Categorical Data Analysis

J.K. Lindsey

March 14, 2000

Contents

Chapter 1

Categorical Variables and Related Distributions

1.1 Categorical Variables

Much of the observed data which a statistician encounters is not in the form of quantitative measurements.

Rather some characteristic or attribute of the individuals is recorded.

Such characteristics take one or more distinct values.

1.1.1 Events

The cases of only one and two values for the variable are of special interest, since they are the most commonly used.

For a single value, the observations are usually summarized as a *count* of the number of occurrences of an event of interest.

Example

If the event is the birth of a child, then the counts might be the number of children in a family. □

Variables with two values are called *binary*. They are often used to record the occurrence and nonoccurrence of an event, usually over time or through space.

Example

Cancer patients are observed to be either alive (coded 0) or dead (coded 1) over a period of time. This binary variable can only change from zero to one. The sequence is called a point or counting process and is equivalent to observing the survival time. □

1.1.2 Nominal Variables

In general, if any one of say I qualitatively different events may occur to an individual, we have a *nominal* variable.

Each different event has a different name, but no mathematical relationship exists among the events.

Each possible characteristic or value of the variable is called a *category* or *level.*

Example

The sex of an individual is a binary categorical variable. More complex nominal variables include the profession of a worker and the type of illness of a hospital patient. □

When a number of individuals are observed, they can be classified into the I possible categories.

The number in each category, n_i, is called the (absolute) *frequency.*

This may also be transformed by dividing by the total number of individuals, $n_{.} = \sum n_i$, to yield the proportion or *relative frequency* in each category.

For clarity of presentation, these numbers are often multiplied by 100 to give *percentages.*

Example

1681 residents of Copenhagen were asked about the type of housing in which they lived. The results are summarized in the following table.

	Type of Housing			
	Tower Block	**Apartment**	**Atrium House**	**Terraced House**
Absolute Frequency	400	765	239	277
Relative Frequency	0.2380	0.4551	0.1422	0.1648
Percent	23.80	45.51	14.22	16.48

□

1.1.3 Ordinal Variables

Often, a categorical variable contains more information than simply the names of the categories.

If the categories can be strictly ordered, we have an *ordinal* variable.

Such variables frequently occur for the preferences of individuals or their state of health.

When available, such information should be used in the statistical analysis.

Example

256 Americans who graduated from high school in 1965 were asked their political party identification in 1982. The absolute frequencies are given in the following table.

Strong Democrat	10
Weak Democrat	59
Leaning Democrat	41
Independent	26
Leaning Republican	44
Weak Republican	47
Strong Republican	29

□

1.1.4 Counts and Frequencies

Counts and (absolute) frequencies are very similar and, indeed, are not always distinguished. Both are numbers of events.

A count is made of events on one individual unit of observation, such as the family above.

A frequency is an aggregation of events on different units of observation, with each unit appearing only once, at least at any given point in time.

Example
Consider the following distribution of accidents:

Accidents	**Frequency**
0	447
1	132
2	42
3	21
4	3
5	2
6	0

The first column is the count per individual; the second column is the frequency with which that count occurs across individuals. □

Similar statistical techniques can often be used for both counts and frequencies.

However, since counts involve events on the same unit, there will often be some form of dependence among these events, which often may need to be taken into account.

In contrast, frequencies refer to numbers of independent events, since they occur on different units.

1.1.5 Other Types of Variables

Any variable can be reduced to a simpler form by ignoring its special characteristics.

A quantitatively measured variable may be cut into a series of distinct categories, usually more or less arbitrarily.

Example

Income is often recorded as a categorical variable, say to the nearest 500 francs. □

In fact, any quantitative variable can only be measured in a categorical way, since all measuring instruments have some finite limit to their resolution.

Example

The length of employment of a certain type of British postal workers was recorded to the nearest month:

Months	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Freq.	22	18	19	13	5	6	3	2	2	1	0	1	0	0	0	1	1	1	3	1	1	0	0	0

□

The question is rather whether the statistical technique applied to the data uses the quantitative information contained in the labels on the categories.

When only the nominal information in a variable is used, no (mathematical) relationships exist among the categories.

Statistical analysis must rely on the frequencies of occurrence of the categories to provide the mathematical structure.

Thus, the less is known or assumed about the relationships among the categories, the more observations are required in order to have sufficiently large frequencies in each category.

1.2 Poisson Distribution

If events of the i^{th} type are independent across individuals and occurring at a uniform rate, τ_i, then the (random) number of such events, say N_i, will have a *Poisson distribution* with probability mass function

$$\Pr(N_i = n_i; \mu_i) = \frac{\mathrm{e}^{-\mu_i}\mu_i^{n_i}}{n_i!}$$

where $\mu_i = 1/\tau_i$ is the mean number of events and where the total number of events, $n_. = \sum n_i$, is not fixed in advance.

This distribution is characterized by the relationship between its mean and its variance:

$$\begin{aligned} \mathrm{E}[N_i] &= \mu_i \\ &= \mathrm{var}[N_i] \end{aligned}$$

Example

Consider the classical data on the numbers of deaths by horse kicks each year between 1875 and 1894 in 14 corps of the Prussian army:

Deaths/Corps/Year	0	1	2	3	4	5
Frequency	144	91	32	11	2	0

Here, the mean is estimated to be $\hat{\mu} = 0.70$ deaths per year per corps. □

If each category of event has a Poisson distribution, then the total number of events of all kinds, $n_{.}$, will also have a Poisson distribution, with mean $\mu = \sum \mu_i$.

The hypotheses of the Poisson distribution may often be reasonable for frequencies since the events are independent across individuals.

The question is whether the (categories of) individuals whose events are grouped in the frequencies are homogeneous enough so that they all have the same rate for the event.

Since a count refers to a number of events all on the same individual unit, the dependency among them must be examined closely.

On the other hand, all of the counted events will usually have the same rate, or they would not have been counted together.

Most often, the Poisson distribution will not be found suitable for counts.

Example

For the deaths by horse kicks, there are, in fact, two types of corps. One may need to investigate if they both have the same death rate. □

Thus, for frequencies, the heterogeneity among individuals must be checked, while, for counts, the dependence among events on an individual plays a more important role.

One indication will be that the mean and variance are substantially different.

According to the direction of the difference, it is known as under- or overdispersion.

The most common correction is to replace the Poisson distribution by the negative binomial.

1.3 Multinomial Distribution

Suppose now that we keep the same hypotheses as for the Poisson distribution, but fix the total number of events, $n_{.}$, before making the observations.

We must now look at the conditional distribution

$$\begin{aligned}
\Pr(n_1,\ldots,n_I|n_{.};\mu_1,\ldots,\mu_I) &= \frac{\prod_{i=1}^{I}\frac{\mathrm{e}^{-\mu_i}\mu_i^{n_i}}{n_i!}}{\frac{\mathrm{e}^{-\mu}\mu^{n_{.}}}{n_{.}!}} \\
&= \binom{n_{.}}{n_1\cdots n_I}\prod_{i=1}^{I}\left(\frac{\mu_i}{\mu}\right)^{n_i} \\
&= \binom{n_{.}}{n_1\cdots n_I}\prod_{i=1}^{I}\pi_i^{n_i}
\end{aligned}$$

where $\pi_i = \mu_i/\mu$ may take values between zero and one, with sum equal to one, and hence are probabilities.

This is known as the *multinomial distribution.*

It describes the distribution of I different types of events occurring independently, each type of event with a constant rate, where the total number of events is fixed.

This relationship between the Poisson and multinomial distributions is important.

It allows us to construct univariate models for categorical data whose frequencies are multivariate, simply by conditioning on the total number of events.

Example

In the Copenhagen housing example, the distribution of the $n_. = 1681$ residents might be taken to be multinomial, with four categories.

However, it can be modelled as Poisson by conditioning on the observed total, $n_.$. □

1.3.1 Binomial Distribution

A special case of the multinomial distribution, when only two types of events are observed, so that the variable is binary, merits mention.

This is the *binomial distribution*:

$$\Pr(N_1 = n_1 | n_.; \mu_1) \quad = \quad \binom{n_.}{n_1} \pi_1^{n_1} (1 - \pi_1)^{n_. - n_1}$$

The mean and variance of the random variable, N_1, are given by

$$\begin{aligned} \mathrm{E}[N_1] &= \mu_1 \\ &= n_. \pi_1 \\ \mathrm{var}[N_1] &= n_. \pi_1 (1 - \pi_1) \end{aligned}$$

1.4 Chi-Squared Distribution

If U_i are random variables having independent standard normal distributions, with mean 0 and variance 1,

$$U_i \sim \mathrm{N}(0, 1)$$

then U_i^2 has a *Chi-squared distribution* with one degree of freedom, χ_1^2 and

$$\begin{aligned} Z_p &= \sum_{i=1}^{p} U_i^2 \\ &\sim \chi_p^2 \end{aligned}$$

a Chi-squared distribution with p degrees of freedom and $\mathrm{E}[Z_p] = p$.

Often, we have a random variable, Y_i, with mean, μ_i, and variance, σ^2, such that

$$U_i = \frac{Y_i - \hat{\mu}_i}{\sigma}$$

so that

$$Z = \frac{\sum(Y_i - \hat{\mu}_i)^2}{\sigma^2}$$

where the variance, σ^2, is known.

For p large, $\chi_p^2 \doteq \mathrm{N}(p, 2p)$.

The Chi-squared distribution is a special case of the *gamma distribution*:

$$f(y) = \frac{y^{\frac{p}{2}-1}\mathrm{e}^{-\frac{y}{2}}}{\Gamma\left(\frac{p}{2}\right) 2^{\frac{p}{2}}}$$

1.4.1 Maximum Likelihood Estimate

The maximum likelihood estimate (m.l.e.), $\hat{\psi}$, has asymptotic distribution $\mathbf{N}[\psi, \mathbf{I}^{-1}(\psi)]$.

Under mild regularity conditions, for n independent observations, we know that the mean and variance of the score, $\mathbf{U}$, are

$$\mathrm{E}[\mathbf{U}(\psi)] = 0$$

and

$$\begin{aligned} \mathrm{E}[\mathbf{U}\mathbf{U}^T] &= \mathrm{E}[-\mathbf{U}'] \\ &= \mathbf{I} > 0 \end{aligned}$$

where $\mathbf{I}$ is the Fisher information.

Expand the score in a Taylor series about the true value, ψ

$$\mathbf{U}(\hat{\psi}) = \mathbf{U}(\psi) + \mathbf{U}'(\psi)\frac{\hat{\psi} - \psi}{1!} + \ldots$$

The left hand side is zero.

By the law of large numbers,

$$\lim_{n \to \infty}[-\mathbf{U}'(\psi)] = \mathbf{I}(\psi)$$

so that

$$(\hat{\psi} - \psi) \doteq \mathbf{I}^{-1}(\psi)\mathbf{U}(\psi)$$

The mean and variance of the right hand side are 0 and $\mathbf{I}^{-1}(\psi)$.

Then, since $\mathbf{U}(\psi)$ is a sum, by the central limit theorem, asymptotically

$$\hat{\psi} \sim \mathrm{MVN}[\psi, \mathbf{I}^{-1}(\psi)]$$

Since, ψ is typically unknown, any consistent estimate of $\mathbf{I}(\psi)$, such as $\mathbf{I}(\hat{\psi})$, can be used without affecting the limiting distribution. □

This implies, asymptotically, that the standard error of the parameter estimates is the square root of the diagonal elements of $\mathbf{I}^{-1}(\psi)$ and that

$$(\hat{\psi}-\psi)^T\mathbf{I}(\psi)(\hat{\psi}-\psi) \sim \chi^2_p$$

where p is the dimension of ψ.

This result is known as Wald's statistic.

Example

For the parameter of the binomial distribution, with

$$\mathrm{I}(\pi_1) \quad = \quad \frac{n_\cdot}{\pi_1(1-\pi_1)}$$

Wald's statistic is

$$\frac{n_\cdot(\hat{\pi}_1-\pi_1)^2}{\pi_1(1-\pi_1)} = \frac{(n_1-n_\cdot\pi_1)^2}{n_\cdot\pi_1(1-\pi_1)}$$

□

Wald's statistic and the asymptotic standard errors have several major handicaps, especially in small samples:

- If the log likelihood is not quadratic (i.e. Gaussian) for a parameter, they can be very misleading.
- They are not invariant under parameter transformations.

Thus, in categorical data analysis, Wald's statistic and the asymptotic standard errors should only be used with great care and as an approximation.

1.4.2 Log Likelihood and Deviance

Expand the log likelihood function as a Taylor series at $\psi = \hat{\psi}$:

$$\begin{aligned}\mathrm{l}(\psi) \quad &= \quad \mathrm{l}(\hat{\psi}) + (\psi-\hat{\psi})^T\mathrm{l}'(\hat{\psi}) \\ &\qquad + \frac{1}{2}(\psi-\hat{\psi})^T\mathrm{l}''(\hat{\psi})(\psi-\hat{\psi}) + \ldots\end{aligned}$$

Since $\mathrm{l}'(\hat{\psi}) = 0$, we have minus two times the log likelihood ratio, $\mathrm{l}(\psi) - \mathrm{l}(\hat{\psi})$ called the *deviance*,

$$\mathrm{D}(\psi) \doteq (\psi-\hat{\psi})^T\mathbf{I}(\hat{\psi})(\psi-\hat{\psi})$$

For n sufficiently large, $\hat{\psi}$ will be close to the true value, ψ, and this will be a good approximation.

As we have seen above,

$$(\psi-\hat{\psi})^T\mathbf{I}(\hat{\psi})(\psi-\hat{\psi}) \sim \chi^2_p$$

so that, asymptotically,

$$\mathrm{D}(\psi) \sim \chi^2_p$$

where ψ is the true value, with dimension p.

Example
For the binomial distribution, the deviance is

$$\begin{aligned} \mathrm{D}(\pi_1) &= -2\left[n_1 \log\left(\frac{\pi_1}{\hat{\pi}_1}\right) + (n_. - n_1)\log\left(\frac{1-\pi_1}{1-\hat{\pi}_1}\right)\right] \\ &= 2\sum_{i=1}^{2} n_i \log\left(\frac{n_i}{n_.\pi_i}\right) \end{aligned}$$

In categorical data analysis, this is often called G^2. □

Now, suppose that we wish to compare this full model to some submodel, ψ_1, of dimension $r < p$, *nested* in Ψ, i.e. where $\Psi_1 \subset \Psi$.

We have

$$\begin{aligned} -2[\mathrm{l}(\hat{\psi}_1) - \mathrm{l}(\hat{\psi})] &= -2\{[\mathrm{l}(\psi) - \mathrm{l}(\hat{\psi})] - [\mathrm{l}(\psi_1) - \mathrm{l}(\hat{\psi}_1)] \\ &\qquad - [\mathrm{l}(\psi) - \mathrm{l}(\psi_1)]\} \\ &= \mathrm{D}(\psi) - \mathrm{D}(\psi_1) + 2[\mathrm{l}(\psi) - \mathrm{l}(\psi_1)] \end{aligned}$$

The first term has a χ^2_p distribution, the second, χ^2_r, and the third is a positive constant, near zero if the correct model is indexed by ψ_1. Then,

$$\mathrm{D}(\psi) - \mathrm{D}(\psi_1) \sim \chi^2_{p-r}$$

under $\psi_1 \in \Psi_1 \subset \Psi$, since sums of Chi-squared variables are Chi-squared.

1.4.3 Score

Since we know that the mean of the score is zero and its variance is the Fisher information, and since the score is a sum, by the central limit theorem, asymptotically

$$\mathbf{U}(\psi) \sim \mathrm{MVN}[0, \mathbf{I}(\psi)]$$

and, hence,

$$\mathbf{U}^T(\psi)\mathbf{I}^{-1}(\psi)\mathbf{U}(\psi) \sim \chi^2_p$$

This is called the *score statistic*.

The same result can be obtained in another way.

From the asymptotic normality of the m.l.e., we know that

$$\hat{\psi} - \psi \doteq \mathbf{I}^{-1}(\psi)\mathbf{U}(\psi)$$

Substituting this into the asymptotic distribution of the deviance, we obtain

$$\mathrm{D}_U(\psi) = \mathbf{U}^T(\psi)\mathbf{I}^{-1}(\psi)\mathbf{U}(\psi)$$

which will have an asymptotic Chi-squared distribution.

The advantage of this statistic, as compared to the deviance and its normal approximation, is that it does not require the calculation of $\hat{\psi}$, but depends only on the fixed value, ψ.

Example
For the binomial distribution, with

$$\mathrm{U}(\pi_1) \quad = \quad \frac{n_1 - n_{.}\pi_1}{\pi_1(1-\pi_1)}$$

and Fisher information as given above, we have the score statistic

$$\frac{(n_1 - n_{.}\pi_1)^2}{n_{.}\pi_1(1-\pi_1)} = \sum_{i=1}^{2} \frac{(n_i - n_{.}\pi_i)^2}{n_{.}\pi_i}$$

which, in this case, is identical to Wald's statistic. □

This is a simple case of the *Pearson Chi-squared statistic*, which is the score statistic approximation to the deviance,

$$\mathrm{D}(\pi_1) = 2\sum_{i=1}^{2} n_i \log\left(\frac{n_i}{n_{.}\pi_i}\right)$$

given above.

Both have an asymptotic Chi-squared distribution.

Example
For the postal workers example, suppose that we entertain the null hypothesis of constant loss over the 24 months.

The constant probability of loss is $\pi_i = \frac{1}{24}$, which gives a Pearson statistic of 243.7 and a deviance of 189.5, both with 23 d.f. □

Chapter 2

Contingency Tables and Independence

2.1 Contingency Tables

Throughout this chapter, we shall concentrate on the relationships between only two variables, since more complex situations are more easily handled by the construction of formal models, presented in the next chapter.

2.1.1 Two-way Tables

Suppose that X and Y are two categorical variables having respectively I and J different levels.

If individuals are classified simultaneously according to both variables, IJ combinations are possible.

This can be displayed as a rectangular table with I rows and J columns, with the cells of the table representing the possible outcomes.

When the cells contain the frequencies, say n_{ij}, of outcomes in a sample, the table is called a *contingency table.*

The *marginal totals* are represented by $n_{.j}$, $n_{i.}$ and $n_{..}$.

Example

Injuries in car accidents in Florida in 1988 are classified as to whether a seat belt was being used at the time or not.

	Injury		
Seat Belt	**Fatal**	**Nonfatal**	**Total**
No	1601	162527	164128
Yes	510	412368	412878
Total	2111	574895	577006

Here, we have a 2 × 2 table. □

In general, we have an $I \times J$ table.

When presenting the frequencies of a contingency table as proportions or percentages, it is important to indicate in which direction they are calculated.

Example

For the car accident data, the percentages are

	Injury		
Seat Belt	**Fatal**	**Nonfatal**	**Total**
No	0.98	99.02	100.00
Yes	0.12	99.88	100.00
Total	0.37	99.63	100.00

Percentages might also be calculated separately for each type of accident (the columns) or globally for the complete table. □

2.1.2 Types of Designs

Prospective Studies

In a *prospective* study, individuals are sampled from a population and then followed over a certain period of time. Two cases may be distinguished.

1. In a *clinical trial*, the subjects are randomly allocated to one of a number of different treatments before the followup.

   Of all the designs mentioned, this is the only one which is experimental.

2. In a *cohort study*, all variables are simply observed as they occur over time.

Cross-sectional Studies

A *cross-sectional study* simply observes all variables on individuals at one given fixed point in time.

The data in the car accident example come from such a study.

Retrospective Studies

In a *retrospective* or *case-control* study, subjects are chosen according to their response values and then the values of the explanatory variables obtained.

Thus, the explanatory variables are random and the response fixed.

Example

58 married women under treatment for myocardial infarction in England and Wales during 1968–1972 were each matched with three control patients in the same hospitals who were being treated for something else.

All subjects were asked if they had ever used contraceptives, yielding the following table:

	Myocardial Infarction	
Contraceptive	**Yes**	**No**
Yes	23	34
No	35	132

□

2.2 Probability and Dependence

2.2.1 Joint and Conditional Probabilities

Suppose, for the moment, that only the total number of events, $n_{..}$, is fixed. This will be the case in cross-sectional and cohort studies.

Denote the probability of outcome (i, j) by π_{ij}.

These probabilities describe the *joint distribution* of X and Y, and might be taken to have a multinomial distribution.

The *marginal distributions* are obtained by summing the joint probabilities to obtain row or column totals.

Denote these by

$$\begin{aligned} \pi_{i.} &= \sum_j \pi_{ij} \\ \pi_{.j} &= \sum_i \pi_{ij} \end{aligned}$$

These marginal probabilities contain no information about the relationships between the variables. Only the joint probabilities do.

Often, one variable, say Y, is taken to be a response and the other, an explanatory variable.

In other words, Y is random, but X is fixed, so that the joint distribution is no longer meaningful. Such will be the case in a clinical trial.

The distribution of Y for fixed X, with probabilities

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i.}}$$

is called the *conditional distribution*.

Then, we wish to compare the conditional distribution of Y at various levels of the explanatory variable, X.

The maximum likelihood estimates can be shown to be

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n_{..}}$$

for the joint distribution,

$$\begin{aligned} \hat{\pi}_{i.} &= \frac{n_{i.}}{n_{..}} \\ \hat{\pi}_{.j} &= \frac{n_{.j}}{n_{..}} \end{aligned}$$

for the marginal distributions, and

$$\hat{\pi}_{j|i} = \frac{n_{ij}}{n_{i.}}$$

for the conditional distribution.

Example

In an American social survey, people were asked about their opinions on the death penalty and gun registration, with the following results:

	Death Penalty	
Gun Registration	**Favour**	**Oppose**
Favour	784	236
Oppose	311	66

The maximum likelihood estimates of the joint probabilities are (0.56, 0.17, 0.22, 0.05), of the marginal probabilities, (0.73, 0.27) for gun registration and (0.78, 0.22) for the death penalty, and of the conditional probabilities, (0.77, 0.23) for those favouring gun registration and (0.83, 0.17) for those opposing it. □

2.2.2 Independence

The variables X and Y are statistically *independent* if all joint probabilities equal the product of the corresponding marginal probabilities:

$$\pi_{ij} = \pi_{i.}\pi_{.j} \qquad \forall i, j$$

This is also equivalent to

$$\pi_{j|i} \quad = \quad \pi_{.j} \qquad \forall i, j$$

Each conditional distribution of Y is equal to the marginal distribution.

Thus, the response, Y, does not depend on the fixed conditions, X, when the probabilities are the same for all of those conditions.

Example

In the death penalty example, neither variable might be taken as a response with the other fixed, so we look at the joint probabilities.

Under independence, they are estimated as (0.57, 0.16, 0.21, 0.06) as compared to (0.56, 0.17, 0.22, 0.05) given above, indicating some dependence.

In the car accident example, the type of injury might be taken as a response, given the fact that a seat belt was being worn at the time or not.

The conditional probability of a fatal accident, given that a seat belt was worn, is estimated as 0.0012 compared with 0.0098 without a seat belt.

Again, this indicates a dependence of type of accident on whether a seat belt was worn or not. □

2.2.3 Comparison of Probabilities

All estimates involved in the comparison of probabilities can be obtained directly from the maximum likelihood estimates of the probabilities, due to their invariance property.

Differences

For the conditional probabilities, any two rows of the table can be compared by taking the appropriate differences of probabilities: $\pi_{j|i} - \pi_{j|i'}$ for rows i and i'.

Such differences must lie between -1.0 and 1.0. If all differences are zero, the conditional probability distributions are identical and the two variables are independent.

The drawback of this rather intuitive approach is that a difference in probabilities of given size may have greater importance when the proportions are close to the limits, 0 or 1, than in the middle, near 0.5.

Relative Risk

The ratio of conditional probabilities under different conditions is known as the *relative risk*, $\pi_{j|i}/\pi_{j|i'}$, which can take any nonnegative real value.

If all relative risks are equal to unity, the variables are independent.

Relative risks will differ depending on which variable is taken as response and which as explanatory.

Thus, it is not appropriate in situations where there is no such distinction among the variables.

Example

In the car accident example, the relative risk of a fatal accident is estimated as

$$\frac{\frac{1601}{1601+162527}}{\frac{510}{510+412368}} = 7.90$$

when not wearing a seat belt as compared to wearing one, while that of a nonfatal accident is

$$\frac{\frac{162527}{1601+162527}}{\frac{412368}{510+412368}} = 0.99$$

Nonseat belt wearers have a higher risk of a fatal accident than seat belt wearers, but not of a nonfatal accident.

This indicates a dependence of type of accident on whether or not a seat belt was worn. □

Odds Ratio

The ratio of probabilities under the same conditions is known as the *odds*,

$$\begin{aligned} \phi_{jj'|i} &= \frac{\pi_{j|i}}{\pi_{j'|i}} \\ &= \frac{\pi_{ij}}{\pi_{ij'}} \end{aligned}$$

which can take any nonnegative real value.

The log odds is often called the *logit*.

$\phi_{jj'|i}$ is greater than unity when response j is more probable than response j' and conversely.

For independence, the vector of odds under each condition, i, must be the same.

Example

In the car accident example, the odds of a fatal as compared to a non-fatal injury is estimated to be 1601/162527=0.0099 without a seat belt and 510/412368=0.0012 with one.

Again, this indicates a dependence of type of accident on whether or not a seat belt was worn. □

Note that the estimation of the odds does not involve the marginal frequencies.

The *odds ratio* or *cross product ratio* is defined as

$$\begin{aligned}\lambda_{ij;i'j'} &= \frac{\phi_{jj'|i}}{\phi_{jj'|i'}} \\ &= \frac{\pi_{j|i}\pi_{j'|i'}}{\pi_{j'|i}\pi_{j|i'}} \\ &= \frac{\pi_{ij}\pi_{i'j'}}{\pi_{ij'}\pi_{i'j}}\end{aligned}$$

which again can take any nonnegative real value.

Degrees of dependence are measured from unity, which indicates independence.

A value greater than unity indicates the same degree of dependence, but in the opposite direction, as its reciprocal, which will be less than unity.

Thus, the ranges are not symmetric, being $(1, \infty)$ above unity and $(0, 1)$ below.

However, the odds ratio is symmetric in the variables, as can be seen from its definition in terms of joint probabilities.

Often, it is more convenient to use the *log odds ratio*,

$$\theta_{ij;i'j'} = \log(\lambda_{ij;i'j'})$$

which can take any real value and is symmetric in measuring dependence on each side of independence (at 0).

Example

In the car accident example, the estimated odds ratio is

$$\frac{1601/162527}{510/412368} = 7.96$$

and the corresponding log odds ratio, 2.075.

Both indicate a positive dependence between fatal injuries and not wearing a seat belt, i.e. that there is a much greater chance of a fatal accident without a seat belt. □

However, one major problem with any ratio of probabilities, such as relative risk and odds, is that its estimate is not defined if a denominator probability is estimated as zero.

The log odds is not defined if any probability is estimated as zero.

Sampling Distributions

In cohort and cross-sectional studies, the total number of observations to be made is usually fixed.

Thus, a multinomial distribution over all combinations of categories is appropriate. This is known as *multinomial sampling.*

In a clinical trial, the marginal distribution of the treatments is fixed.

Thus, the frequencies for each fixed value of the explanatory variables will have a multinomial distribution.

This is known as *independent* or *product multinomial sampling.*

However, when a distinction is to be made between response and explanatory variables, it usually makes sense to treat all sampling schemes as if they were product multinomial.

2.3 Characteristics of the Odds Ratio

2.3.1 Retrospective Studies

As we have seen, the odds ratio is symmetric in the variables and its estimation does not involve the marginal frequencies.

Due to these characteristics, it has a further useful property.

It can measure dependence even when the study is performed "backwards", as in a retrospective or case-control study.

There, the marginal distribution of the response variable is fixed by the design.

Example

In the myocardial infarction example, the marginal distribution of myocardial infarction is fixed by the design of the study.

The dependence of infarction on contraceptive use, as measured by the log odds ratio, is

$$\log\left(\frac{23 \times 132}{34 \times 35}\right) = 0.937$$

indicating a strong positive relationship between them. □

2.3.2 Relation to Relative Risk

For a 2 × 2 table, we have

$$\begin{aligned}\lambda_{11;22} &= \frac{\pi_{1|1}}{\pi_{1|2}} \times \frac{\pi_{2|2}}{\pi_{2|1}} \\ &= \frac{\pi_{1|1}}{\pi_{1|2}} \times \frac{1-\pi_{1|2}}{1-\pi_{1|1}}\end{aligned}$$

The first factor is the relative risk.

If the conditional probability of response one, $\pi_{1|i}$, is small for both groups, the second factor will be close to unity and the relative risk and odds ratio will be very similar.

Example

In the car accident example, the conditional probabilities of fatal injury are 0.0099 for nonseat belt wearers and 0.0012 for seat belt wearers.

The odds ratio was found to be 7.96 while the relative risk of a fatal accident is 7.90. □

This result is especially important in retrospective studies where the appropriate conditional probability estimates are not available, so that the relative risk cannot be directly estimated.

2.3.3 $I \times J$ Tables

In the 2×2 table, all four possible odds ratios are simply permutations of the frequencies in the numerator and denominator.

For larger tables, a number of distinct odds ratios can be calculated.

The $(I-1)(J-1)$ *local* odds ratios

$$\begin{aligned}\lambda_{ij;i+1,j+1} &= \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}} \\ & \quad i = 1, \ldots, I-1, j = 1, \ldots, J-1\end{aligned}$$

between adjacent categories determine all possible odds ratios and contain all of the information in them.

However, the construction of a minimal set of odds ratios is not unique.

Another possibility would be to make comparisons with the first category:

$$\lambda_{11;ij} = \frac{\pi_{11}\pi_{ij}}{\pi_{1j}\pi_{i1}} \qquad i = 2, \ldots, I, j = 2, \ldots, J$$

2.4 Tests

2.4.1 Goodness of Fit

If the statistician has some specific model in mind for the data, its *goodness of fit* can be tested.

Any of the statistics discussed in Chapter 1 might be used. The deviance gives a *likelihood ratio test* and the score the *Pearson Chi-squared test.*

In the simplest cases, the complete model is known from theory, so that all probabilities can be calculated without knowledge of the data.

Example

In a genetic experiment, with two gene types, G-g and H-h, a number of *Pharbitis* plants were bred, yielding the table

	G	**g**
H	123	27
h	30	21

The theoretical probabilities are $(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$.

The deviance, using the multinomial distribution, is

$$2\left\{123\log\left(\frac{123\times 16}{9\times 201}\right)+30\log\left(\frac{30\times 16}{3\times 201}\right)\right.$$
$$\left.+27\log\left(\frac{27\times 16}{3\times 201}\right)+21\log\left(\frac{21\times 16}{1\times 201}\right)\right\}=10.61$$

Since the total number of plants is fixed, there are three degrees of freedom, corresponding to three of the four observed frequencies.

Then, the Chi-squared value is large enough to indicate significant departure from the model.

The Pearson statistic is

$$\frac{\left(123-\frac{9\times 201}{16}\right)^2}{\frac{9\times 201}{16}}+\frac{\left(30-\frac{3\times 201}{16}\right)^2}{\frac{3\times 201}{16}}$$
$$+\frac{\left(27-\frac{3\times 201}{16}\right)^2}{\frac{3\times 201}{16}}+\frac{\left(21-\frac{1\times 201}{16}\right)^2}{\frac{1\times 201}{16}}=11.14$$

giving the same conclusion. □

2.4.2 Independence

Independence is a special model which very often is of interest.

Recall that it is defined by

$$\pi_{ij}=\pi_{i.}\pi_{.j}\qquad\forall i,j$$

Although the probabilities are not completely defined by the theory, as they were in the previous section, this relationship among them is specified and places a constraint on their values.

We can proceed by estimating the required marginal probabilities from the data and using them in our deviance or Pearson statistic.

However, the degrees of freedom must be adjusted to allow for each probability estimated.

Example

For the car accident example, the deviance is 2041 and the Pearson statistic 2338.

For the death penalty and gun registration example, they are respectively 5.32 and 5.15.

For the myocardial infarction example, they are respectively 7.87 and 8.33.

In each case, two marginal probabilities are estimated so that the degrees of freedom equal one.

In all cases, the hypothesis of independence is rejected. □

2.4.3 Fisher's Exact Test

The preceding Chi-squared tests require the asymptotic assumption that the sample size is very large.

In many situations, especially where it is very costly to obtain observations or where the phenomenon under study is very rare, only small frequencies will be available in a table.

In such cases, it is often even more important to make an accurate inference about the meaning of the results.

Under the null hypothesis of independence, an exact distribution of the observations can be obtained by conditioning on both sets of marginal frequencies.

The result is a *hypergeometric distribution*, which, for the 2 × 2 table, may be written

$$\frac{\binom{n_{1.}}{n_{11}}\binom{n_{2.}}{n_{.1}-n_{11}}}{\binom{n_{..}}{n_{.1}}}$$

Here, the only random element is n_{11} which, when the margins are fixed, determines all frequencies in the table

To obtain a test, all possible tables with the given marginal frequencies must be enumerated.

Those with probabilities at least as small as for that observed are retained and those probabilities summed to give a P-value.

Example

For the myocardial infarction example, the P-value for Fisher's exact test can be calculated to be 0.0052.

This compares with an asymptotic P-value of 0.0050 for the deviance and 0.0039 for the Pearson statistic.

The similarity among the values is not surprising, given the relatively large number of observations in this table. □

Chapter 3

Log Linear and Logistic Models

The logistic and log linear models for categorical data use respectively the binomial and Poisson distributions for regression analysis.

Hence, they are generalized linear models, using respectively the logit and the log links.

In fact, logistic regression is just a special case of a log linear model and all logistic models can be fitted as log linear models.

3.1 Log Linear Models

3.1.1 Poisson Regression

To introduce log linear models, we shall first look at the simplest case, when there is only one variable, a one-dimensional table of frequencies or counts.

Poisson regression, as the name implies, uses the Poisson distribution.

With the log link, this can be written

$$\log(\mu_i) = \sum_k \beta_k x_{ik}$$

where $\mu_i = \mathrm{E}[N_i]$ is the mean of the Poisson distribution.

In the special case of an ANOVA type situation, the x_{ik} will be indicator or *factor* variables.

If the values of the variable are numerical quantities, three simple models are possible.

The simplest, or null, model, with only $x_{i0} = 1$, fits a common mean to all categories.

The most complex, or saturated, fits a different parameter value to each category using a factor variable, or, equivalently, a series of indicator variables.

In between are situated the usual regression models, of which the most common is a simple linear Poisson regression:

$$\log(\mu_i) = \beta_0 + \beta_1 x_i$$

Note that, in terms of the mean values, this is an exponential curve:

$$\mu_i = \beta_0' \mathrm{e}^{\beta_1 x_i}$$

where $\beta_0' = \mathrm{e}^{\beta_0}$.

Since, in all models, the total number of observations is fixed, Poisson regression is equivalent to fitting a multinomial distribution.

Comparison of models is customarily performed using the deviance.

Example

Consider a study where subjects were asked to recall one recent stressful event.

The number of months prior to the study when the event occurred was recorded:

Months	1	2	3	4	5	6	7	8	9
Subjects	15	11	14	17	5	11	10	4	8
Months	10	11	12	13	14	15	16	17	18
Subjects	10	7	9	11	3	6	1	1	4

The model with a common mean for all 18 months has a deviance of 50.84 and 17 d.f., indicating a poor fit.

The log mean is estimated as 2.100 with s.e. 0.08248.

As always, the saturated model has zero deviance and zero d.f., since it fits perfectly, having a different mean for each category.

The linear regression model, where x_i is the number of months, has a deviance of 24.57 with 16 d.f., indicating a reasonable fit and a very significant improvement over the null model.

The parameters are estimated as $\hat{\beta}_0 = 2.803$ and $\hat{\beta}_1 = -0.08377$ showing that the number of subjects recalling an event decreases over time. □

3.1.2 Two-way Tables

In a two-way table, we have two categorical variables which must be related to the mean.

Thus, the saturated log linear model for a two-way table may be written as a Poisson regression:

$$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

in the familiar ANOVA-style notation.

As usual, some arbitrary constraints must be placed on the parameters for them to be identifiable.

The "conventional" constraints are $\sum_i \alpha_i = 0$, etc., although very few statistical computer packages use them.

Here, we choose to set the first element to zero: $\alpha_1 = 0$, etc.

If there is no interaction between the variables, they are independent, as discussed in the previous chapter.

Thus, when $\gamma_{ij} = 0 \quad \forall i, j$, we obtain the independence model.

The deviance of this model is that given in the previous chapter.

The estimates of γ_{ij} in the saturated model are a minimal set of log odds ratios, the set produced depending on the constraints chosen.

This implies that log linear models may be fitted to data from any of the study designs described in the previous chapter, and, in particular, to retrospective studies.

The choice of which variable (or both) is the response does not affect the estimate of the interaction log odds ratio parameter.

Example

For the myocardial infarction example of the previous chapter, the saturated model yields parameter estimates, $\hat{\mu} = 3.135$ (s.e. 0.2085), $\hat{\alpha}_2 = 0.4199$, (s.e. 0.2684), $\hat{\beta}_2 = 0.3909$, (s.e. 0.2700), $\hat{\gamma}_{22} = 0.9366$, (s.e. 0.3302).

As expected, the value of $\hat{\gamma}_{22}$ is identical to the log odds ratio calculated in the previous chapter.

The deviance of 7.8676 with 1 d.f. for the independence model is also identical to that obtained there. □

If one or more of the variables in the table refer to measurements, as in the event recall example above, the categorical variables can be replaced by continuous ones in the interactions in the log linear model.

However, in order to fix the marginal totals, factor variables should be used for the main effects.

Example

Consider data on the number of albinos in families of different sizes.

Number of	**Size of family**			
Albinos	**4**	**5**	**6**	**7**
1	22	25	18	16
2	21	23	13	10
3	7	10	18	14
4	0	1	3	5
5	–	1	0	1
6	–	–	1	0

This is a 4×6 table, but of a special form since three categories are impossible.

These are called *structural zeroes* and should not be included in the data set when the models are fitted.

The other zeroes in the table are called *sampling zeroes* since, in another, perhaps larger, sample positive frequencies might be observed.

If we fit the independence model,

$$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j$$

the deviance is 24.326 with 12 d.f., which gives a P-value of 0.01836, so that we reject the hypothesis of independence.

If we use a linear interaction between family size and number of albinos, the model is

$$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j + \gamma x_{1i} x_{2j}$$

where x_{1i} refers to the number of albinos and x_{2j} to the family size.

For this model, the deviance is 15.774 with 11 d.f. for a P-value of 0.1497 so that the model fits acceptable well.

The interaction parameter is estimated as $\gamma = 0.2076$ (s.e. 0.07283), revealing a positive relationship between family size and albinism. □

3.1.3 Multi-way Tables

The extension to higher dimensional tables is direct, as in the classical ANOVA case.

Here, it is useful to introduce a different notation. In addition to the '+', the symbols '.' and '*' will be used.

The '+' has the usual meaning, while the '.' signifies an interaction.

The '*' is a more complex operator, with the following meaning:

$$\mathrm{W*X=W+X+W.X}$$

This will indicate a saturated model, with interaction, for a two-way table, such as that used in the previous section.

Thus,

$$\mathrm{W*X*Z=W+X+Z+W.X+W.Z+X.Z+W.X.Z}$$

is the saturated model for a three-way table, and so on.

This is known as the Wilkinson and Rogers notation.

With an increased number of variables indexing the table, the ways of choosing the response variables becomes more complex.

Thus, all variables might be taken to be responses, with no explanatory variables, as in the gun registration and death penalty example, or any smaller number down to only one response variable.

Usually, all marginal totals for the explanatory variables are taken to be fixed, so that a minimal model would be

$$\mathrm{R1+R2+R3+\cdots+X1*X2*X3*\cdots}$$

where Rn indicates a response variable and Xn an explanatory variable.

This is a model for independence among all responses and of responses on explanatory variables.

Association among responses can be introduced as R1.R2, etc., and dependence of responses on explanatory variables as R1.X1, etc.

Any necessary degree of interaction can be included.

Example

Consider a study of the dependence of delinquency on socioeconomic status and on whether the person concerned had been a boy scout.

Socioeconomic Status	**Boy Scout**	**Delinquent**	
		Yes	**No**
Low	**Yes**	10	40
	No	40	160
Medium	**Yes**	18	132
	No	18	132
High	**Yes**	8	192
	No	2	48

This is a $2 \times 2 \times 3$ contingency table.

If we let D, BS, and SS signify, respectively, the variables delinquent, boy scout, and socioeconomic status, the minimal model is

D+BS*SS

which has a deviance of 32.752 with 5 d.f.

Thus, we reject the hypothesis that delinquency is simultaneously independent of socioeconomic status and having been a boy scout.

If we introduce the dependence of delinquency on boy scout,

D+BS*SS+D.BS

the deviance is reduced by 6.882 with 1 d.f.

The parameter estimate of -0.579, corresponding to D.BS, indicates that delinquency is lower among former boy scouts.

If, instead, we introduce the dependence on socioeconomic status,

D+BS*SS+D.SS

it is reduced by 32.75 to about zero with 3 d.f.

Thus, the boy scout variable is no longer needed in the model when socioeconomic status is present. By itself, it explains differences in delinquency, but this is because it is linked with socioeconomic status.

The parameter estimates for dependence of delinquency on socioeconomic status, corresponding to D.SS, are $(0.000, 0.6061, 1.792)$, showing that nondelinquency is higher in the higher statuses. □

3.2 Logistic Models

The logistic model is a special case of log linear models when there is only one response variable, and that variable has only two categories.

The binomial distribution is used with the logit link.

3.2.1 Binary Data

This model can be more easily applied to individual data which have not been grouped into the frequencies of a contingency table than can a log linear model.

Such data are known as *binary data.*

If there are continuous variables available, having many distinct values, this individual approach will be the only one available to analyze such data, unless the values of those variables are grouped into a small number of categories.

The general logistic regression model is now

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_k \beta_k x_{ik}$$

and the special cases are as for log linear models.

However, in distinction to log linear models, the response variable is not included among the x_{ik}.

As we shall see, all terms included in the model are, in fact, 'interactions' with the binary response variable.

Example

Let us look at a small data set with 7 individuals, which will illustrate the relationships among the various approaches.

X1	X2	Y
1	1	0
1	2	1
1	2	0
2	1	0
2	2	1
1	2	1
1	1	1

These are individual data, not grouped into a contingency table.

For the response variable, Y, a one indicates the occurrence of the event of interest.

When we fit the independence model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mu$$

we obtain a deviance of 9.561 with 6 d.f.

However, in contrast to the case of frequency data in a contingency table, here, for binary data, the deviance gives no indication of goodness of fit.

Adding X1 reduces this deviance by 0.058, and X2 by 1.185, each with 1 d.f.

As an example, the parameter value for X2 is 1.792.

These differences in deviances are interpretable in the usual way.

The saturated model has a deviance of 6.592 with 3 d.f., in contrast to the zero deviance of saturated models in contingency tables.

The difference in deviance between the saturated and the null model is 2.969 with 3 d.f. □

3.2.2 Grouped Binomial Data

Logistic models can also be applied to the frequency data of contingency tables when there is one response variable and it is binary.

The same procedures are used as for individual binary data and the results will be identical in cases where the data could be classified into a contingency table.

Example

Our binary data example can be grouped into the following $2 \times 2 \times 2$ contingency table:

		Y	
X1	**X2**	**0**	**1**
1	**1**	1	1
1	**2**	1	2
2	**1**	1	0
2	**2**	0	1

The deviance for the null model is now 2.969 with 3 d.f., which was our difference in deviance above.

This may here be interpreted as a goodness of fit.

The same reductions in deviance are found as previously and the parameter value for X2 is again 1.792, so that all of our results are identical.

However, we can also fit this table as a log linear model.

The independence model

$$Y+X1*X2$$

gives a deviance of 2.969, as might be expected.

The parameter estimate for the term, Y.X2, in the model

$$Y+X1*X2+Y.X2$$

is 1.792 as previously.

Thus, all three approaches give absolutely identical results. □

3.2.3 Alternative Link Functions

Binary models can sometimes be interpreted as arising when some underlying continuous stimulus is present which only gives a positive response after some critical level is reached.

If this underlying continuous variable has a logistic distribution, the resulting binary response will follow a logistic regression.

The underlying continuous distribution can be altered by specifying a different link function.

The two most commonly used are the probit, corresponding to a normal distribution, and the complementary log log, for an extreme value distribution.

3.3 Ordinal Variables

3.3.1 Fixed Scales

All of the models so far presented in this chapter impose no structure on the values of the variables.

The categories can be reordered in any way without changing the results.

If a variable does have an ordering, this will lead to a loss of information.

If reasonable, the simplest approach is to assign numerical values to the categories, often just a linear scale involving the consecutive integers.

If such a scale can be derived, the methods already described can be used directly, since the variable has been promoted to being continuous.

Example

Consider the classification of schizophrenic patients in a London institution, where the types of visit are (A) goes home or visited regularly, (B) visited less than once a month and does not go home, and (C) never visited and never goes home.

Length	**Type of Visit**		
of Stay	**A**	**B**	**C**
2-10	43	6	9
10-20	16	11	18
>20	3	10	16

Here, both variables might be taken to be ordinal.

The independence model has a deviance of 38.353 with 4 d.f.

The model with a linear scale for visit and nominal for length has deviance 6.46 while that with a linear scale for length and nominal for visit has 0.02, both with 2 d.f.

This indicates that the linear scale is acceptable for length of stay, but not for type of visits allowed. □

3.3.2 The Log Multiplicative Model

The logical extension of the fixed scale model is to estimate the position of the categories on an arbitrary scale.

This model will have the form

$$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j + \gamma x_i \delta_j$$

where δ_j is an unknown scale for the ordinal variable indexed by j.

The last term of this model contains a product of two unknown parameters, hence the name, *log multiplicative* model, so that it is not a log linear model and cannot be estimated by standard software.

Example

When this model is applied to the schizophrenic data, the estimated scale is $\hat{\delta}_j = (0.00, 0.98, 1.00)$, with deviance 0.02 on 2 d.f., when length has a linear scale as above.

This indicates that the second and third categories of patients, who never go home, are similar and might be classed together.

The regression coefficient is $\hat{\gamma} = 1.63$, showing that the longer is the length, the more chance there is of the patient being higher on the visit scale. □

3.3.3 The Continuation Ratio Model

A second type of approach to ordinal variables regroups the categories of response instead of creating a scale.

It is only applicable to a response variable.

In the *continuation ratio* model, each successive category is considered in turn and the frequency of response at least up to that point is compared to that for the next higher category.

In this way, the original contingency table, with a J category ordinal scale is converted into a series of $J-1$ subtables, each with a binary categorization, lower/higher than that given point.

Since this is only a reparametrization of the multinomial distribution for the table, a standard logistic model can be applied to the reconstructed table.

Example

For the schizophrenic data, the reconstructed table is

Length	**Type of Visit**	
of Stay	**A**	**B**
2-10	43	6
10-20	16	11
>20	3	10
	A+B	**C**
2-10	49	9
10-20	27	18
>20	13	16

The logistic model

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \mu + \alpha x_i + \beta_j$$

gives a deviance of 2.69 with 3 d.f.

The parameter for length of stay is $\hat{\alpha} = -2.36$, indicating less chance of being in the lower category as the length of stay increases. □

3.3.4 The Proportional Odds Model

The *proportional odds* model, the continuation ratio model, except that the frequency up to a given point is compared to that for all points higher.

Again, a new table is constructed, but, this time, it is not a simple reparametrization of the multinomial distribution, so that the logistic model cannot be applied. Special software is required.

Example
For the schizophrenic data, the reconstructed table is

Length	**Type of Visit**	
of Stay	**A**	**B+C**
2-10	43	15
10-20	16	29
>20	3	26
	A+B	**C**
2-10	49	9
10-20	27	18
>20	13	16

This model gives a deviance of 3.55 with 6 d.f.

The parameter for length of stay is -3.05, again indicating less chance of being in the lower category as the length of stay increases. □

3.4 Square Tables

One special type of table which is frequently encountered is the square table of two or more dimensions.

This may arise, for example, in panel studies, where the same question is asked to the same people at two or more different points in time.

It is often useful for mobility and migration studies, and for changes in voter preferences.

3.4.1 Quasi-independence and the Mover-Stayer Model

One characteristic of such tables is that the frequencies on the main diagonal are usually very large.

This arises because a large majority of individuals do not change categories between time points.

In many cases, the responses would be independent at different time points if it were not for these high frequencies.

Such a model fitted without the diagonal is known as *quasi-independence*.

Two type of people may be distinguished in a given population: those who may potentially change (the movers) and those who will never change (the stayers).

This is called the *mover-stayer* model.

However, between any two time points when observations are made, some of the movers will not have changed and will be inextricable mixed up with the stayers.

Thus, we know that individuals off the main diagonal are movers. But, movers and stayers are mixed up on the diagonal.

For this reason, we estimate the model from the off-diagonal frequencies only.

The number of movers who did not move can then be estimated.

Example

A study was made of migration among four areas of Britain between 1966 and 1971. We immediately notice the large diagonal frequencies.

1971	**Central**	**Lancs.**	**West**	**Greater**
1966	**Clydes.**	**& Yorks.**	**Midlands**	**London**
Central Clydes.	118	12	7	23
Lancs. & Yorks.	14	2127	86	130
West Midlands	8	69	2548	107
Greater London	12	110	88	7712

The usual independence model gives a deviance of 19884 with 9 d.f.

When we fit the same model, but without the main diagonal, the deviance is only 4.37 with 5 d.f.

This result is somewhat surprising since it means that the arrival point is independent of the origin, and thus of the distance travelled.

The number of potential movers who did not move between 1966 and 1971 is estimated to be $(1.6, 95.2, 60.3, 154.6)$. □

3.4.2 Symmetry

One may wish to know if the probability of change between two categories between two time points is the same in both directions.

This is called the *symmetry* model.

$$\log(\mu_{ij}) = \gamma_{ij} \qquad \text{with } \gamma_{ij} = \gamma_{ji}$$

It implies that the marginal distributions are identical, instead of being fixed at the observed values, as is usually the case for log linear models.

A less demanding model is produced if the exchange is identical in both directions within the limits imposed by the observed marginal distributions.

This is known as *quasi-symmetry*:

$$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \qquad \text{with } \gamma_{ij} = \gamma_{ji}$$

On the other hand, if the marginal distributions are identical but there is not reciprocal exchange, we have *marginal homogeneity*.

This is not a log linear model, although the other two are.

Example

For the migration example, the symmetry model gives a deviance of 9.13 with 6 d.f.

Since this is an acceptable fit, the quasi-symmetry model will also fit well: 2.67 with 5 d.f.

The parameter values $(0.00, -0.55, 0.30, 1.79, 2.22, 2.01)$ indicate that the highest migration is between Lancashire/Yorkshire and London, and the lowest between Central Clydesdale and the West Midlands. □

Chapter 4

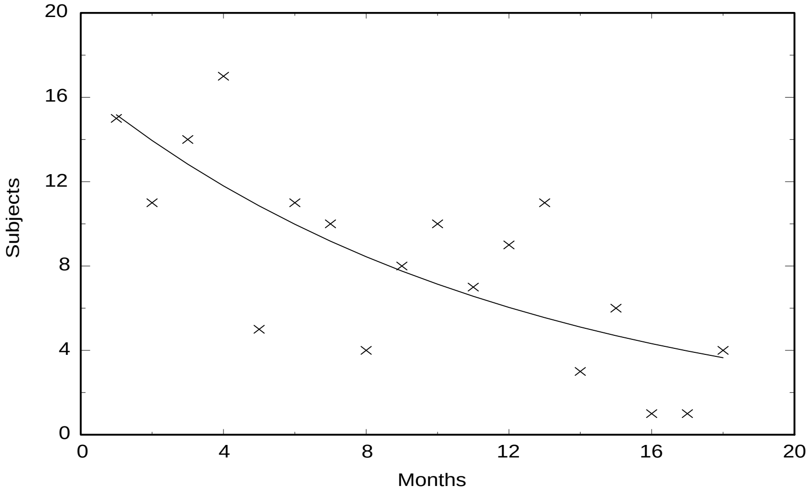
Diagnostics

A first step, where possible, is always to plot the model along with the data.

Example

Throughout this chapter, we shall take as an example the Poisson regression for the recall of events over 18 months.

We plot our fitted log linear regression model with the observations:



□

Most of the standard diagnostic techniques for normal theory models can easily be extended to generalized linear models, and, hence, to log linear and logistic models.

4.1 Residuals

In the study of departures from a model, the role of residuals is essential.

Plots of residuals are very useful in detecting departures from the model.

4.1.1 Raw Residuals

The *raw residual* for each observation is its difference from its estimated expected value:

$$\begin{aligned}\varepsilon_i^R &= y_i - \mathrm{E}[Y_i] \\ &= y_i - \hat{\mu}_i \\ &= y_i - \hat{y}_i\end{aligned}$$

For categorical data, such residuals are of little use, since their variability depends on $\mathrm{E}[Y_i]$.

4.1.2 The Hat Matrix

We have

$$\begin{aligned}\varepsilon^R &= \mathbf{y} - \hat{\mu} \\ &= \mathbf{y} - \hat{\mathbf{y}} \\ &= (\mathbf{I}_n - \mathbf{H})\mathbf{y}\end{aligned}$$

so that $\mathbf{H}$ is called the *hat matrix*, since it puts the hat on y.

It is idempotent and symmetric.

For generalized linear models,

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{\frac{1}{2}}$$

where $\mathbf{W}$ is the diagonal of the information matrix for the linear predictor.

For the Poisson distribution, it contains the elements, μ_i, and for the binomial distribution, $n_i\pi_i(1-\pi_i)$.

4.1.3 Studentized Residuals

Since, in generalized linear models, the variance is a function of the mean, it is useful to standardize the raw residuals by dividing them by the standard error to obtain a *standardized studentized residual*:

$$\varepsilon_i^S = \frac{y_i - \hat{\mu}_i}{\sqrt{w_{ii}(1-h_{ii})}}$$

where w_{ii} and h_{ii} are the i^{th} diagonal elements of the weight and hat matrices.

This is also sometimes called the *standardized Pearson residual* because $(y_i - \hat{y}_i)^2/w_{ii}$ is the contribution of the i^{th} observation to the generalized Pearson (score) statistic.

4.1.4 Deviance Residuals

Standardized deviance residuals are defined as the contribution of the i^{th} observation to the (lack of fit) deviance:

$$\varepsilon_i^D = \frac{\text{sign}(\tilde{\eta}_i - \hat{\eta}_i)\sqrt{2\text{l}(\tilde{\eta}_i; y_i) - 2\text{l}(\hat{\eta}_i; y_i)}}{\sqrt{1 - h_{ii}}}$$

where $\tilde{\eta}_i$ is the value of the linear predictor, η, which maximizes the unconstrained likelihood for the data.

4.1.5 Likelihood Residuals

Another possibility is to compare the deviance for a fitted model for the complete set of observations with that when each observation, in turn, is omitted.

Since this requires a great deal of calculation, it may be approximated by

$$\varepsilon_i^L = \text{sign}(\tilde{\eta}_i - \hat{\eta}_i)\sqrt{h_{ii}(\varepsilon_i^S)^2 + (1 - h_{ii})(\varepsilon_i^D)^2}$$

a weighted average of the previous two.

4.1.6 Residual Plots

Residuals can be plotted against a variety of statistics, each providing different information about departures from the model.

In an *index plot*, the residuals are shown against the corresponding observation number.

Ordering in this way may make identification of departures from the model easier.

If the order has intrinsic meaning, for example, as the order of collection of the data in time, the plot may indicate systematic variability in this sense.

Residuals can be plotted against the estimated means or estimated linear predictor.

They may also be plotted against each of the explanatory variables.

Finally, a *normal probability plot* shows the residuals, arranged in ascending order, against an approximation to their expected values, which is given by a standard normal distribution, $\Phi^{-1}[(i - \frac{3}{8})/(n + \frac{1}{4})]$.

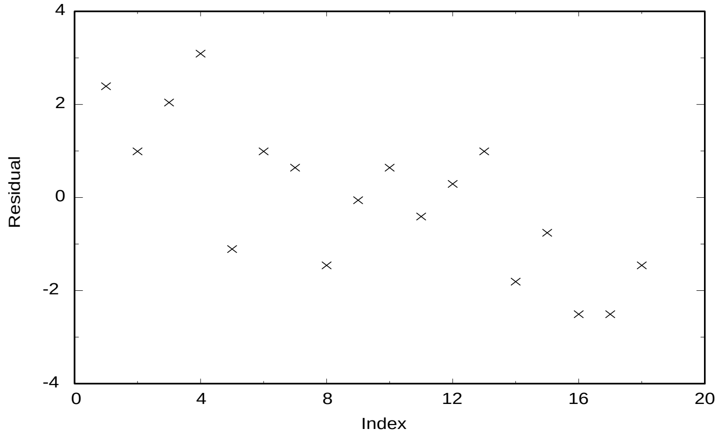
If the model fits well, this should yield a straight line at 45 degrees.

A *half-normal plot* uses the absolute values of the residuals against $\Phi^{-1}[(i + n - \frac{1}{8})/(2n + \frac{1}{2})]$.

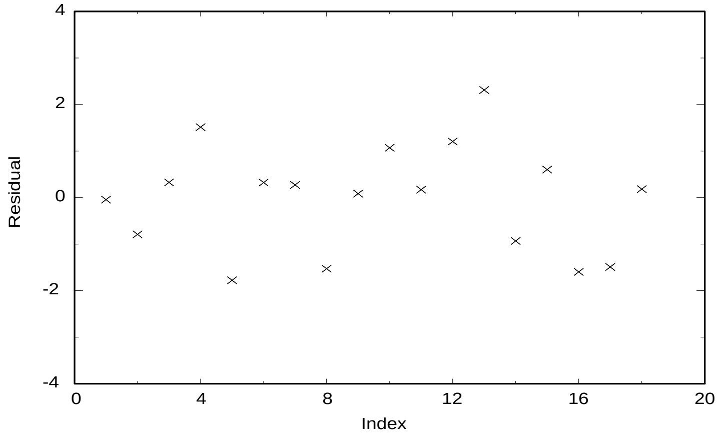
Example

Let us compare the residual plots for the regression model with those for the null model when $\beta_1 = 0$.

The index plots of the studentized residuals are, for the null model,



and, for the regression model,

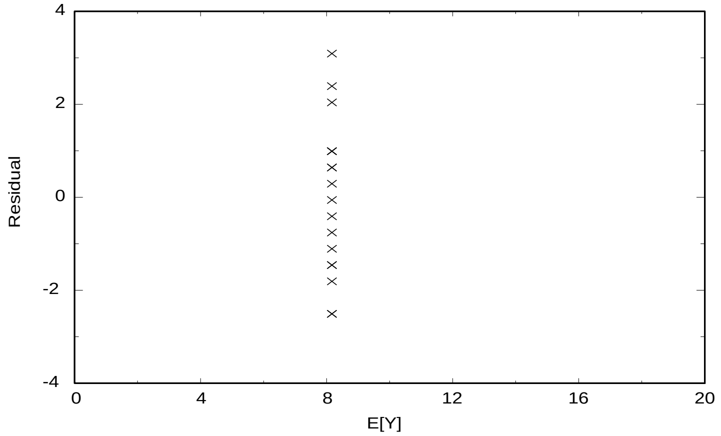


The null model shows positive residuals on the left and negative residuals on the right.

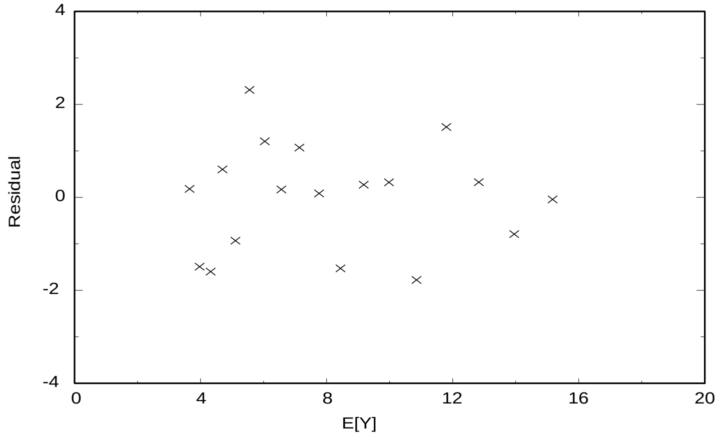
With a constant mean, early values are underestimated and later values overestimated.

The plot for the regression model shows no such trend.

The plots of residuals against the expected values are, for the null model,

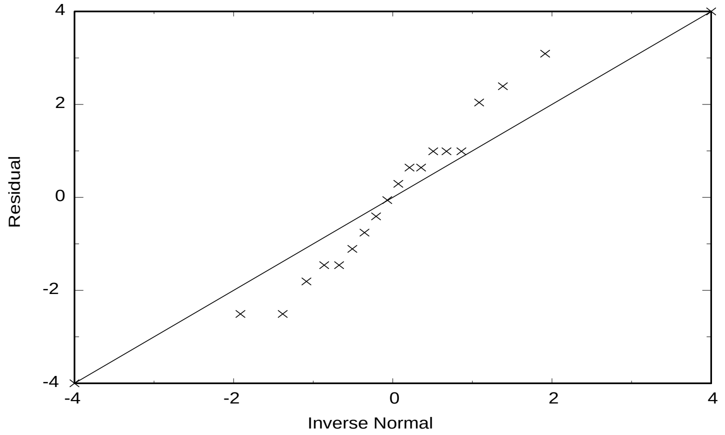


and, for the regression model,

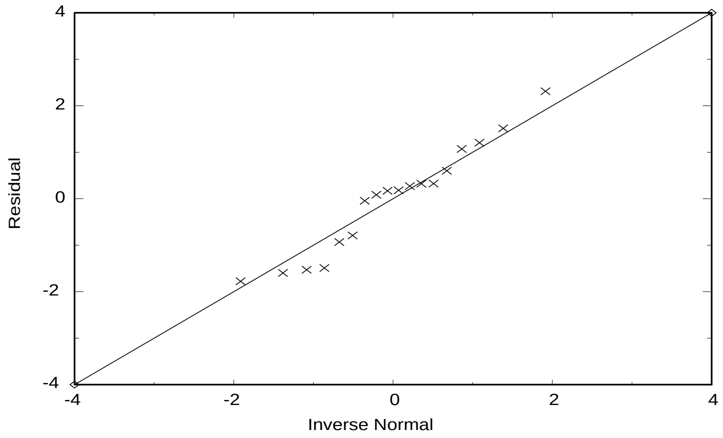


For the null model, all observations have the same estimated value. The regression model shows no abnormalities. The residual plot against the explanatory variable is identical to the index plot.

The normal probability plots are, for the null model,



and, for the regression model,



The departure from a straight line for the null model indicates that it fits poorly.

The regression model is much closer to the straight line, although it is still curved. □

4.2 Isolated Departures from a Model

When only a very few observations do not fit the model, several possibilities may be considered.

1. there may be some error in choosing certain members of the population sampled or it may not be homogeneous for the factors considered,

2. there may be some error in recording the results, either on the part of people doing the recording or transcribing it, or on the part of the respondents, not understanding a question,

3. some rare occurrence may have been observed,

4. the model may not be sufficiently well specified to account for completely acceptable observations, thus, pointing to unforeseen aspects of the phenomenon under study.

If there is no error, one will eventually have to decide if the departure is important enough to modify the model to take it into account. This will be covered in the next section.

4.2.1 Outliers

Any individual observation which departs in some way from the main body of the data is called an *outlier*.

It will not be well fitted by the model.

Outliers may be due to extreme values of the random variable (the response) or of one or more of the explanatory variables.

The likelihood that the i^{th} observation is an outlier is obtained by fitting the model without that observation. This yields a reduction in deviance for the possibility that it is an outlier, which can be approximated by

$$\varepsilon_i^O = \sqrt{(1 - h_{ii})(\varepsilon_i^D)^2 + h_{ii}(\varepsilon_i^S)^2}$$

However, in complex situations, it is rarely wise simply to eliminate an outlier, unless it is known to be an error.

Eliminating one outlier and refitting the model will quite often result in a second outlier appearing, and so on.

It is usually preferable, either to find out why the model cannot easily accommodate the observation or to accept it as a rare value.

4.2.2 Influence

An *influential* observation is one which, when changed by a small amount or omitted, will modify substantially the parameter estimates of the model.

It is an observation which may have undue impact on conclusions from the model.

However, it may not be an outlier, in the sense that it may have a small residual.

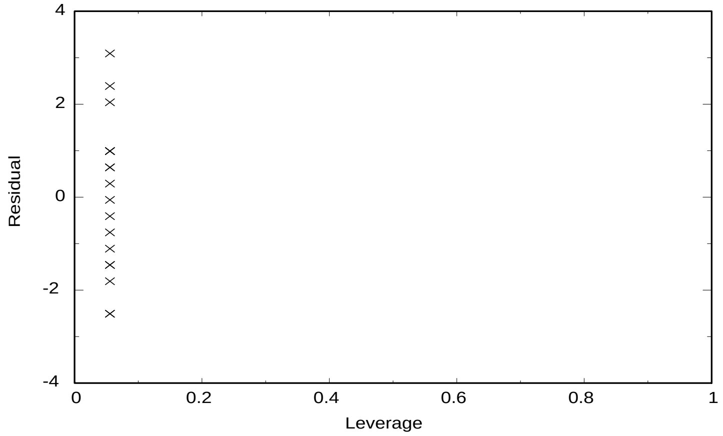
Leverage is an indication of how much influence an observation has.

A measure of leverage is the diagonal element of the hat matrix, h_{ii}, since it is the effect of the observation, y_i, in the determination of $\hat{\mu}_i$.

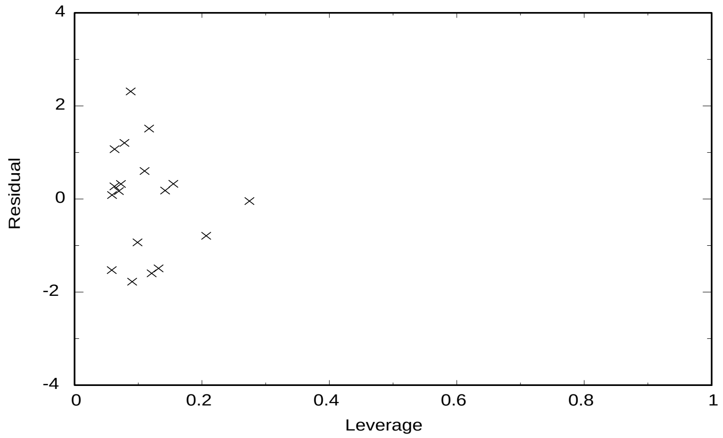
It is a measure of the distance of that observation from the remaining ones.

Example

The plots of residuals against leverage are, for the null model,



and, for the regression model,



All points have the same leverage in the null model.

No points show large leverage in the regression model.

All of the residual plots seem to point to the regression model fitting well.□

Cook's statistic is used to examine how each observation affects the complete set of parameter estimates.

The parameter estimates, with and without each observation, are compared using

$$\mathrm{C}_i = \frac{1}{p}(\hat{\beta} - \hat{\beta}_{(i)})^T \mathbf{X}^T \mathbf{W} \mathbf{X}((\hat{\beta} - \hat{\beta}_{(i)})$$

where $\hat{\beta}_{(i)}$ is the parameter estimate without the i^{th} observation.

This statistic measures the squared distance between $\hat{\beta}$ and $\hat{\beta}_{(i)}$.

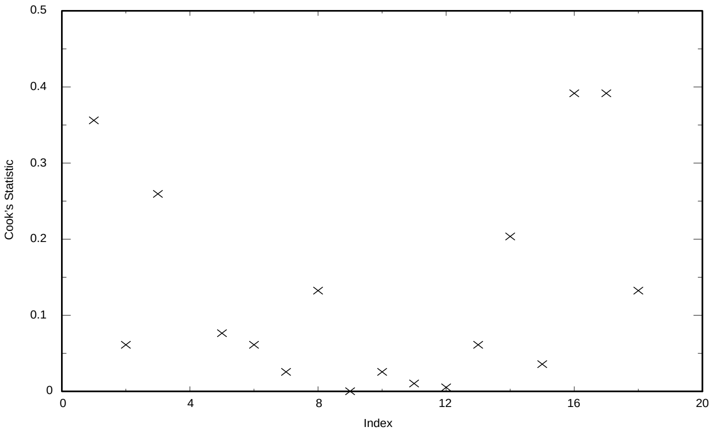
This can be approximated by

$$\mathrm{C}_i \doteq \frac{h_{ii}(\varepsilon_i^S)^2}{p(1 - h_{ii})}$$

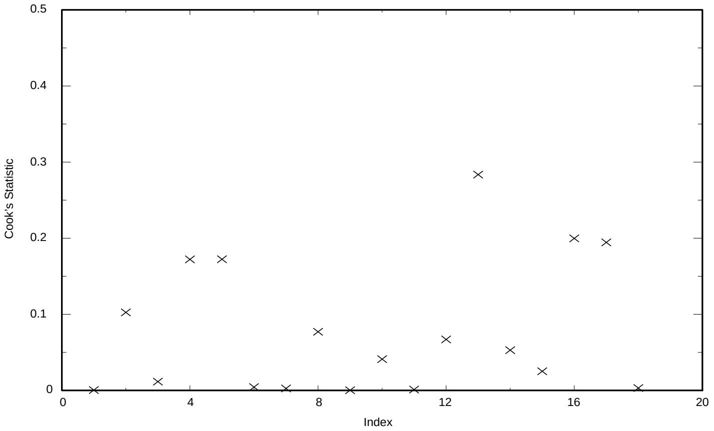
These are most usefully presented as a plot against index values.

Example

The index plots of Cook's statistic are, for the null model,



and, for the regression model,



We see that observations 1, 16, and 17 influence most the parameter estimates for the null model and 13 for the full model. □

4.3 Systematic Departures from a Model

Systematic departures from a model can often be detected from the residual plots already described.

Certain patterns will appear when the residuals are plotted against some other statistic.

Misspecification of a model may come about in a number of ways:

1. an incorrect probability distribution (for example, overdispersion),
2. an incorrect specification of the way in which the mean changes with the explanatory variables,
   - the systematic component may be misspecified,

- the link function may not be appropriate,

3. missing variables,

4. incorrect functions of the explanatory variables in the model or not enough such different functions,

5. dependence among the observations, for example over time.

These can be verified by fitting the appropriate models and comparing the likelihoods.

Sometimes, the appropriate score statistics are checked or plotted, since they do not involve actually fitting the new, more complex model.

For categorical data, two of the most important things to verify are overdispersion and the appropriateness of the link function.

Bibliography

[1] Agresti, A. (1984) **Analysis of Ordinal Categorical Data**. New York: John Wiley.

[2] Agresti, A. (1990) **Categorical Data Analysis**. New York: John Wiley.

[3] Aickin, M. (1983) **Linear Statistical Analysis of Discrete Data**. New York: John Wiley.

[4] Andersen, E.B. (1980) **Discrete Statistical Models with Social Science Applications**. Amsterdam: North Holland.

[5] Andersen, E.B. (1990) **Statistical Analysis of Categorical Data**. Berlin: Springer Verlag.

[6] Everitt, B.S. (1977) **The Analysis of Contingency Tables**. London: Chapman & Hall.

[7] Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975) **Discrete Multivariate Analysis: Theory and Practice**. Cambridge: MIT Press.

[8] Christensen, R. (1990) **Log-Linear Models**. Berlin: Springer Verlag.

[9] Collett, D. (1991) **Modelling Binary Data**. London: Chapman & Hall.

[10] Cox, D.R. (1970) **The Analysis of Binary Data**. London: Methuen.

[11] Cox, D.R. & Snell, E.J. (1989) **The Analysis of Binary Data**. London: Chapman & Hall.

[12] Fienberg, S.E. (1977) **The Analysis of Cross-Classified Categorical Data**. Cambridge: MIT Press.

[13] Fingleton, B. (1984) **Models of Category Counts**. Cambridge: Cambridge University Press.

[14] Haberman, S.J. (1974) **The Analysis of Frequency Data**. Chicago: University of Chicago Press.

[15] Haberman, S.J. (1978) **Analysis of Qualitative Data**. Vol. I. **Introductory Topics**. New York: Academic Press.

[16] Haberman, S.J. (1979) **Analysis of Qualitative Data**. Vol. II. **New Developments**. New York: Academic Press.

[17] Hosmer, D.W. & Lemeshow, S. (1989) **Applied Logistic Regression**. New York: John Wiley.

[18] Knoke, D. & Burke, P.J (1980) **Log-linear Models**. Beverly Hills: Sage.

[19] Lindsey, J.K. (1973) **Inferences from Sociological Survey Data: A Unified Approach**. Amsterdam: Elsevier.

[20] Lindsey, J.K. (1989) **The Analysis of Categorical Data Using GLIM.** Berlin: Springer Verlag.

[21] Maxwell, A.E. (1961) **Analysing Qualitative Data**. London: Methuen.

[22] Plackett, R.L. (1974) **The Analysis of Categorical Data**. London: Griffin.

[23] Reynolds, H.T. (1977) **The Analysis of Cross-Classifications**. New York: Macmillan.

[24] Santner, T.J. & Duffy, D.E. (1989) **The Statistical Analysis of Discrete Data**. Berlin: Springer Verlag.

[25] Upton, G.J.G. (1978) **The Analysis of Cross-Tabulated Data**. New York: John Wiley.