta)$.

Note that the canonical parameter is $\log[\pi/(1-\pi)]$, the logit function of π .

B-1.3 Properties

The moments of Y are directly determined by the functions a and b , as follows. Although not needed for this report, the formulas are given here to complete the discussion. First, write the log-likelihood $\log f(y; \theta, \phi)$ as $l(\theta, \phi; y)$. For any density satisfying standard regularity conditions, we have

$$\begin{aligned}
E\left(\frac{\partial l}{\partial \theta}\right) &= \int \left(\frac{\partial \log f(y; \theta, \phi)}{\partial \theta}\right) f(y) dy \\
&= \int \left(\frac{\partial f(y; \theta, \phi)}{\partial \theta}\right) dy \\
&= \left(\frac{\partial}{\partial \theta}\right) \int f(y; \theta, \phi) dy \\
&= \left(\frac{\partial}{\partial \theta}\right) 1 = 0
\end{aligned}$$

with the integral replaced by a sum when the distribution is discrete. The third line is the crucial step, and uses the regularity conditions to guarantee that the order of differentiation and integration can be reversed. Similarly one can show

$$E\left(\frac{\partial^2 l}{\partial \theta^2}\right) = -E\left(\frac{\partial l}{\partial \theta}\right)^2.$$

From Equation (B-2) we have $l(\theta; y) = y\theta/a(\phi) - b(\theta)/a(\phi) + \log c(y, \phi)$. Therefore, the derivatives are

$$\frac{\partial l}{\partial \theta} = [y - b'(\theta)]/a(\phi)$$

and

$$\frac{\partial^2 l}{\partial \theta^2} = -b''(\theta)/a(\phi).$$

It follows that $E(Y) = b'(\theta)$ and $\text{var}(Y) = E[Y - b'(\theta)]^2 = a^2(\phi)E[\partial l/\partial \theta]^2 = b''(\theta)a(\phi)$. These results hold when the distribution is expressed in the canonical form (B-2).

It is straightforward to confirm that these expressions for the mean and variance are satisfied in the examples presented above.

B-2. GENERALIZED LINEAR MODELS

B-2.1 Definition

Generalized linear models were introduced by Nelder and Wedderburn (1972) as a means of uniting the theory of many models. The basic elements of this approach are presented here, following Agresti (1990), with the notation that is now in wide use.

The model has three elements, a *random component*, a *systematic component*, and a *link*, a function relating the random and systematic components. The random component consists of N independent observations $\mathbf{Y} = (Y_1, \dots, Y_N)'$. Each observation Y_i has a distribution of the second form of the exponential family, Equation (B-2) with parameters θ_i and ϕ . The parameter θ_i may vary with i , so that the Y_i 's do not all have the same mean.

The systematic component of the model is an N -dimensional vector $\boldsymbol{\eta}$, related to a vector of unknown parameters $\boldsymbol{\beta}$ and to known explanatory variables in a linear way:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} .$$

Here $\mathbf{X}$ is sometimes called the *design matrix*.

The link is a function relating the systematic component and the mean of the random component. The link is a monotonic differentiable function g with $\eta_i = g[E(Y_i)]$. If g is the identity function, so that $\eta_i = E(Y_i)$, g is called the *identity link*. If $g[E(Y_i)] = \theta_i$, with θ_i the canonical parameter, so that $\eta_i = \theta_i$, g is called the *canonical link*.

B-2.2 Examples

B-2.2.1 Standard Linear Model

The usual linear model, with the usual notation, has Y_i following a $\text{normal}(\mu_i, \sigma^2)$ distribution, with $\boldsymbol{\mu}$, the vector of means μ_i , equal to $\mathbf{X}\boldsymbol{\beta}$. In the notation of generalized linear models, we must have $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, with $\eta_i = g(\mu_i)$. Therefore, let g be the identity link, and the usual linear model is a special case of the generalized linear model. In this example, g is both the identity link and the canonical link.

The identity link is not appropriate for random variables that are restricted to positive values, such as counts. The identity link would say that $E(Y_i)$ is an element of $\mathbf{X}\boldsymbol{\beta}$. If $\boldsymbol{\beta}$ is unrestricted and $\mathbf{X}$ has no

rows consisting of all zeros, each component of $X\beta$ can take all possible values, including negative values. Therefore, other links, typically involving logarithms, are used for count data.

B-2.2.2 Loglinear Model

If the link is the natural logarithm, $g[E(Y_i)] = \log[E(Y_i)] = \sum_j x_{ij} \beta_j$, the model is loglinear. When Y_i is Poisson(μ_i), we have the standard loglinear model introduced in the body of this report. (The standard notation in categorical data literature uses N_i and m_i instead of Y_i and μ_i .) Because the parameter θ in Equation (B-2) is $\log \mu$ in this example, the log link is the canonical link.

Now suppose that μ_i equals $\lambda_i t_i$, for some known t_i . This treats the event *rate* λ as the fundamental parameter, and assumes that the count Y_i is based on a known exposure time t_i , which is not the same for all i . Assume that this fundamental parameter is related to the explanatory variables through the loglinear model $\log(\lambda_i) = \sum_j x_{ij} \beta_j$. There are two ways to formulate the rest of the model.

One way, following the approach given as an example of a generalized linear model, is to let Y_i denote the Poisson count divided by t_i . As shown above, the canonical parameter θ_i is $\log(\lambda_i)$. Then $E(Y_i)$ equals λ_i , and the logarithmic link $g[E(Y_i)] = \log(\lambda_i)$ is the canonical link. We then have

$$g[E(Y_i)] = \log(\lambda_i) = \sum_j x_{ij} \beta_j .$$

The other way is to let Y_i be the Poisson count, with parameter $\mu_i = \lambda_i t_i$, and to use the logarithmic link. Then we have

$$g[E(Y_i)] = \log(\mu_i) = \log(\lambda_i) + \log(t_i) = \sum_j x_{ij} \beta_j + \log(t_i) .$$

This is called a loglinear model with an offset, $\log(t_i)$. The difference between the two models is the definition of Y_i .

B-2.2.3 Logit Model

Suppose that the observable random variable is binomial(n_i, π_i), for n_i known and π_i unknown. As in the earlier example, let the observable random variable be denoted $n_i Y_i$, so that Y_i is the binomial random variable divided by n_i . As shown above, the canonical parameter is $\log[\pi/(1-\pi)]$, the logit of π . Because $E(Y_i)$ equals π_i , the canonical link is the logit function. The generalized linear model sets

$$g[E(Y_i)] = g\{\text{binomial}(n_i, \pi_i)/n_i\} = g(\pi_i) = \log[\pi_i/(1-\pi_i)] = \sum_j x_{ij} \beta_j .$$

Alternatively, we could formulate the model in terms of the n_i individual Bernoulli(π_i) random variables Y_{ik} that form the binomial(n_i, π_i) random variable $n_i Y_i$: $n_i Y_i = \sum Y_{ik}$. Consider the Bernoulli random variables as the observable quantities, and use the canonical link, to obtain

$$g[E(Y_k)] = g\{E[\text{Bernoulli}(\pi_i)]\} = g(\pi_i) = \log[\pi_i/(1-\pi_i)] = \sum_j x_{ij} \beta_j.$$

This avoids explicit use of the number n_i , but uses n_i implicitly because there are n_i such equations, identical except for the value of k .

B-3. RELATION OF POISSON AND MULTINOMIAL DISTRIBUTIONS

All the remaining sections of Appendix B use the notation of the body of this report, rather than the notation of Sections B-1 and B-2.

The examples in the body of this report involve Poisson or multinomial data, with binomial data being a special case of multinomial data. (The definition and basic facts about multinomial distributions are given in many intermediate-level statistics books.) This section shows that the Poisson and multinomial distributions are very closely related. This exercise has three functions: (a) It helps show the unity of the subject. (b) It enables the analyst to use computer software that appears to assume one distribution when the data correspond to the other distribution. (c) It sometimes allows the analyst to use conditioning to eliminate nuisance parameters from the model.

B-3.1 Statement of Relation

Let N_i be independent Poisson random variables with mean m_i , where i is a generic notation standing for a set of subscripts ranging over some set. Although the subscript is denoted here by i , in an example the subscript could be i ranging over all possible values, or ij ranging over all possible values, or ij with i variable and j fixed at a specific value, or some other set of subscripts with an associated range.

In the Unplanned-Demands Example, N_i is the number of events in calendar year i , and the Poisson assumption is very natural. Here $m_i = \lambda_i t_i$, where λ_i is the rate of occurrence of unplanned demands per plant operating year during calendar year i and t_i is the number of plant operating years in calendar year i . In the RCIC/HPCI or MOV-Test Example, on the other hand, N_{ij} and N_{2j} denote the numbers of failures and successes, respectively, corresponding to the j th level of the explanatory variables of interest, where j could index plant, system, calendar year, or some combination of the explanatory variables. In these two examples, the Poisson assumption seems unnatural, because we do not ordinarily think of failures and successes as having independent Poisson distributions. However, the Poisson distribution will shortly lead to the more natural binomial distribution.

Consider some fixed collection of the N_i 's—more precisely, the i 's are a fixed set, and the corresponding N_i 's are Poisson random variables. This collection could be every cell of the entire table, or it could correspond to a subtable. For example, the collection might be all of the N_i 's, such as all of the event counts in the MOV-Test Example. Possible subtables of the MOV-Test Example would be all the failure and success counts for a single year, with the plant ignored, or the set of failure counts and success counts for a single combination of year and plant, or the set of failure counts and success counts for some specified collection of year-plant combinations. In these examples, i indexes the outcome and

possibly some explanatory variables, with the other explanatory variables either specified or ignored. Define $N_+ = \sum N_i$ and $m_+ = \sum m_i$, summing over the assumed range of i , and let n_+ be the observed value of the random variable N_+ . Then conditional on n_+ , the N_i 's have a multinomial distribution, with the conditional mean of N_i equal to $n_+ m_i / m_+$. Before giving the proof, in Section B-3, we illustrate the relation by the following examples.

In the Unplanned-Demands Example, it is natural to consider N_i , the event count in calendar year i as a Poisson random variable. However, it is also possible to condition on the total number of events and treat the event counts, conditionally, as multinomial. Write the expectation of N_i as $m_i = \lambda_i t_i$, where t_i is the number of plant operating years in the i th calendar year. The occurrence rate for the i th calendar year is λ_i , with units 1/plant-operating-year. This normalization by t_i is reasonable because the HPCI system can be demanded only when the plant is operating. Let H_0 be the hypothesis that the event rates λ_i are all equal to a common λ , so that $m_i = \lambda t_i$ and $m_i / m_+ = t_i / t_+$, where $t_+ = \sum t_i$. The fact stated above says that, when H_0 is true, conditional on the total number of demands, 63, N_i is binomial(63, $p_i = t_i / 116.60$), and the conditional joint distribution of the N_i 's is multinomial. For example, the first cell, 1987, has N_1 binomial(63, 0.1285) because $14.98 / 116.60 = 0.1285$. The other cells are similar, and the joint distribution of $(N_1, \dots, N_7)$ is multinomial(63, 0.1285, ..., 0.1536).

In the RCIC/HPCI Example, suppose that we begin with the unnatural assumption that the failure count $N_{j\#}$ and the success count $N_{2\#}$ in plant j and system k are independent Poisson random variables. For example, $j = 5$ and $k = 1$ corresponds to the fifth plant, denoted Plant E, in the RCIC system. This had $n_{151} = 3$ failures and $n_{251} = 23$ successes, for a total of $n_{+51} = 26$ demands. Suppose that for each plant-system we condition on the number of demands, that is, we treat the total number of demands during each plant-system as fixed. This approach considers each subtable as corresponding to a single plant and system. Then the number of failures for that plant and system, conditional on the number of demands, is multinomial, which is binomial because there are only two possibilities, success and failure. Conditional on 26 demands for the RCIC system at plant E, the number of failures N_{151} is binomial and so is the number of successes. The parameters are proportional to the assumed Poisson means. When all the plant-systems are considered together, the overall distribution is that of independent multinomials (actually binomials), each multinomial describing the distribution within one subtable. Sometimes this joint distribution is called the product-multinomial distribution, because the likelihood is a product of the multinomial distribution functions for the various subtables.

This relation between independent Poisson, multinomial, and independent multinomial distributions is used extensively to obtain general asymptotic results without worrying much about which distribution is in effect.

B-3.2 Proof of the Relation

We now show that, conditional on n_+ , the N_i 's have a multinomial distribution, with the conditional mean of N_i equal to $n_+ m_i / m_+$.

The probability that $N_i = n_i$ for all i is

$$\begin{aligned} P(\dots, N_i = n_i, \dots) &= \prod_i \left[e^{-m_i} m_i^{n_i} / n_i! \right] \\ &= e^{-m_+} (\prod m_i^{n_i}) / (\prod n_i!) \end{aligned}$$

Because the N_i 's are independent Poisson, their sum, N_+ , is also a Poisson random variable, with parameter equal to the sum of the individual parameters, m_+ :

$$P(N_+ = n_+) = e^{-m_+} m_+^{n_+} / n_+!$$

The conditional probability of the n_i 's given n_+ is the quotient of the two probabilities:

$$\begin{aligned} P(\dots, N_i = n_i, \dots \mid N_+ = n_+) &= \frac{(\prod m_i^{n_i}) / (\prod n_i!)}{(m_+^{n_+}) / n_+!} \\ &= [\prod (m_i / m_+)^{n_i}] [n_+! / \prod n_i!] \\ &= \binom{n_+}{\dots, n_i, \dots} \prod \pi_i^{n_i} \end{aligned}$$

where the first term in the last line is the multinomial coefficient, and π_i is defined as m_i / m_+ . This is the probability of a multinomial distribution. Note in particular that, if i takes only two values (say, 1 and 2), then the distribution is binomial.

B-4. SOME PROPERTIES OF LOGLINEAR MODELS

B-4.1 Hierarchical Models and Marginal Constraints

The material of this section is explained in more detail by Bishop, Fienberg, and Holland (1975).

Consider loglinear models with unordered categorical explanatory variables. Let the response variable have either a Poisson distribution, a multinomial distribution, or a product-multinomial distribution. In every case, the cell counts are random. The first case has no constraints on the totals of the cell counts. The second case has one constraint: the total of all the cell counts is fixed. The third case has more constraints: the cell counts in various subtables of the overall table have independent multinomial distributions, so the total count for each subtable is fixed. The results below will be true, regardless of which distribution is assumed.

A loglinear model is *hierarchical* if, for every higher order effect in the model, the lower order effects composed of the variables in the higher effect are also in the model. For example

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)}$$

is hierarchical, because, corresponding to the interaction term $u_{12(ij)}$, the model contains the lower order effects: u , $u_{1(i)}$, and $u_{2(j)}$. If any of these three lower order terms had been missing from a model that contains the higher order term $u_{12(ij)}$, the model would not be hierarchical.

Corresponding to any interaction term is the marginal count with the same indices. In the above example, the main effect $u_{1(i)}$ corresponds to the count n_{i++} and the mean m_{i++} . The interaction $u_{12(ij)}$ corresponds to the count n_{ij+} and the mean m_{ij+} .

Hierarchical models have an important feature: if any effect is in the model, the corresponding estimated mean count equals the observed count. Here, "estimated" means "estimated by maximum likelihood." In the above example, $\hat{m}_{ij+} = n_{ij+}$, $\hat{m}_{+j+} = n_{+j+}$, and so forth. Note that knowledge of n_{ij+} for all i and j determines n_{i++} , n_{+j+} , and n_{+++} . When the model is hierarchical, the estimated cell counts are related in the same way.

When the model is hierarchical with an interaction term, the estimated cell counts are the same whether or not we condition on the observed values of the marginal totals corresponding to the interaction term. In the example with $u_{12(ij)}$, the estimated cell means are the same in the two cases (a) all the counts are independent Poisson, and (b) for each value of ij , the counts are multinomial with n_{ij+} fixed at the observed values. It does not matter which distribution is assumed—the point estimates are the same.

B-4.2 Relation between Logit and Loglinear Models

The logit model can be written as a loglinear model, as follows. As an example, let N_{ijk} be the number of outcomes of type i , system j , and plant k , with type 1 outcomes being failures and type 2 outcomes being successes. Assume that N_{ijk} is Poisson(m_{ijk}), and that they follow the loglinear model

$$\log(m_{ijk}) = u + u_{2(j)} + u_{3(k)} + u_{23(jk)} + u_{1(i)} + u_{12(ij)} + u_{13(ik)} + u_{123(ijk)} . \quad (\text{B-3})$$

It is essential that, for every higher order term, the corresponding lower order terms be included, so that the model is *hierarchical*, as defined above in Section B-4.1. The terms that do not involve variable 1 are written first, and all of them are included. The highest order term that does not involve variable 1, $u_{23(jk)}$, must be included in the model. Inclusion of this term forces the estimation procedure to make the fitted cell totals $\hat{m}_{+jk}$ equal to the observed total counts n_{+jk} , as explained in Section B-4. Forcing the fitted total to equal the observed total is equivalent to conditioning on the observed total, which is needed to obtain equivalence between the Poisson and multinomial distributions. Some of the terms that involve variable 1 could be dropped, as is discussed below.

When the relation was given between the Poisson and multinomial (or in this case binomial) distributions, the relation between the parameters was seen to be $\pi_{jk} = m_{1jk}/m_{+jk}$, where π_{jk} is the probability of a type 1 outcome, and the sampling distribution is product multinomial with each n_{+jk} fixed. Therefore, the logit function of π_{jk} is

$$\log[\pi_{jk}/(1 - \pi_{jk})] = \log[m_{1jk}/(m_{+jk} - m_{1jk})] = \log[m_{1jk}/m_{2jk}] = \log(m_{1jk}) - \log(m_{2jk}) .$$

Substitute the values from Equation (B-3) into this expression. Then the logit function of π_{jk} is

$$\text{logit}(\pi_{jk}) = u_{1(i)} + u_{12(ij)} + u_{13(ik)} + u_{123(ijk)} - u_{1(2)} - u_{12(2j)} - u_{13(2k)} - u_{123(2jk)} .$$

All the other terms, the ones that do not involve variable 1, cancel each other in the subtraction, and so are not shown. Recall that the terms in a loglinear model satisfy constraints such as $\sum_i u_{12(ij)} = 0$. Therefore, the equation becomes

$$\text{logit}(\pi_{jk}) = 2u_{1(i)} + 2u_{12(ij)} + 2u_{13(ik)} + 2u_{123(ijk)} = w + w_{2(j)} + w_{3(k)} + w_{23(jk)} ,$$

which has the form of a logit model. If the loglinear model (B-3) had not contained the third-order interaction, $u_{123(ijk)}$, the resulting logit model would have included the main effects, but not the interaction $w_{23(jk)}$. If it had contained neither $u_{123(ijk)}$ nor $u_{12(ij)}$, the resulting logit model would not contain the terms $w_{23(jk)}$ or $w_{2(j)}$.

B-5. DEVIANCE RESIDUALS

The deviance D is defined as $D = G^2 = -2\log\Lambda$, where Λ is the likelihood ratio test statistic for testing some hypothesis H_0 against H_a , the hypothesis of no unnecessary restrictions on the means of the random variables. The deviance residuals are defined so that the deviance D is the sum of the squared residuals. This is not quite as simple as one might suppose, because terms in D can cancel, and some of the terms in some representations of D can be negative. Therefore, we return to basics, and retain all contributors to D that depend on the parameters.

When the N_i 's are independent Poisson(m_i), the log-likelihood, in terms of m_i , is

$$\Sigma[-m_i + n_i \log(m_i) - \log(n_i!)]$$

so that

$$D = 2\Sigma\{n_i[\log(n_i) - \log(\hat{m}_i)] - n_i + \hat{m}_i\}.$$

This reduces to Equation (5) in the body of this report

$$D = 2\Sigma n_i[\log(n_i) - \log(\hat{m}_i)] ,$$

because under all reasonable models $\Sigma \hat{m}_i$ equals Σn_i . (More precisely, this fact holds for all hierarchical models, as explained in Section B-4, because any such model contains an overall effect u .) Therefore, the i th residual is defined as the square root of $2\{n_i[\log(n_i) - \log(\hat{m}_i)] - n_i + \hat{m}_i\}$, with a positive sign if $n_i > \hat{m}_i$ and a negative sign otherwise.

When the N_i 's are multinomial(n_+ , ..., p_i , ...), the log-likelihood is

$$\Sigma n_i \log(p_i) - \log(\text{multinomial coefficient}) ,$$

where the multinomial coefficient does not depend on the p_i 's. Therefore, we have

$$D = 2\Sigma n_i[\log(n_i/n_+) - \log(\hat{p}_i)] ,$$

which reduces to Equation (5) in the body of this report when $\hat{m}_i$ is set to $n_+ \hat{p}_i$. Therefore, the i th residual is defined as the square root of $2n_i[\log(n_i/n_+) - \log(\hat{p}_i)]$, with a positive sign if $n_i > n_+ \hat{p}_i$ and a negative sign otherwise.

Finally, suppose that Y_i is normal(μ_i, σ^2), for i from 1 to n , and let H_0 be a hypothesis about the μ_i 's. The log-likelihood is

$$-\frac{1}{2}\Sigma(y_i - \mu_i)^2/\sigma^2 - (n/2)\log(2\pi\sigma^2) .$$

Under H_∞ , the hypothesis of no restrictions on the μ_i 's, the MLE of μ_i is y_i and the first term vanishes. The second term is the same under H_0 and H_∞ . Therefore, the deviance is

$$D = \Sigma(y_i - \mu_i)^2/\sigma^2 ,$$

and the i th deviance residual is $(y_i - \mu_i)/\sigma$. These quantities are familiar from the analysis of linear models, and help motivate the deviance residuals of generalized linear models.

B-6. SMALL-SAMPLE BEHAVIOR OF LIKELIHOOD RATIO AND PEARSON TEST STATISTICS

A data analyst must consider whether the sample is large enough to use the χ^2 approximation. For small samples, when failures on demand are rare, it can happen that X^2 is far in the *right* tail of the χ^2 distribution while G^2 is far in the *left* tail of the χ^2 distribution—the two asymptotic approximations give strongly contradictory results.

The question of required sample size has been studied more for goodness-of-fit tests than for tests of nested hypotheses. This work with goodness-of-fit statistics is mentioned first. Moore (1986) cites work by Larntz (1978) and Read (1984), and concludes that the Pearson chi-square statistic, X^2 , approaches its asymptotic χ^2 distribution faster than does the likelihood ratio deviance, G^2 . Moore also discusses the power of the tests, and concludes that, for testing H_0 against the hypothesis of no restrictions on the expected counts, the test based on X^2 is preferable to that based on G^2 .

When is the sample size large enough to justify the χ^2 approximation for X^2 ? In the multinomial context, Moore (1986) cites Roscoe and Byars (1971) and recommends (a) with equiprobable cells, the average expected frequency (m_j in the notation of this report) should be at least 1 when testing fit at the $\alpha = 0.05$ level, and (b) when the cells are not approximately equiprobable, the average expected frequency should be doubled. Of course, any such rule based on an average can be defeated by a sufficiently skewed assignment of cell probabilities. In a $2 \times J$ contingency table, Everitt (1992, Sec. 3.3) cites Lewontin and Felsenstein (1965), and states that the χ^2 approximation is adequate if all the m_{ij} 's are 1.0 or greater, and in "the majority of cases" it is sufficient for the m_{ij} 's to be 0.5 or greater. These findings are also summarized by Engelhardt (1994) and Atwood (1994a), respectively.

Agresti (1990, Sec. 7.7.3) cites Larntz (1978), Koehler and Larntz (1980), and Koehler (1986), and states that the distribution of G^2 is usually poorly approximated by the χ^2 distribution when the average count per cell is less than 5. When most of the expected cell frequencies m_i are less than 0.5, G^2 tends to be smaller than the corresponding χ^2 random variable, and the reported p -values tend to be too large. If most of the expected frequencies m_i are between 0.5 and 5, G^2 tends to be larger than the χ^2 random variable, and the reported p -values tend to be too small.

McCullagh and Nelder (1989, pp. 118-122, 174) echo these concerns for using G^2 with moderately sparse data. They say, "It is good statistical practice, however, not to rely on either D or X^2 as an absolute measure of goodness of fit in these circumstances. It is much better to look for specific deviations from the model of a type that is easily understood scientifically. For instance, we may look for interactions among the covariates or non-linear effects by adding suitable terms to the model and

observing the reduction in deviance. The reduction in deviance thus induced is usually well approximated by a χ^2 distribution."

The above warnings about the use of D and X^2 are the reason why the SAS program GENMOD refuses to print a p -value corresponding to the calculated statistics, although it prints p -values corresponding to reductions in the deviance resulting from adding parameters to the model.

Now let us move from goodness-of-fit tests to tests of nested hypotheses. Agresti (1990, p. 250) points out that $G^2(H_0 | H_1)$ and $X^2(H_0 | H_1)$ depend on the data only through the fitted values, which in most loglinear models depend only on marginal totals. These marginal totals are larger than the individual cell counts. Therefore the test statistics approach their asymptotic distributions faster than do the corresponding goodness-of-fit statistics D and X^2 . Agresti suggests that if the marginal totals that determine the fits under H_1 are all in the range 5 to 10, the χ^2 approximation is likely to be adequate. In the passage just cited, McCullagh and Nelder assert that the χ^2 approximation "usually" works well for tests of nested hypotheses, at least when the difference in dimensions is small.

If both statistics are printed and they do not give the same conclusion, the sample is probably too small to use the χ^2 distribution for G^2 . A commonly used rule of thumb is that if the two tests agree, the sample size is probably large enough. I do not know if pathological examples exist in which this rule results in wrong conclusions.

The commercial computer package StatXact (1994) computes the exact p -values corresponding to the goodness-of-fit statistics X^2 , G^2 , and several others. It uses modern fast algorithms, although with large tables it is limited by the available machine memory. When it cannot calculate the exact p -value, it can calculate a confidence interval for the exact p -value by Monte Carlo simulation. One version of StatXact can interface with SAS, producing a single output stream. It can handle two-way contingency tables, and stratified $2 \times C$ contingency tables, but apparently not tables of higher orders, such as a general 3-way or 4-way table. According to the promotional material, it apparently does not perform goodness-of-fit tests for arbitrary loglinear models, even if they have only two explanatory variables. The companion package LogXact analyzes data using the logit model. If the MLEs fail to converge because the data are sparse (see the analysis of the RCIC/HPCI Example in the body of this report), it nevertheless can find exact upper confidence limits on the model parameters. The limitations of LogXact are not clear from the promotional material.

B-7. POINTWISE AND SIMULTANEOUS CONFIDENCE STATEMENTS

B-7.1 Derivations

This section derives formulas for pointwise confidence intervals for the fitted response and for simultaneous confidence bands for the fitted response. The formulas are all approximate, because they depend on the asymptotic normality of the parameter estimators.

The facts needed here for dealing with multivariate normal random variables are these. In this paragraph, let $\mathbf{X}$ be a (column) vector random variable with mean μ . Most of the applications below will be to a multivariate maximum likelihood estimator $\hat{\beta}$, but the facts stated in this paragraph are completely general. The covariance matrix of $\mathbf{X}$ is defined as $E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T]$, where T denotes the transpose. This matrix has variances on the diagonal and covariances off the diagonal. The covariance matrix, denoted $\mathbf{V}$ here, is symmetric positive definite, and therefore has a "square root"; that is, there is a matrix $\mathbf{U}$ such that $\mathbf{U}^T\mathbf{U} = \mathbf{V}$. Then $\mathbf{V}^{-1} = \mathbf{U}^{-1}(\mathbf{U}^T)^{-1}$. Finally, let $\mathbf{A}$ be a matrix with the same number of columns as $\mathbf{X}$ has rows, so that the matrix product $\mathbf{AX}$ is defined. Then $E(\mathbf{AX}) = \mathbf{A}\mu$ and $\text{cov}(\mathbf{AX}) = \mathbf{A}\mathbf{V}\mathbf{A}^T$, and if $\mathbf{X}$ is multivariate normal $\mathbf{AX}$ is also normal. In particular, if $\mathbf{A}$ is a single row, written now as $\mathbf{c}^T$, then $\mathbf{c}^T\mathbf{X}$ is one-dimensional, with mean $\mathbf{c}^T\mu$ and variance $\mathbf{c}^T\mathbf{A}\mathbf{c}$.

Consider first the generalized linear model setting; applications to loglinear and logit models will come later. The systematic component η corresponding to an observation Y is modeled as

$$\eta = \mathbf{c}^T\beta$$

where $\mathbf{c}$ is a column vector of values of explanatory variables, β is a column vector of parameters, and T denotes the transpose. For example, if η is $\beta_0 + x_1\beta_1 + x_2\beta_2$, where x_1 and x_2 are two explanatory variables, the expression can be written in vector notation with $\mathbf{c}^T = (1, x_1, x_2)$ and $\beta^T = (\beta_0, \beta_1, \beta_2)$. The maximum likelihood estimator $\hat{\beta}$ is asymptotically normal with mean β and covariance matrix denoted $\mathbf{V}$ here.

Let $\hat{\eta}$ denote the fitted value, $\mathbf{c}^T\hat{\beta}$. Therefore, we have

$$\begin{aligned}\hat{\eta} - \eta &= \mathbf{c}^T\hat{\beta} - \mathbf{c}^T\beta \\ &= \mathbf{c}^T(\hat{\beta} - \beta),\end{aligned}$$

which is asymptotically normally distributed with mean 0 and variance $\mathbf{c}^T \mathbf{V} \mathbf{c}$. Therefore, an approximate 90% confidence interval on η is $\hat{\eta} \pm 1.645(\mathbf{c}^T \mathbf{V} \mathbf{c})^{1/2}$. The number 1.645 is the 95th percentile of the standard normal distribution.

The above formula is valid for any single η and corresponding $\mathbf{c}$. It cannot be used simultaneously for the entire fitted line or surface, because the probability that a set of confidence intervals all contain the true values is less than the probability that any one interval contains the true value. Instead, the following idea is used, developed by Working, Hotelling, and Scheffé, and explained by Miller (1981). This notation uses $|\mathbf{v}| = (\mathbf{v}^T \mathbf{v})^{1/2}$ to denote the length of a vector $\mathbf{v}$. We have

$$\begin{aligned} |\hat{\eta} - \eta| &= |\mathbf{c}^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})| \\ &= |(\mathbf{U} \mathbf{c})^T [(\mathbf{U}^T)^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]| \quad \text{for any square invertible matrix } \mathbf{U} \\ &\leq |\mathbf{U} \mathbf{c}| |(\mathbf{U}^T)^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|. \end{aligned}$$

This inequality is the Cauchy-Schwarz inequality, proved in linear algebra books. The inequality holds simultaneously, for all values of $\mathbf{c}$. In particular, let $\mathbf{U}$ be such that $\mathbf{U}^T \mathbf{U} = \mathbf{V}$. The lengths of the two vectors in the final product can be rewritten, leading to

$$|\hat{\eta} - \eta| \leq (\mathbf{c}^T \mathbf{V} \mathbf{c})^{1/2} [(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{V}^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]^{1/2}.$$

The portion in square brackets, involving $\hat{\boldsymbol{\beta}}$ and its mean and covariance matrix but not $\mathbf{c}$, is $\mathbf{Z}^T \mathbf{Z}$, where $\mathbf{Z} = (\mathbf{U}^T)^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is approximately normal random with mean 0 and covariance matrix $\mathbf{I}$. Thus, $\mathbf{Z}^T \mathbf{Z}$ is approximately χ^2 distributed, with degrees of freedom equal to the dimension of $\mathbf{Z}$, which is the number of unknown parameters β_i . Denote this number of parameters by p . It follows that 90% confidence limits on η , simultaneously valid for all $\mathbf{c}$, are

$$\hat{\eta} \pm (\mathbf{c}^T \mathbf{V} \mathbf{c})^{1/2} [\chi^2_{0.90}(p)]^{1/2}.$$

The pointwise and simultaneous confidence intervals can be translated back to the original model. In the Poisson case, we have $\eta = \log[E(N/t)]$, and the mean response $E(N/t)$ is equal to $\exp(\eta)$. Therefore, pointwise or simultaneous confidence intervals for η translate into the corresponding intervals for the mean response. In the binomial case, η is the logit of π . Therefore, $\pi = e^\eta / (1 + e^\eta)$, a monotonically increasing function of η , and confidence bounds on η translate into the corresponding confidence bounds on π .

It is awkward to use the equivalence of loglinear and logit models, discussed in Section B-4.2, and to obtain confidence bounds for π based on the loglinear parameterization. One reason is that the number of parameters is different: two Poisson counts N_1 and N_2 correspond to two parameters, but conditional on the sum n_+ , the binomial fractions N_1/n_+ and $N_2/n_+ = 1 - N_1/n_+$ involve only one parameter. Therefore, the dimension of β is different for the two setups, and the covariance matrices are different.

B-7.2 Discussion

The formulas derived above, both the pointwise and the simultaneous formulas, involve V , the asymptotic covariance matrix of $\hat{\beta}$. Any good software package finds an estimate of this matrix, using the asymptotic theory for maximum likelihood estimation. The details are not given here. To apply the formulas in practice, this estimate of V must be used. Some packages give two options, based on the "observed" and "expected" Fisher information. Efron and Hinkley (1978) argue that the observed information is generally more appropriate for confidence interval calculations.

It is interesting to compare these formulas to the corresponding formulas for linear models. Under the usual normal theory for linear models, the pointwise confidence intervals involve the estimated standard deviation s and a percentile of Student's t distribution. The simultaneous confidence band involves s and a percentile of the F distribution. The difference between those formulas and the ones here are that the loglinear model does not have the extra parameter σ to estimate. Instead, the Poisson or binomial mean determines the corresponding variance. When the variance is known, Student's t distribution is replaced by a normal distribution and the F distribution is replaced by a χ^2 distribution. On the other hand, there is a price for this simplification. The normal assumption for linear models results in exact confidence statements, while all the results here use asymptotic normality, and so are only approximate.

Sometimes, for application in probabilistic risk assessments, a Bayesian distribution for the mean response may need to be found. A confidence interval for η is based on the asymptotic normality of $\hat{\eta}$; these calculations are all non-Bayesian. If η were assigned the same Bayesian distribution, normal with the same calculated variance, and mean equal to the calculated $\hat{\eta}$, the Bayesian intervals would be numerically identical to the confidence intervals. For example, the 90% Bayes intervals would coincide with the 90% confidence intervals, the 95% intervals would coincide, etc., because both kinds of interval are calculated using the same normal distribution. Therefore, in the absence of prior information about η , it is reasonable to use this normal distribution as the Bayes distribution for η . This argument could be formalized by using a noninformative (uniform) prior on η , and updating it with the single normally distributed observation $\hat{\eta}$, but we will not do this here.

This Bayesian distribution for η results in a Bayesian distribution for the parameter of interest. In the loglinear model with Poisson data, the corresponding distribution for the mean response $E(N/t) = \exp(\eta)$ is lognormal. In the logit model with binomial data, the distribution for $\pi = \exp(\eta)/[1 + \exp(\eta)]$ is logistic-normal (see Titterington 1989 for references).

B-8. REFERENCES

- Agresti, Alan, 1990, *Categorical Data Analysis*, New York: Wiley-Interscience, John Wiley & Sons.
- Atwood, Corwin L., 1994a, *Hits per Trial: Basic Analysis of Binomial Data*, EGG-RAAM-11041.
- Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul W. Holland, 1975, *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press.
- Efron, Bradley, and David V. Hinkley, 1978, "Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information" (with discussion), *Biometrika*, **65**, pp. 457-487.
- Engelhardt, Max, 1994, *Events in Time: Basic Analysis of Poisson Data*, EGG-RAAM-11088.
- Everitt, B. S., 1992, *The Analysis of Contingency Tables, Second Edition*, London: Chapman & Hall.
- Koehler, K., 1986, "Goodness-of-Fit Tests for Log-Linear Models in Sparse Contingency Tables," *Journal of the American Statistical Association*, **81**, pp. 483-493.
- Koehler, K. and K. Larntz, 1980, "An Empirical Investigation of Goodness-of-Fit Statistics for Sparse Multinomials," *Journal of the American Statistical Association*, **75**, pp. 33-344.
- Larntz, K., 1978, "Small-Sample Comparisons of Exact Levels for Chi-Squared Goodness-of-Fit Statistics," *Journal of the American Statistical Association*, **73**, pp. 253-263.
- Lewontin, R. C., and J. Felsenstein, 1965, "The Robustness of Homogeneity Tests in $2 \times N$ Tables," *Biometrics*, **21**, pp. 19-33.
- McCullagh, P., and J. A. Nelder, 1989, *Generalized Linear Models, Second Edition*, London: Chapman and Hall.
- Miller, Rupert G., Jr., 1981, *Simultaneous Statistical Inference, Second Edition*, New York: Springer-Verlag, pp. 63-67.
- Moore, David S., 1986, "Tests of the Chi-Squared Type," in *Goodness-of-Fit Techniques*, edited by Ralph B. D'Agostino and Michael A. Stephens, New York: Marcel Dekker, Inc., pp. 63-96.
- Nelder, J., and R. W. M. Wedderburn, 1972, "Generalized linear models," *Journal of the Royal Statistical Society, Series A*, **135**, pp. 370-384.

Read, T. R. C., 1984, "Small Sample Comparisons for the Power Divergence Goodness-of-Fit Statistics," *Journal of the American Statistical Association*, 79, pp. 929-935.

Roscoe, J. T., and J. A. Byars, 1971, "An Investigation of the Restraints with Respect to Sample Size Commonly Imposed on the Use of the Chi-square Statistic," *Journal of the American Statistical Association*, 66, pp. 755-759.

SAS/STAT Software: The GENMOD Procedure, Release 6.09, 1993, SAS Technical Report P-243, Cary NC: SAS Institute, Inc.

StatXact for SAS, 1994, Cambridge, MA: CYTEL Software Corp.

Titterington, D. M., 1989, "Logistic-normal distribution," in *Encyclopedia of Statistical Sciences, Supplement Volume*, S. Kotz and N. L. Johnson, editors, New York: Wiley-Interscience, pp. 90-91.

Appendix C

Index

Appendix C

Index

This index gives the locations of definitions, and some of the locations where terms are mentioned. If you cannot find a term here, look in the table of contents, or in the review section. If a phrase is not listed under the initial word, it may be indexed under the noun. Commonly used terms, such as "Poisson," are not indexed; they occur throughout the report. Also, many of the statistical concepts defined early in the report are used extensively during the example analyses; those usages are too numerous to index here.

- Analysis of covariance 8
- Analysis of variance 8
- Bayes
 - distribution for mean response B-21
 - empirical 47
- Categorical variable
 - definition 1
 - see also Explanatory variable
- Confidence bands, pointwise and simultaneous 33—36, 39, A-9—A-11, B-19—B-21
- Crossed effects, definition 44
- D*, deviance, definition 21
- Deviance residuals, definition 21, B-15
- Dimension of a model 16—18
- Distributions
 - examples of exponential family B-4—B-5
 - relation of binomial and multinomial 14, 15
 - relation of Poisson and multinomial 15, B-10—B-12, B-13
- Effects
 - see Main, Interaction, Nested, Crossed, Fixed, Random
- Empirical Bayes modeling 47
- Examples
 - MOV-Test 3—4, 8, 13, 14, 36—40, A-16—A-22, B-10
 - Prenatal care 5—6, 43
 - RCIC/HPCI 1—3, 8, 13, 14, 16, 18—20, 25—29, A-4—A-8, B-10, B-11, B-18
 - SFL 4—5, 8, 14
 - Unplanned-Demands 3, 8, 12, 13, 17, 29—36, A-12—A-15, B-10, B-11
- Explanatory variable
 - definition 1
 - numerical 3, 8, 13, 14
 - ordered 3, 8
 - unordered categorical 3, 8, 10, 13, 25
- Exponential family B-3—B-6, B-7
- Exposure times
 - definition 3
 - unequal 12
- Fixed effect 47
- G^2 , likelihood ratio goodness-of-fit statistic, definition 21

- $G^2(H_0 \mid H_1)$, definition 21
- Generalized linear model 10, 47, B-7
 - see also Model
- H_∞ , definition 19
- Hypothesis
 - definition 17
 - graph for showing 41
 - see also Model
- Interaction
 - definition 11
 - graph for showing 41—42
- Level of a variable, definition 1
- Likelihood, definition 19—20
- Likelihood ratio test
 - definition 19—20
 - relation with Pearson test 22
- Linear model
 - as a generalized linear model B-7
 - compared to loglinear or logit model 8, 9, 34—35
- Link function 10, 14, B-7
- Logarithm, definition as natural log 10
- Logistic regression
 - definition 14
 - example 36—39
- Logit function, definition 14
- Logit model
 - as a generalized linear model B-8
 - definition 14
 - relation with loglinear model 15, B-14
 - see also Generalized linear model, Linear model
- Loglinear model
 - as a generalized linear model B-8
 - definition 10
 - relation with logit model 15, B-14
 - see also Generalized linear model, Linear model
- Main effect, definition 11
- Maximum likelihood estimator (MLE) 16, B-13
- Model
 - definition 17
 - hierarchical 12, B-13—B-14
 - nested 18—21, 22
- Necessary restrictions, definition 17
- Nested effect, definition 44
- Nested models or hypotheses, definition 18
- Numerical variable
 - see Explanatory variable
- Odds ratio 14
- Offset 12, 30, B-8
- Overdispersion 29, 36

- Pearson chi-square test
 - definition 22
 - relation with likelihood ratio test 22
- Pearson residuals, definition 22
- Product-multinomial distribution, definition B-11
- Random effect 47—48
- Regression
 - linear model 8
 - logistic 14, 36—39
 - Poisson 13, 29—36
- Residuals
 - deviance 21, B-15
 - Pearson 22
- Response variable
 - definition 1
 - see also Explanatory variable
- Simpson's paradox 5—6, 42—44
- Small samples 23, B-17—B-18
 - see also Zero counts
- Software
 - CATMOD 25
 - for small samples 24, B-18
 - FRAC 48
 - GENMOD 7, 23, 25, A-3—A-22
 - GENSTAT 48
 - GLIM 25, 45
 - GLIMMIX 47
 - LogXact B-18
 - NLIN 25
 - SAS programs A-3—A-22
 - StatXact B-18
- X^2 , Pearson chi-square statistic, definition 22
- $X^2(H_0 | H_1)$
 - calculation A-3
 - definition 22
- Zero counts 26—27, 44—47
 - see also Small samples