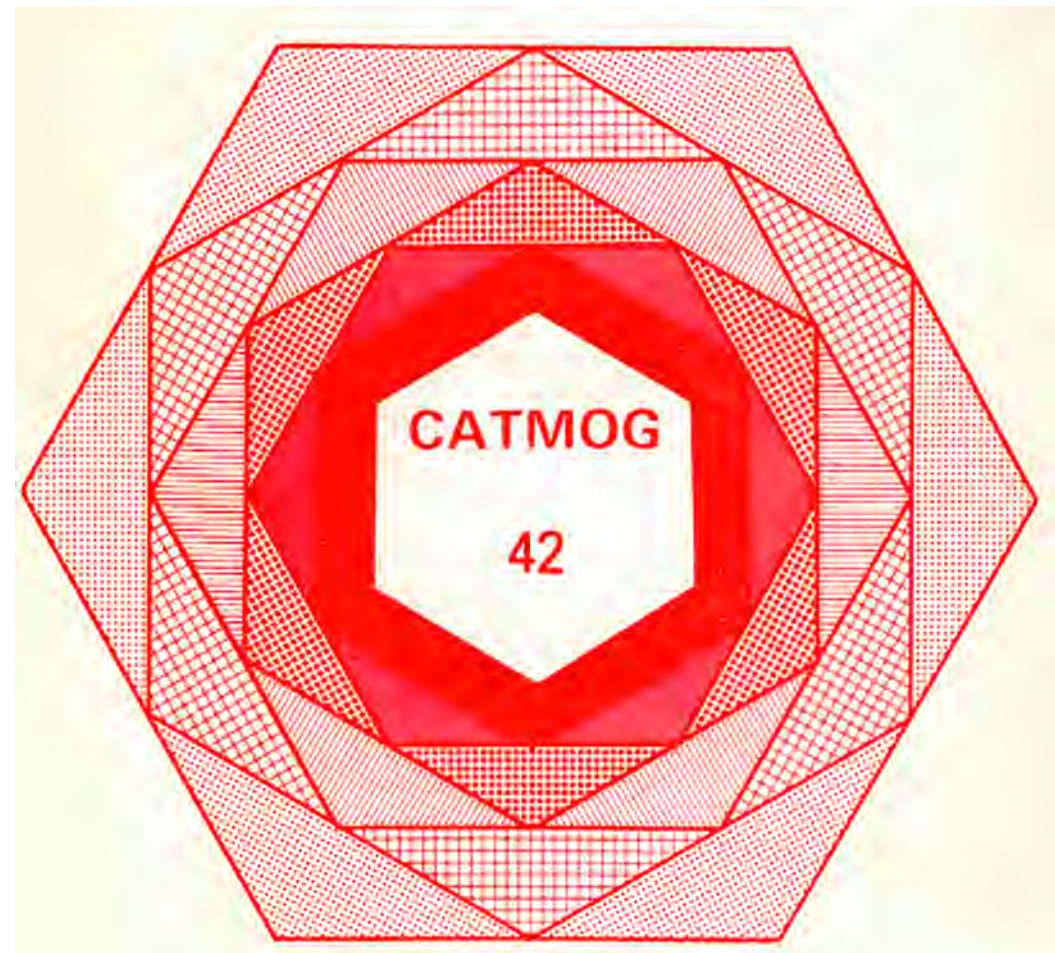


AN INTRODUCTION TO LIKELIHOOD ANALYSIS

Andrew Pickles



ISSN 0306 6142

ISBN 0 86094 190 6

© Andrew Pickles

Published by Geo Books, Norwich

Printed by W.H. Hutchins & Sons, Norwich

CAT MOG - Concepts and Techniques in Modern Geography

CATMOG has been created to fill in a teaching need in the field of quantitative methods in undergraduate geography courses. These texts are admirable guides for teachers, yet cheap enough for student purchase as the basis of classwork. Each book is written by an author currently working with the technique or concept he describes.

1.	Introduction to Markov chain analysis	- L. Collins
2.	Distance decay in spatial interactions	- P.J. Taylor
3.	Understanding canonical correlation analysis	- D. Clark
4.	Some theoretical and applied aspects of spatial interaction shopping models	- S. Openshaw
5.	An introduction to trend surface analysis	- D. Unwin
6.	Classification in geography	- R.J. Johnston
7.	An introduction to factor analysis	- 3.13. Goddard & A. Kirby
8.	Principal components analysis	- S. Daultrey
9.	Causal inferences from dichotomous variables	- N. Davidson
10.	Introduction to the use of logit models in geography	- N. Wrigley
11.	Linear programming: elementary geographical applications of the transportation problem	- A. Hay
12.	An introduction to quadrat analysis (2nd edition)	- R.W. Thomas
13.	An introduction to time-geography	- N.J. Thrift
14.	An introduction to graph theoretical methods in geography	- K.J. Tinkler
15.	Linear regression in geography	- R. Ferguson
16.	Probability surface mapping. An introduction with examples and FORTRAN programs	- N. Wrigley
17.	Sampling methods for geographical research	- C.J. Dixon & B. Leach
18.	Questionnaires and interviews in geographical research	- C.J. Dixon & B. Leach
19.	Analysis of frequency distributions	- V. Gardiner & G. Gardiner
20.	Analysis of covariance and comparison of regression lines	- J. Silk
21.	An introduction to the use of simultaneous-equation regression analysis in geography	- D. Todd
22.	Transfer function modelling: relationship between time series variables	- Pong-wai Lai
23.	Stochastic processes in one dimensional series: an introduction	- K.S. Richards
24.	Linear programming: the Simplex method with geographical applications	- James E. Killen
25.	Directional statistics	- G.L. Gaile & J.E. Burt
26.	Potential models in human geography	- D.C. Rich
27.	Causal modelling: the Simon-Blalock approach	- D.G. Pringle
28.	Statistical forecasting	- R.J. Bennett
29.	The British Census	- J.C. Dewdney
30.	The analysis of variance	- J. Silk
31.	Information statistics in geography	- R.W. Thomas
32.	Centographic measures in geography	- A. Kellerman
33.	An introduction to dimensional analysis for geographers	- R. Haynes
34.	An introduction to Q-analysis	- J. Beaumont & A. Gatrell
35.	The agricultural census - United Kingdom and United States	- G. Clark
36.	Order-neighbour analysis	- G. Aplin
37.	Classification using information statistics	- R.J. Johnston & R.K. Semple
38.	The modifiable areal unit problem	- S. Openshaw
39.	Survey research in underdeveloped countries	- C.J. Dixon & B.E. Leach
40.	Innovation diffusion: contemporary geographical approaches	- G. Clark
41.	Choice in Field Surveying-	Roger P. Kirby

This series is produced by the Study Group in Quantitative methods, of the Institute of British Geographers.

For details of membership of the Study Group, write to the Institute of British Geographers, 1 Kensington Gore, London SW7 2AR, England.

The series is published by:

Geo Books, Regency House, 34 Duke Street, Norwich NR3 3AP, England, to whom all other enquiries should be addressed.

AN INTRODUCTION TO LIKELIHOOD ANALYSIS

by
Andrew Pickles
(Northwestern University)

	CONTENTS	
Section I	INTRODUCTION	Page 3
	(i) Purpose	3
	(ii) Prerequisites	3
	(iii) Chance Variation in Social and Geographic Data	4
	(iv) Statistical Models and Explanation	5
Section II	LIKELIHOOD AND LIKELIHOOD METHODS : SINGLE PARAMETER MODELS FOR BINARY DISCRETE DATA	8
	(i) The Binomial Model - Holiday Homes in West Wales	8
	(ii) Likelihood, Log-likelihood and Maximum Likelihood	9
	(iii) Relative Likelihood, Relative Log-Likelihood Confidence Intervals and Likelihood Ratio Test Statistics	11
	(iv) Sample Size and Additivity of the Log-likelihood	13
	(v) Sufficient Statistics	14
	(vi) Extending the Example	14
Section III	MODELS WITH SEVERAL PARAMETERS	17
	(i) Transformations of the Parameter Space - The Logit Transformation of the Binomial Model	17
	(ii) The Logit Model - Travel Mode Choice in Sydney	18
Section IV	THE SCORE FUNCTION AND OTHER TEST STATISTICS	24
	(i) The Score Function	24
	(ii) The Test Statistics	25
	(iii) The LM Test, Newton Search Method and Convexity	29
Section V	LIKELIHOOD METHODS FOR CONTINUOUS DATA	30
	(i) The Exponential Model - Times Between Earthquakes in California	30
	(ii) The Normal Model - Urbanisation and Per Capita Income	32
Section VI	ANALYSIS OF COUNTS	36
	(i) The Gravity Model	36
	(ii) Distance Decay and University Attendance	38
	(iii) Contingency Tables	42
Section VII	SCIENTIFIC METHOD AND STATISTICAL INFERENCE	42
	(i) Alternative Statistical Frameworks	42
	(ii) Likelihood and Entropy	44
Section VIII	BIBLIOGRAPHY	44

$STN = \text{signal} + \text{noise}$

I. INTRODUCTION

(i) Purpose

Methods which make use of likelihood are becoming a dominant means of inference within much of the statistical research of the social and medical sciences. Yet introductory courses in statistics for geography rarely mention, let alone fully discuss, the concept or its use. This is all the more remarkable in that likelihood provides both a rigorous and intuitively appealing approach to the subject of statistical analysis, and an intuitive understanding is of the essence for the non-specialist student. In addition, likelihood provides a unifying framework within which work previously seen as diverse and unrelated by the majority of students may be seen to be intimately and logically related. Simple least squares regression, contingency tables and gravity models, for example, need not be presented as isolated models requiring different, sometimes *ad hoc*, estimation procedures and different statistics for their interpretation.

Likelihood theory is just one of several schools of thought within statistics. Whilst virtually all statisticians make use of some, often much, likelihood theory, likelihood purists are rare. The monograph presents the main elements of the theory as commonly used'. Possible inconsistencies and problems with this 'impure' approach are discussed in Section VII on alternative statistical frameworks.

A considerable amount of effort in statistics is spent not in determining the value of a particular quantity of interest but in determining the level of uncertainty that exists about this value. Much theory which allows us to estimate this level of uncertainty performs better when we have large numbers of observations rather than few, though what is implied by a large number depends upon the context. In some simple situations of a normal linear regression model, like that applied to per capita income data in the text, 30 observations may be large enough for our theory to perform well. In other situations the theory may provide poor guidance even with several hundred observations. Some caution is therefore required before applying such 'asymptotic theory, as it is known, to more complicated models such as those in the earthquake incidence or student attendance examples. In practice, where the number of observations may be quite small, some adjustments may be desirable which are specific to the particular study. All too frequently, we know little about such adjustments and tend to rely on asymptotic theory. I have deliberately chosen to do so in this introductory text to allow the generality of likelihood theory to come to the fore. Nevertheless, the reader should be aware that the methods to be described, whilst of general applicability, may be improved upon in some situations.

(ii) Prerequisites

The reader, for the most part, is not assumed to have much previous knowledge of statistics. The mathematical prerequisites extend only to simple calculus and linear algebra: An attempt has been made to explain all statistical terms as they arise and illustrations have been used for many of the basic concepts. However, to maintain brevity and fluency, some of these explanations are more suggestive of the underlying idea than exhaustive and rigorous introductions. Thus, for example, some readers may prefer to use a more standard text to obtain a more formal introduction to frequency and probability density functions.

(iii) Chance Variation in Social and Geographical Data

In statistical analysis our concern is to draw conclusions from limited information about some population of interest. A population refers to the universe or set of all possible elements or individuals which are the subject of study. In general we might like our study to be of broad relevance, suggesting that the population in question should be defined as all mankind, all towns or all earthquakes. In practice we are usually unable to observe the greater part of such populations, for many of the elements of such populations may have vanished long before records were kept, may be geographically isolated or otherwise inaccessible. Unless we can persuade ourselves that such unobservable elements would not differ materially from those that we can observe, a population is more properly defined by the set of all possible elements from which it is possible to obtain information.

The information gathered from any social or geographical survey or experiment is always likely to be subject to some, often considerable, chance variation. Such chance variation may be thought of as arising from two sources: *stochastic* and *sampling* variations.

(a) Stochastic variation

It is not unusual to observe changes in the response of an individual to a repeated survey question or changes in the behaviour of an individual in apparently identical conditions. Such variation might be explained by some hypothetical biological process of random firing neurons giving rise to truly random behaviour in the manner of a rolling die giving random responses of one to six. A more appealing explanation, however, may be found in a more critical examination of the conditions of observation. Where the experimenter or social investigator may perceive identical observational conditions, the subject may perceive subtle but important differences, differences sufficient to explain changes in response or behaviour. For example, variation in concentration, hunger and mood of the subject or the context, tone and intonation of the question being asked can all influence the responses of the same individual. When observations are made of social or physical aggregates, such as countries, the exact value of an observed variable will be subject to a myriad of small perturbations. Any particular value which theory might suggest should prevail. These perturbations are the result of the countless idiosyncratic decisions among the unobserved individuals and firms or chance events among environmental factors, each of which contributes to the eventual value of the macro or aggregate variable.

This notion of unobserved differences amongst observations is also useful in understanding the differences in response of individuals who, whilst seemingly identical in all respects to the analyst, behave or respond in a non-identical fashion. The individuals may differ in respect of "unobservable" characteristics such as "motivation" or characteristics which whilst quite measurable, such as income, were not recorded by the observer. Such considerations give rise to the second source of chance variation, that of sampling.

(b) Sampling Variation

In studying a particular population rarely is there the time or the money to allow every member of that population to be examined. More often only a proportion or sample of the members of the population can be included within

the study. Sampling describes the process whereby individuals within the population are selected for inclusion within the set of data to be analysed. Much statistical theory rests heavily upon the fact that everybody in the population should have some chance or probability of being included within the sample. Under simple random sampling each individual has an equal probability of being selected and selection itself does not disqualify the individual from being selected for a second time. The procedure for such sampling usually involves the ordering of all individuals in the population on a list (for example, an electoral register, list of premium bond holders, or a spatially ordered list such as a map) and the selection of individuals from the list by means of random numbers. The particular set of individuals studied will therefore be the result of a chance process. Due to the observed and unobserved differences between individual members of the population, this introduces chance variation into the responses obtained from the study.

(iv) Statistical Models and Explanation

The data should therefore be thought of as values from a set of random variables, which whilst sharing some common tendency, are subject to chance variation. The presence of such chance variation means that we cannot have complete confidence in the meaningfulness of the exact values obtained from the survey or experiment. It is as if we have been presented with a hard outline (our actual observations or data) which has been drawn from a blurred image (an image made up of the potential sets of data we might have drawn, some likely, others less likely). The task of statistics and statistical modelling is to identify what the underlying object in the drawing might be and to determine something about its measurements.

In univariate statistics the concern might be to obtain some guidance as to the value of a particular characteristic of the population from values of this characteristic from a sample. In multivariate statistics the concern might be to determine which of the observed characteristics of individuals and their environment influence or explain their responses or behaviour and in what manner. This may involve several steps though they may all be performed simultaneously. Firstly, especially when we have little knowledge about the data and the likely relationships to be found, an exploratory stage may be undertaken, examining simple plots and associations amongst variables. Then, with this knowledge, the performance of particular probability models are examined. In a *probability mode* the form of the stochastic or sampling variation and the structure of the relationship amongst variables are specified, though the quantitative strength of those relationships are usually represented by *parameters*, numbers of unknown value but which are to be estimated from the available data. A probability model, with parameter values set equal to some values, allows the calculation of the probability of each of the possible responses for a single individual drawn at random from the population of interest. Model estimation generally involves determining those values of the parameters which allow the model to predict the responses of a particular sample of individuals as well as possible. More formally we are concerned with

- (1) identifying an appropriate family of models
- (2) estimating values of unknown parameters and the confidence we can place in those values, on the provisional assumption that the family of models being considered is correct.

Often, at any particular moment our attention is focussed upon only a subset of the assumed relationships and parameters of the model. We may wish to test hypotheses, that is to say to compare the performance of alternative specifications for this subset of the model. The rest of the model then consists of 'maintained' hypotheses, or assumptions about the relationships and parameter values which are not immediately in question. Each of the maintained hypotheses may, in their turn, be the object of such closer scrutiny. By the examination of a series of comparisons or hypothesis tests, much may be learnt about the data.

What distinguishes a likelihood approach from other approaches is its more exclusive focus solely on the data at hand. Our objective is essentially to find a model and set of parameter values for which the data would seem a natural or unsurprising outcome. The evidence for one hypothesis over another is determined simply by their relative performance with the data at hand. This is measured by the relative probability of the data under the model specified by one hypothesis as compared to that under the alternative hypothesis. It is the probability of the data given the model which is central, and statements about absolute truths are avoided through limiting ourselves to a comparison of the performance of alternatives only. The latter is important in view of the fact that since all models are approximations only, no model can ever be true in every detail. By contrast, the more usual 'frequentist' or 'classical' approach assesses the performance of a hypothesis by an explicit consideration of the pattern of all possible data sets that might have been generated under it. If the data at hand do not appear extreme by comparison with the pattern of possible data sets, then the hypothesis is not rejected. In essence, calculations are undertaken to assess the probability of the hypothesis being correct, given the data and the model, but statements about absolute truths are side-stepped by interpreting a low probability as evidence against the hypothesis, but a high probability not as evidence for it, but simply as a lack of evidence against it. This fundamental lack of elegance within the classical approach is quickly hidden by technical terminology, but remains apparent where such basic concepts as confidence intervals are 'explained'.

Although representing profoundly different philosophical viewpoints, the procedures used and results obtained by these alternative approaches are in many cases similar. Frequently, the same decision criterion may be calculated from the data for deciding between hypotheses. However, in neither approach should it be thought that statistical analysis involves simply the application of a potentially automated process of numerical comparison of alternative hypotheses. Indeed, it is not even possible to lay down precise rules for the specification of the probability model, though Cox and Hinkley (1974) suggest that the following are important considerations:

- (1) theoretical knowledge about the process and previous empirical results
- (2) consistency with known limiting behaviour; if we know that car ownership rates decline to zero with decreasing values of household income then the model should possess this characteristic - it should not predict negative values of car ownership rate for very small values of household income

- (3) the model should be in a form such that each parameter has a clear theoretical interpretation
- (4) a model should be parsimonious; it should have as few parameters as possible, consistent with an adequate explanation of the data
- (5) the statistical theory should be as simple as possible.

It is not unusual for at least some of these considerations to lead to conflicting demands for model structure.

In the social sciences the relative weakness and informality of much theory has allowed the simplicity of the statistical theory to dominate many aspects of model selection. The potential of likelihood methods and advanced computers to extend the range of potential probability models that can be examined gives increased scope for the development of theory, relatively unrestricted by such consideration. The following sections illustrate likelihood methods for some simple models of wide interest and possessing potential for generalization to more complex situations. Several themes will be found to run through all the examples. One less obvious one is the role of the concepts of frequency and probability functions as an economical means of describing the data and the models. These are introduced here by way of example.

Consider a population in which individuals may possess one of three values, 1, 2 or 3 for a particular characteristic. A number n_1 may possess the value 1, n_2 the value 2 and n_3 the value 3. The function which tells us the number n_i , in any particular category i , ($i = 1, 2$ or 3) is called the frequency distribution function, say $f(i)$

$$f(i) = n_i \quad \text{for } i = 1, 2 \text{ or } 3$$

Such frequency distribution functions are often displayed graphically by means of histograms.

The function which tells us the number in all the categories from category 1 up to and including category i , is the cumulative distribution function, $F(i)$

$$F(i) = \sum_{j=1}^i n_j \quad \text{for } i = 1, 2 \text{ or } 3$$

If we standardize the area under the histogram to be equal to one, by dividing by the total number of individuals in all categories, N , we can obtain two new functions. The function $g(i)$, where

$$g(i) = n_i/N$$

tells us the proportion of individuals in category i . The function $G(i)$, where

$$G(i) = \sum_{j=1}^i n_j/N$$

tells us the proportion of individuals in all categories from 1 to i inclusive. Such standardized functions are referred to as a discrete probability

distribution function and a discrete cumulative distribution function respectively.

Many characteristics do not have discrete values such as 1, 2 or 3, but are measured on a continuous scale. However, similar functions to those described above for discrete variables can still be obtained.

Conceptually the measurements of the characteristic may be thought of as grouped into discrete intervals, such as years for the dwelling data of Figure 3, but then simple limit theory is used to examine the form of the function as the intervals are reduced in width to nothing. Such a process gives rise to a probability density function, such as that sketched in Figure 12, and a probability cumulative distribution function.

II LIKELIHOOD AND LIKELIHOOD METHODS: SINGLE PARAMETER MODELS FOR BINARY DISCRETE DATA

(i) The Binomial Model - Holiday Homes in West Wales

We are frequently concerned with determining how many members of a population possess a particular attribute, and how many do not. From a sample of members of the population the number of members possessing the attribute, r , must be used to make inference about the unknown number R in the population. For example dwellings sampled from a list of all dwellings in a particular area may be examined to determine the proportion of holiday homes.

For a *simple random sample* of size n from a population of size N (sampled with replacement) the *probability model* that a dwelling will possess the attribute, holiday home, is

$$p = R/N \quad (1)$$

where R/N is the proportion in the population with the attribute. The probability that an individual dwelling sampled will not possess the attribute is, of course, $(1-p)$. Table 1 gives the results from the examination of 10 hypothetical dwellings.

Table 1: Hypothetical Dwelling Survey Results

Dwelling	Holiday Home		Outcome probability
	Yes = 1	No = 0	
1	1		p
2	0		$1-p$
3	1		p
4	1		p
5	1		p
6	1		p
7	0		$1-p$
8	1		p
9	0		$1-p$
10	1		p
10	7		

The first column indicates the order in which the dwellings were observed, the second column whether it was a holiday home and the third column the probability of the observed outcome as given by the probability model. Since each dwelling was examined independently of the others the probability of observing this particular set of outcomes is found from the model by multiplying the probability of each outcome to obtain, in this case, $p \times (1-p) \times p \times p \times p \times p \times (1-p) \times p \times (1-p) \times p = p^7(1-p)^3$.

We can write a general expression for the values in column 3 by making use of an indicator variable. Let δ_i be a variable such that

$$\delta_i = \begin{cases} 1 & \text{if dwelling } i \text{ in the sample possesses the attribute} \\ 0 & \text{otherwise.} \end{cases}$$

Then the probability of any particular sample observation on dwelling i is

$$p^{\delta_i}(1-p)^{(1-\delta_i)}$$

The probability that r dwellings in the sample will possess the attribute and $(n-r)$ will not, is the product of the independent probabilities that each member of the sample is found to have or not have the attribute, multiplied by the number of possible permutations of obtaining r from n . This sample probability $P(p; r, n)$ is therefore given by

$$\begin{aligned} P(p; r, n) &= \prod_{i=1}^n p^{\delta_i}(1-p)^{(1-\delta_i)} \frac{n!}{(r!(n-r)!)} \quad (2) \\ &= p^r(1-p)^{(n-r)} \frac{n!}{(r!(n-r)!)} \end{aligned}$$

This sample probability can be seen to be a function of the data through r and n , and the parameter p , which may take on any value within certain constraints defined by the parameter space. (In this instance, since p is a probability, the parameter space is the interval bounded between 0 and 1). This is a *binomial probability model*. For a given value of the parameter p the model allows us to calculate the probability of obtaining any particular number r with the attribute from a simple random sample of size n . However, from our sample data we know the values of r and n and our interest is typically in learning what we can about the unknown value of the parameter p , the proportion in the population with the attribute.

(ii) Likelihood, Log-likelihood and Maximum Likelihood Estimation of Parameters

The sample likelihood function $L(p; r, n)$, is a function of both the parameters of the model and the data but it is usually abbreviated to $L(p)$. It describes the relative probability or odds of obtaining the observed data, for all points or values of the parameters within the parameter space, given that the model is correct. For the binomial probability model the likelihood function is given by

$$L(p) = p^r(1-p)^{(n-r)} \quad (3)$$

It may be noted that the sample probability and sample likelihood are equal for a sample size of 1, providing a method of constructing the likelihood function from knowledge of the probability process.

It is often more convenient to work with the natural logarithm of the probability and likelihood functions.

$$\log[P(p; r, n)] = r \log(p) + (n-r) \log(1-p) + \log(n! / (r!(n-r)!)) \quad (4)$$

$$\log[L(p)] = r \log(p) + (n-r) \log(1-p) \quad (5)$$

It may be noted that the likelihood and *log-likelihood* functions do not include any terms whose values are independent of the parameters, such as the factorial parts of the binomial probability model.

Sample data provide us with values of r and n from which we are to determine the most likely value and plausible range of values for the parameter p . For our example data of Table 1 the likelihood function is given by

$$L(p) = p^7(1-p)^3 \quad (6)$$

The values of this likelihood function for various values of p are given in column 3 of Table 2.

Table 2: values of the Likelihood, log-likelihood and Relative log-likelihood Functions

p	$1-p$	$L(p) = p^7(1-p)^3$	$\ell(p) = \log[L(p)]$	$\log[R(p)] = \ell(p) - \ell(\hat{p})$
0.0	1.0	0	$-\infty$	$-\infty$
0.1	0.9	7.3×10^{-8}	-16.43	-10.32
0.2	0.8	6.6×10^{-6}	-11.94	-5.83
0.3	0.7	7.5×10^{-5}	-9.50	-3.39
0.4	0.6	3.5×10^{-4}	-7.95	-1.84
0.5	0.5	9.8×10^{-4}	-6.93	-0.82
0.6	0.4	1.8×10^{-3}	-6.32	-0.21
0.7	0.3	2.2×10^{-3}	-6.11	0
0.8	0.2	1.7×10^{-3}	-6.39	-0.28
0.9	0.1	4.8×10^{-4}	-7.64	-1.53
1.0	0.0	0	$-\infty$	$-\infty$

The likelihood $L(p)$ first increases then decreases with a maximum at 0.7 of 2.2×10^{-3} . Since \log is a monotonic transformation, the *log-likelihood* $\ell(p)$ also first increases and then decreases with a maximum at the same value of p of 0.7. That value of p which gives the highest value for the

likelihood and *log-likelihood* functions is called the *maximum likelihood estimate* (ML estimate) of p , and is written as \hat{p} . The likelihood function has been plotted in Figure 1(a) and the *log-likelihood* function in Figure 1(b).

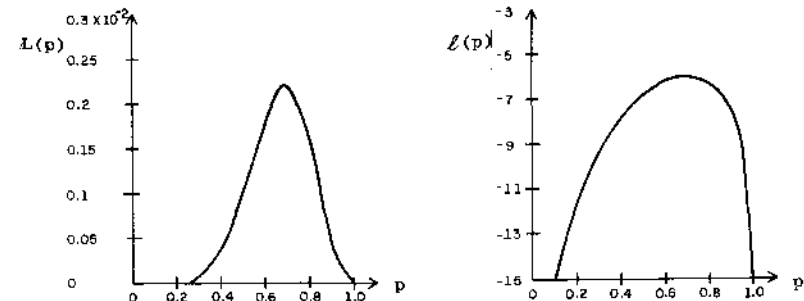


Figure 1(a) Likelihood Function 1(b) Log-Likelihood Function

The exact value of \hat{p} , may be found by solving the first order conditions for a maximum, by finding that point on the *log-likelihood* function with zero slope (the flat point just at the very top of the curve). This can be done by using simple calculus, to differentiate the *log-likelihood* function already given in Equation (5) as

$$\ell(p) = r \log(p) + (n-r) \log(1-p)$$

This gives us the equation of the slope $\ell'(p)$

$$\ell'(p) = \frac{d\ell(p)}{dp} = \frac{r}{p} - \frac{n-r}{1-p} = \frac{r - np}{p(1-p)} \quad (7)$$

This slope can only be zero when $p = r/n$. Further examination of the second derivative would ensure that this was a maximum value of the *log-likelihood* function rather than a minimum. The maximum likelihood estimate of the proportion of holiday homes in the area is $\hat{p} = 7/10 = 0.7$. In general, given a simple random sample in which the proportion possessing the attribute is (r/n) the most plausible value for the proportion possessing the attribute in the population is also (r/n) . We are consistent with commonsense!

(iii) Relative Likelihood, Relative log-likelihood, Confidence Intervals and Likelihood Ratio Test Statistics

The actual value of the *log-likelihood* function at the maximum depends upon the data, through r and n , and not just on \hat{p} . It is therefore convenient to introduce functions with standardized maxima. The *relative likelihood* function $R(p)$ is the value of the likelihood function at p , divided by its maximum value, that at \hat{p} .

$$R(p) = L(p)/L(\hat{p}) \quad (8)$$

Remembering that the likelihood function is proportional to the sample probability, $R(p)$ is the ratio of the sample probability at p to that at \hat{p} , and ranges from 0 to 1.

Around the ML (maximum likelihood) parameter estimates \hat{p} ; there will be other plausible values of p . The relative performance of these other values may be examined using the logarithm of the relative likelihood function

$$\log LR(p) = \log [L(p)/L(\hat{p})] = \lambda(p) - \lambda(\hat{p}) \quad (9)$$

This *relative log-likelihood* function has also been tabulated for the example in Table 1, and plotted in Figure 2 as curve (a). It has a maximum of 0 at \hat{p} . As values of p further from \hat{p} are considered, so the log-relative likelihood declines, reflecting the fact that the data support values of p which are close to \hat{p} rather than further away. As the log-relative likelihood declines so does the relative probability that the sample data could have been generated by the model with corresponding value of p . The log-relative likelihood measures the departure from the most likely, or the plausibility of a particular, value of p , and is itself directly interpretable. However, many statisticians argue that we should go further and calculate the probability of obtaining such a degree of departure from the most likely, using a conceptual model in which we have a large number of hypothetical samples of size n , drawn from the same population. Such a 'frequentist' approach, as it is called, argues that in the long-run the probability of obtaining a particular value of r , say r_0 , giving a log-relative likelihood equal to or lower than any specified value is given by a from the Chi-square distribution. The proof of this is complex but it concludes that

$$-2 \log [R(\frac{r_0}{n})] = -2 \log [R(p_0)] = -2 \{ \log [L(p_0)] - \log [L(\hat{p})] \} = \chi_{1,\alpha}^2 \quad (10)$$

or $\log [R(p_0)] = \chi_{1,\alpha}^2 / 2$

where $\chi_{1,\alpha}^2$ is the point on the χ_1^2 distribution (where 1 denotes 1 degree of freedom) at which the area under the distribution to the right of the value of $\log [R(p_0)]$ is equal to α . Some typical values of α and their corresponding values of χ_1^2 are shown below.

Significance Level α	0.50	0.25	0.10	0.05	0.01
χ_1^2 value	0.45	1.32	2.71	3.84	6.64

The relationship is only approximate in small samples but is often taken as the basis for 'hypothesis tests'.

In the example, to examine the hypothesis that the proportion of holiday homes in the area is actually 0.5, rather than the ML estimate 0.7, we calculate the quantity $-2 \log ER(p)$, known as the *likelihood ratio* (LR) test statistic. The hypothetical value of p , denoted by p_0 , is equal to 0.5 and with the information from Table 2 gives

$$-2 \log ER(p_0) = -2 \{ \lambda(0.5) - \lambda(0.7) \} = -2 \{ -6.93 + 6.11 \} = 1.64 \quad (11)$$

The corresponding value of α , obtained by consulting tables of the χ^2 distribution, is $\alpha = 0.20$. We would expect to obtain a sample giving a LR statistic of 1.64 or larger 20% of the time if the population proportion was actually 0.5. Thus 0.5 seems to be a reasonably plausible value for p , even though the sample proportion was actually equal to 0.7.

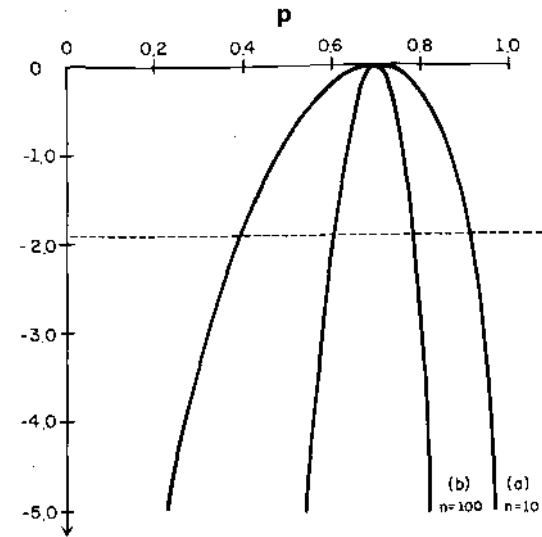


Figure 2 Relative log-likelihood Function for the Binomial Model
Curve (a) for sample size 10, 7 with attribute
Curve (b) for sample size 100, 70 with attribute

It is often convenient to be able to define a plausible region within which we have some confidence that the population parameter lies. A horizontal line drawn across Figure 2, say at -2, separates that range of parameter values which are at least 1/7 as probable as the ML parameter estimate (if $\log R(p) = -2$, $R(p) = 0.135 = 1/7$), from those less probable outside. By making use of the LR test we are able to assign to such a range of values or *likelihood interval* a level of confidence reflecting our belief that the specified range contains the true parameter value. The interval defined by the line $\log R(p) = \gamma$ is called a $100(1-\alpha)\%$ confidence interval where α is that value for which $-\gamma = \chi_{1,\alpha}^2 / 2$. For example, the interval defined by the line at -1.92 is the 95% confidence interval ($\chi_{1,0.05}^2 / 2 = 3.84 / 2 = 1.92$), and includes values of p between 0.39 and 0.92.

(iv) Sample Size and Additivity of the Log-Likelihood

With more information we would expect our knowledge of a situation to be more detailed. If instead of a sample of 10 dwellings with 7 holiday homes a sample of 100 dwellings with 70 holiday homes was available, how would our knowledge change? The log-relative likelihood function for such a sample is also plotted on Figure 2 as curve (b). Although the ML estimate \hat{p} , remains 0.7, the plausibility of other values decreases much more quickly away from \hat{p} . For any given value of α the confidence interval associated with the parameter p is reduced with increasing sample size. In the example, for $\alpha = 0.05$, the confidence interval is now $0.61 < p < 0.78$. Information from two independent samples may be examined together simply by the addition of the appropriate log-likelihood functions for each sample. For the binomial

model the joint likelihood for two samples in the example may be combined

$$L(p) = 7 \log(p) + 3 \log(1-p) + 70 \log(p) + 30 \log(1-p) \\ = 77 \log(p) + 33 \log(1-p) \quad (12)$$

(v) Sufficient Statistics

The likelihood function contains all the information about p that the sample possesses. The binomial likelihood function of Equation (3) is a function of the data, but only through n , the sample size, and r , the number in the sample possessing the attribute. No other information from the sample, for example, the sequence or order of finding those with the attribute amongst those without, is of use. The numbers n and r are referred to as *sufficient statistics* and they embody all our knowledge about the parameter p . Sufficient statistics are quantities calculated from the data which completely define the likelihood function and which may be used as the basis of all subsequent inference. The analyst may, therefore, ignore all other aspects of the data except those that contribute to the value of the sufficient statistics, provided that the model assumed is correct.

Information within the sample, not included in the likelihood function, is often termed *ancillary statistics*. Ancillary statistics may be used within a likelihood approach for model criticism. For example, the binomial probability model assumes each observation to be independent, that the finding of one dwelling to be a holiday home does not affect the probability that the next dwelling will be one. For the binomial model the sequence of finding those dwellings which are holiday homes and which are not is an ancillary statistic. If we had obtained a sequence of 70 holiday homes followed by 30 first homes we might begin to suspect some inconsistencies between the assumptions of the model and the data

(vi) Extending the Example

Concern about the presence of holiday homes stems from their possible deleterious impact on local housing markets, community life and, in certain areas, the indigenous language. Not only is the current proportion of holiday homes a factor of importance, but also the rate of turnover of property ownership, since this provides information about how long into the future such properties may remain unusable for local housing. Table 3 gives information on durations of ownership of holiday homes in Cemaes, West Wales (Davies, 1983).

Table 3: Duration of Holiday Home Ownership (Davies, 1983)

Number of Dwellings	Duration of Holiday Home Tenure (years)																Dwelling Years	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16		
Number sold	124-	10	14	19	14	17	7	11	9	7	4	3	4	2	3	0	0	679
Number Still Owned	343	39	14	38	32	44	29	17	25	21	17	17	21	9	6	2	12	2083

The first row contains information for those dwellings for which both the date of purchase and eventual sale are known. The frequency recorded for each duration of tenure has been plotted in Figure 3.

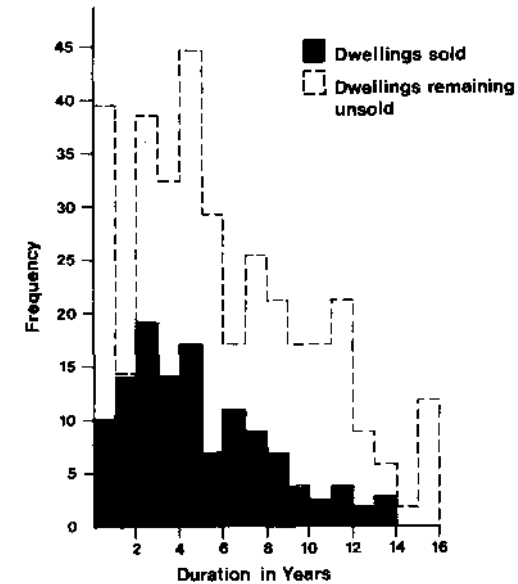


Figure 3 Histogram of Duration Frequency for Dwelling Data of Table 3

However, for many dwellings only the date of purchase is known, so that the observed durations are incomplete. Intuitively we would expect proportionately more of the longer durations to remain incomplete than the shorter, an expectation which is supported by Figure 3. This implies that a simple analysis of only those dwellings which have been sold will be biased towards those sold more quickly.

We need to analyse both completed and incomplete durations and to do this we must make use of a probability model. The simplest model is one in which the probability of sale, p , is constant for all dwellings and over all time periods. Consider a randomly sampled dwelling i observed as sold after a duration of tenure of t_i years. The probability of the observation is given by the model as the probability of not being sold, $(1-p)$, for each of the first t_i years, times the probability, p , of being sold in the $t_i + 1$ th year. Thus each dwelling from the first row of Table 3 has a probability of occurrence or likelihood $L_i(p)$ of the following form

$$L_i(p) = p(1-p)^{t_i} \quad (13)$$

The dwellings in the second row are observed to have simply 'survived without sale for t_j years. Their likelihood is therefore given by

$$L_j(p) = (1-p)^{t_j} \quad (14)$$

If the sale of each dwelling can be assumed independent, the likelihood for the whole sample is obtained, as before, by multiplying the likelihood of each observation

$$L(p) = \prod_i p(1-p)^{t_i} \times \prod_j (1-p)^{t_j} \quad (15)$$

giving a log-likelihood

$$\ell(p) = \sum_i [\log(p) + t_i \log(1-p)] + \sum_j [t_j \log(1-p)] \quad (16)$$

The first term in these functions is for completed durations and the second for incomplete durations. In Figure 4 the relative log-likelihoods for the complete and incomplete durations are plotted separately, together with that for the data combined. The peaks of the functions, and the corresponding ML estimates of the rate of sale, for the complete and incomplete durations are clearly displaced from that for the combined data. Indeed the maximum for incomplete durations data occurs on a boundary of the parameter space at $p = 0$. The displaced maxima illustrate the bias that would have occurred had we not used the probability model to allow us to combine the data.

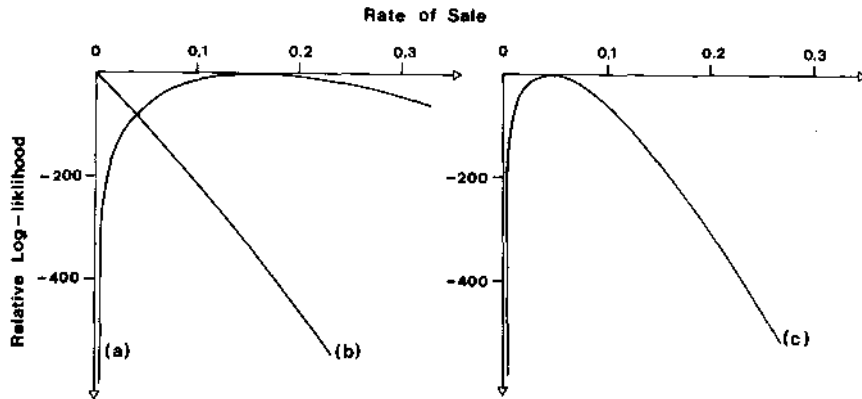


Figure 4 Relative Log-likelihood Functions for (a) Complete, (b) Incomplete and (c) All Durations

Equation (16) for the log-likelihood can be further manipulated to give us the following

$$\ell(p) = \sum_i \log(p) + (\sum_i t_i + \sum_j t_j) \log(1-p) \quad (17)$$

This is just the simple binomial likelihood function that we began with where r is given by the total number of dwellings sold during the observation period and n is given by the total number of dwelling years observed (including the year in which a sale took place). We already know that the ML estimate of p for such a model is r/n . Thus the ML estimate of the probability of sale each year, \hat{p} is obtained from Table 3 as

$$\hat{p} = \frac{\text{number of sales}}{\text{number of dwelling years}} = \frac{124}{679 + 2083 + 124} = 0.043 \quad (18)$$

II. MODELS WITH SEVERAL PARAMETERS

The reader will have been aware that the models so far introduced have at best been highly simplified approximations of reality. In general the probability of occurrence, that is the parameter p of the binomial model, cannot be assumed constant for all individuals in the sample or over all time periods. Fortunately it is a relatively straightforward matter to extend the model to allow the probability of occurrence to vary with the value of characteristics of the individual and their environment. In the previous example the probability of sale may depend upon the condition and location of the dwelling or changing holiday taking habits and money interest rates. Indeed the determination of the possible importance of such causal variables is of prime interest. The obvious path for developing the binomial model is to make the parameter p , the mean of the distribution, a function of these variables. Before examining this development a transformation of the parameter space is introduced.

(i) Transformations of the Parameter Space - The Logit Transformation for the Binomial Model

Any probability model and likelihood function may be written in terms of a variety of alternative parameters. For example, the binomial model may be re-parameterised by the use of a *logit transformation* with parameter θ where

$$p = \frac{\exp(\theta)}{1 + \exp(\theta)} \quad \text{and} \quad 1 - p = \frac{1}{1 + \exp(\theta)} \quad (19)$$

or

$$\text{logit}(p) = \log\left[\frac{p}{1-p}\right] = \theta$$

The logit transformation is illustrated in Figure 5, where as θ varies from $-\infty$ to $+\infty$, p varies from 0 to 1.

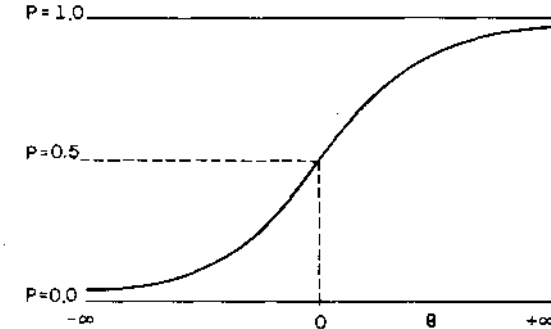


Figure 5 The Logit Function

The log-likelihood function for the binomial model in this transformed parameter space becomes on substituting for p using Equation (19) into Equation (3)

$$\begin{aligned} \ell(\theta) &= r \log\left\{\frac{\exp(\theta)}{1 + \exp(\theta)}\right\} + (n-r) \log\left\{\frac{1}{1 + \exp(\theta)}\right\} \\ &= r\theta - n \log[1 + \exp(\theta)] \end{aligned} \quad (20)$$

As before the ML parameter estimate is obtained by differentiating this log-likelihood function and finding the value of θ which makes this derivative zero. Using standard procedures for differentiating a function of a function, (Wilson and Kirby, 1980, p. 143) we obtain

$$\lambda'(\theta) = d\lambda(\theta)/d\theta = r - n(\exp(\theta)/[1 + \exp(\theta)]) \quad (21)$$

and
$$\hat{\theta} = \log[r/(n-r)] = \log[\hat{p}/(1-\hat{p})] \quad (22)$$

For the introductory example with $r = 7$ and $n = 10$, $\hat{\theta} = \log[7/(10-7)] = 0.847$, which exactly corresponds to a value of p of 0.7. The relative log-likelihood function for this logit formulation has been plotted in Figure 6. Notice that although the log-likelihood of each point along θ gives the same likelihood as the corresponding point in the p parameter space, the fact that θ and p are measured in different metrics means that the actual shape of the log-likelihood is different. Some transformations lead to log-likelihoods possessing shapes which are computationally and statistically attractive, whilst others may create severe problems. In addition different transformations of the parameter space may offer parameters of varying degrees of interest and ease of interpretation within the particular theoretical context of the analysis.

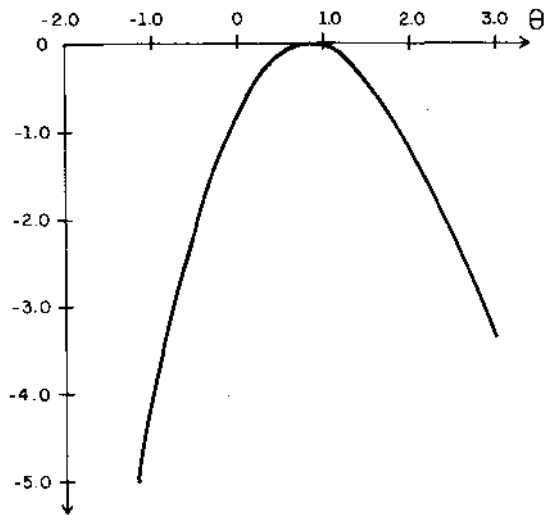


Figure 6 The Relative Log-Likelihood Function for the Binomial Model Using logit Transformation: sample size 10, 7 with attribute

(ii) The Logit Model - Travel Mode Choice in Sydney

The parameter θ of the logit function, since the parameter space is not bounded can be more simply linked to causal variables than the original (0,1) bounded binomial parameter p . In particular a simple additive function of the following form may be used in which

$$\theta_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_n x_{ni} \quad (23)$$

Thus, the value of the logit parameter for individual i , θ_i , depends upon the values of the variables x_1, x_2 to x_n and some new parameters. The parameter α , usually referred to as the constant, gives the value of θ for an individual (possibly hypothetical) with zero values for all the x variables. The β parameters, one for each variable, determines the degree to which variation in the value of each variable gives rise to the variation in θ . A value of zero for a β parameter implies that variation in the corresponding x variable has no impact on θ , and hence the occurrence probability. Large positive or negative values of β may give rise to considerable variation in θ , although this also naturally depends upon the amount of variation between individuals in the values of the x variable. The simple additive form of Equation (23) is referred to as linear in the parameters, in this case the α and β 's. The details of this kind of model are illustrated with an example from transportation analysis.

Table 4 gives information about the travel behaviour and environment of 41 work commuters to the central business district of Sydney, Australia (a subsample from Hensler and Johnson 1981). The first column is the case number. The second column is a single x -variable describing the ratio of travel cost by car to that of train from the respondents home to the central business district. The third column is a standardised cost ratio obtained from the second by the subtraction of its mean value 1.479. This is in no way a necessary step in applying the model but allows a zero value for this new x variable to represent the cost-ratio environment of a reasonably 'average' individual. The fourth column describes whether the individual commuted by car or train.

For each individual the probability of travel by car p_i is given by

$$p_i = \exp(\theta_i)/[1 + \exp(\theta_i)] = \frac{\exp(\alpha + \beta \times \text{standardised cost ratio})}{1 + \exp(\alpha + \beta \times \text{standardised cost ratio})} \quad (24)$$

We want to find the values of both α and β that give rise to values of p_i as close as possible to the observed outcome for that individual, of 1 or 0. As before we make use of an indicator variable δ_i , $\delta = 1$ if travelled by car or 0 otherwise, to write down the likelihood for individual i

$$L_i(\alpha, \beta) = p_i^{\delta_i} (1-p_i)^{(1-\delta_i)} = \left[\frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \right]^{\delta_i} \left[\frac{1}{1 + \exp(\alpha + \beta x_i)} \right]^{1-\delta_i} \quad (25)$$

and for the sample of 41 commuters

$$L(\alpha, \beta) = \prod_{i=1}^{41} L_i(\alpha, \beta) = \prod_{i=1}^{41} p_i^{\delta_i} (1-p_i)^{(1-\delta_i)} = p_1 \times p_2 \times (1-p_3) \times \dots \times p_{41} \quad (26)$$

Table 4: Travel Mode Choice in Sydney, Australia

Commuter Number	Cost Ratio	Standardised Cost Ratio	Car = 1 Train = 0	Predicted Probability
1	0.7179	-0.7610	1	.8545
2	1.0000	-0.4789	1	.7875
3	2.146	0.6671	0	.3635
4	1.935	0.4561	0	.4462
5	1.143	-0.3359	1	.7459
6	0.8421	-0.6368	1	.8275
7	1.857	0.3781	0	.4779
8	1.868	0.3891	0	.4734
9	2.000	0.5211	0	.4202
10	0.3750	-1.104	1	.9113
11	0.4762	-1.003	1	.8971
12	4.167	2.688	0	.0207
13	0.5000	-0.9789	1	.8934
14	3.667	2.188	0	.0455
15	0.4167	-1.062	1	.9057
16	1.500	0.0211	0	.6211
17	0.500	-0.9789	1	.8934
18	3.077	1.598	0	.1111
19	1.429	-0.0499	1	.6479
20	1.967	0.4881	0	.4334
21	1.389	-0.0899	0	.6627
22	0.8333	-0.6456	1	.8295
23	2.121	0.6421	1	.3730
24	0.5172	-0.9617	1	.8907
25	2.115	0.6361	0	.3753
26	1.515	0.0361	0	.6153
27	3.194	1.715	0	.0936
28	0.6154	-0.8635	1	.8741
29	2.364	0.8851	1	.2858
30	0.5333	-0.9456	1	.8881
31	0.5556	-0.9233	1	.8845
32	2.500	1.021	1	.2427
33	1.800	0.3211	1	.5011
34	2.000	0.5211	1	.4202
35	2.051	0.5721	1	.4001
36	1.020	-0.4589	1	.7820
37	1.440	-0.0389	1	.6438
38	1.061	-0.4179	0	.7704
39	0.4545	-1.024	1	.9003
40	0.5714	-0.9075	0	.8818
41	0.4000	-1.079	1	.9080

The log-likelihood function is often more convenient to use, and after substituting for p_i according to Equation (24) this is given by

$$l(\alpha, \beta) = \sum_{i=1}^{41} \{ \delta_i (\alpha + \beta x_i) - \log[1 + \exp(\alpha + \beta x_i)] \} \quad (27)$$

For a given set of data the value of this log-likelihood function varies as the value of α is changed or the value of β or both. The log-likelihood function is now a surface over the two dimensional parameter space of α and β . An isometric view of the likelihood surface is shown in Figure 7a and of the log-likelihood surface in 7b. A contour plot of the relative log-likelihood (the log-likelihood surface standardised to have a maximum of zero) is shown in Figure 7c. The figures show a hill which peaks at $\hat{\alpha} = 0.5285$ and $\hat{\beta} = -1.632$. As the 'hat' notation indicates, these are the ML parameter estimates or the most plausible values of α and β .

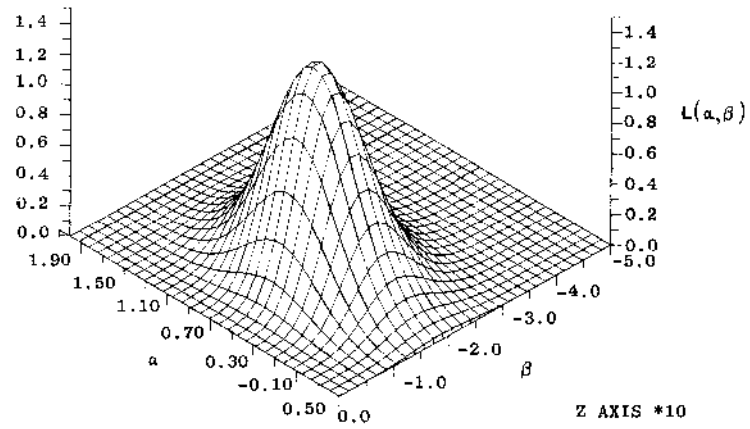
Substitution of the ML estimates of α and β into Equation (24) gives a predicted probability of choosing to travel by car for any individual with known value for the cost ratio variable. Column 5 of Table 4 gives this predicted choice probability for each individual in the sample. The variation in choice probabilities from individual to individual about the mean value of 0.6098 is the result of their cost-ratio environment. Those individuals for whom car travel is unusually expensive in comparison to train travel, tend to have a lower predicted probability of car travel, whilst those for whom it is unusually cheap have a higher probability. Summing these probabilities gives the expected number of car users, in this case 25.

If we were able to predict the travel mode of every individual exactly the value of the likelihood function of Equation (26) would be 1. Of course, that is an extraordinarily difficult task. This is reflected in practice by the very small value of the likelihood that is normally achieved. For this particular example the maximum of the likelihood function at $(\hat{\alpha}, \hat{\beta})$ is 1.4452×10^{-9} ! This might sound as if we are doing pretty poorly, though in fact it is not too bad. The way to assess how good this is is explained in the final example on page 41.

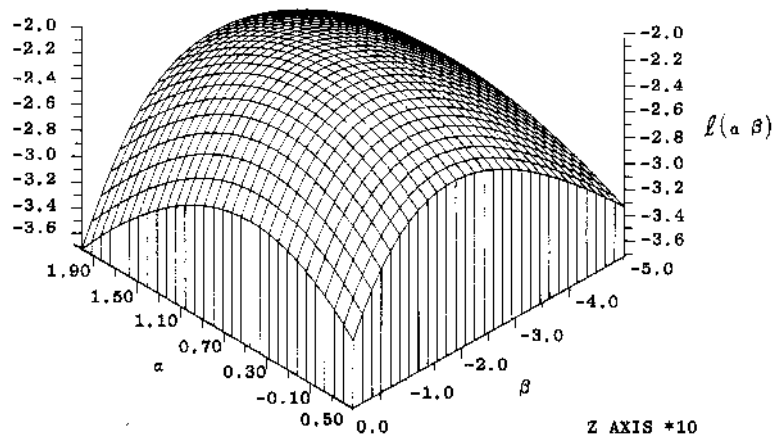
As already emphasized, according to likelihood theory the likelihood surface contains all the information useful for making inference. But the surface becomes increasingly complex, indeed increases in dimensionality, as more complex models are examined. Are there certain aspects of the surface of particular interest?

In this example we have a particular interest in the value of the parameter β , since this informs us as to how sensitive commuters might be expected to be to changes in rail fares or gasoline cost. To begin with, is the value $\beta = 0$, the value which would prevail if commuters were completely price insensitive, a plausible value? To answer this we examine forms of the model for which β is constrained to equal 0. Such models are located on the south west (bottom left) face of the isometric plots of Figures 7(a) and (b) and on the northern (uppermost) edge of the contour plot of 7(c). From Figure 7(c) it can be seen that the highest point that can be achieved with $\beta = 0$, or price insensitivity, is -7 , which shows $\beta = 0$ is implausible.

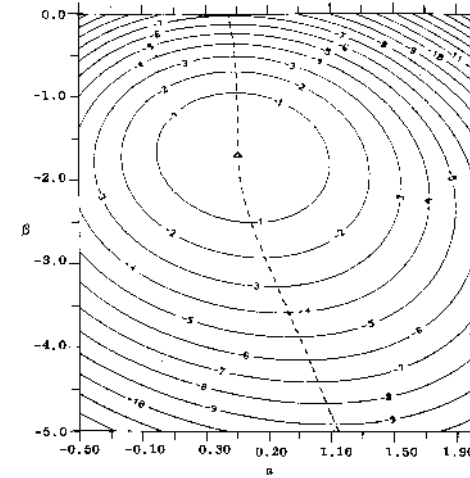
(a) LIKELIHOOD SURFACE



(b) LOG-LIKELIHOOD SURFACE



(C) RELATIVE LOG-LIKELIHOOD SURFACE



(C) PROFILE LOG-LIKELIHOOD

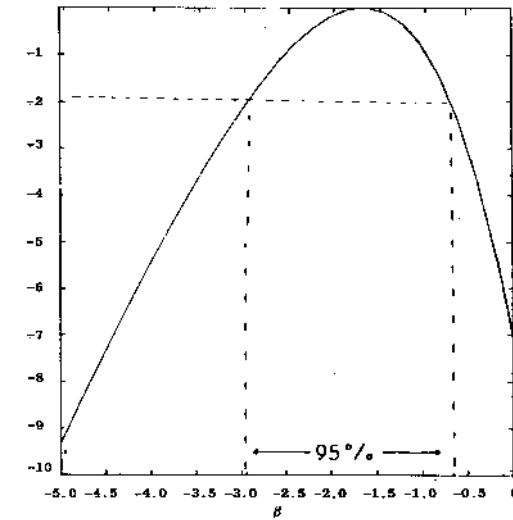


Figure 7 Travel Mode Choice in Sydney: Isometric Views of the Two-parameter Likelihood Surfaces

- (a) Likelihood Surface
- (b) Log-likelihood Surface

Figure 7(continued) Contour and Profile Views of the Two Parameter Log-likelihood Surface

- (c) Relative log-likelihood surface
- (d) Profile Log-likelihood

But what range of values of β is plausible? Assessing this is not entirely straightforward since the value of the log-likelihood achieved with any particular value of β depends upon the value of α considered. This interdependence is a very general and important problem and several methods have been proposed for the elimination of 'nuisance parameters', in this case just α , from our consideration of the parameter of interest, in this case β . In the preceding paragraph the approach was adopted of comparing the best that could be achieved with $\beta = 0$, regardless of α , with the maximum of the surface. This approach can be applied to other values of β , other than zero, and may be visualised as tracking out along the dashed line of Figure 7(c) on a path which follows the highest point of the surface for any given value of β . However, our main interest is not in the path itself but in the relative height of the path. Mathematically this can be difficult and laborious to calculate (though this can be done without too much difficulty in some statistical packages such as GLIM) but graphically simply involves an isometric view of the relative log-likelihood surface from a position without elevation and with the β axis running left to right. This could be obtained directly by suitable adjustment of the viewing position of Figure 7(b) using, for example, a cartographic computer package like Surface II (Sampson, 1976). Such a view would show the *profile log-likelihood* function (Aitken, 1982) or maximum relative log-likelihood function (Kalbfleisch, 1979). Figure 7(d) shows this function, which may be thought of as the transect obtained along the dashed path of Figure 7(c), obtained by viewing the surface from the east (right hand side). From this viewpoint the a-dimension is depth, and we are quite familiar with the notion that a profile view eliminates the depth dimension. Hence the profile log-likelihood, which will be written $\max_{\alpha} [L(\alpha, \beta)]$, provides a way of eliminating α , the nuisance parameter. A likelihood interval may then be constructed in the usual fashion as shown in Figure 7(d). This information would allow us to estimate what the 'likely' best and worst impacts of a new fares policy might be.

IV THE SCORE FUNCTION AND OTHER TEST STATISTICS

From the previous discussion the idea should be emerging that we can understand data from an examination of the likelihood, log-likelihood or 'profile' surfaces. Different models of the data are located according to the restrictions they impose on the parameters (e.g. $\beta = 0$) and their relative performance assessed by means of the relative height of the surface at these points. This concern with differences in height tends naturally to a concern with the steepness and other aspects of the shape of the surface. Can these other aspects be formalised to be of use in understanding?

(i) The Score Function

Consider some arbitrary log-likelihood function $\ell(\theta)$, a function of a single parameter θ . It is sketched below in Figure 8(a) and has a maximum at $\hat{\theta}$. The slope of the log-likelihood function, known as the *score function* is the first derivative of the log-likelihood function with respect to θ .

$$\text{Score function} = \ell'(\theta) = d \log L(\theta) / d\theta \quad (29)$$

The score function for the log-likelihood function of Figure 8(a) is sketched below in 8(b). For small θ the log-likelihood function slopes strongly upwards, indicating the score function to be large and positive. As the log-likelihood approaches the maximum at $\hat{\theta}$ so the score function falls to zero. For values of θ beyond $\hat{\theta}$, the log-likelihood slopes downwards with increasing steepness, indicating a negative and still falling score function.

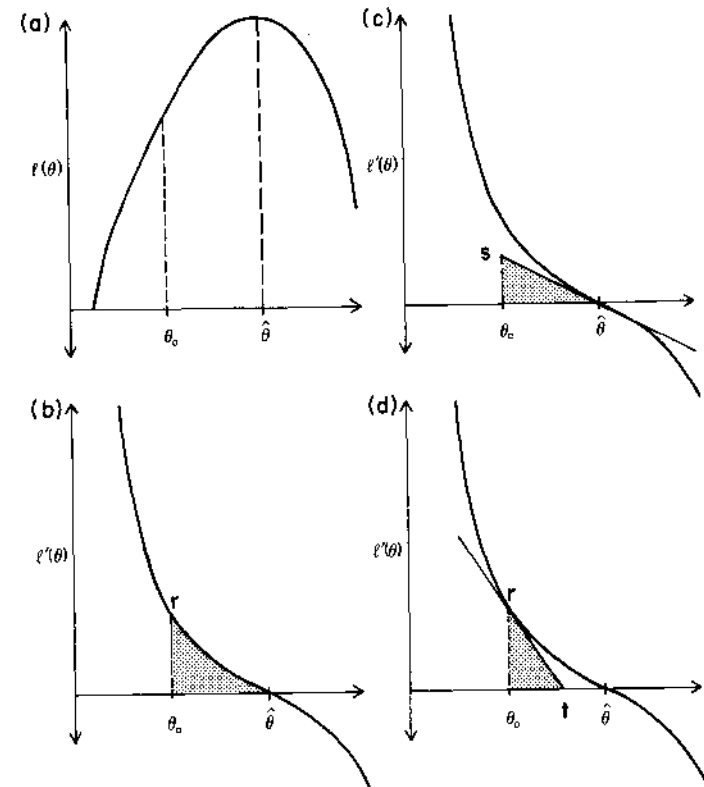


Figure 8 (a) Typical Log-likelihood Function
 (b) Score Function and Likelihood Ratio Test
 (c) Score Function and Wald Test
 (d) Score Function and Lagrange Multiplier or Score Test

(ii) The Test Statistics

The previous sections have introduced the LR statistic for testing the plausibility of two alternative values of the parameter θ , say $\hat{\theta}$ and θ_0 , in which $-2 \log [LR(\theta_0)]$ is considered as a χ^2 distributed statistic. In Figure 8(a) the value of $\log [LR(\theta_0)]$ is represented by the difference in the height of the log-likelihood function at θ_0 and $\hat{\theta}$. How is it represented in Figure 8(b)? Since integration is the reverse of differentiation we may note that

$$-2 \log R(\theta_0) = -2[\log L(\theta_0)] - \log L(\hat{\theta}) = 2 \int_{\theta_0}^{\hat{\theta}} \frac{d \log L(\theta)}{d\theta} d\theta = 2 \int_{\theta_0}^{\hat{\theta}} \lambda'(\theta) d\theta \quad (30)$$

The integral $\int_{\theta_0}^{\hat{\theta}} \lambda'(\theta) d\theta$ is represented on Figure 8(b) by the area under the score function between θ_0 and $\hat{\theta}$. Thus the LR statistic is represented by twice the almost triangular area $\theta_0, r, \hat{\theta}$.

This triangular area may be estimated in two other ways by making use of the relationships

$$\text{Area of triangle} = (\text{height} \times \text{distance})/2 \quad (31)$$

and

$$\text{height} = \text{slope} \times \text{distance}. \quad (32)$$

Just as the slope of the log-likelihood function was given by its derivative, $\lambda'(\theta)$ or score function, so the slope of the score function itself is given by its derivative $\lambda''(\theta)$, the second derivative of the log-likelihood function. Thus Equation (32) may be rewritten as

$$\lambda'(\theta) = \lambda''(\theta) \times (\theta_0 - \hat{\theta}) \quad (33)$$

The *Wald* (W) statistic is obtained from information about the shape of the log-likelihood function at $\hat{\theta}$ only. It is given by

$$W = -(\theta_0 - \hat{\theta}) \lambda''(\hat{\theta}) (\theta_0 - \hat{\theta}) \quad (34)$$

where $\lambda''(\hat{\theta})$ is the second derivative of the log-likelihood function evaluated at $\hat{\theta}$ which is the slope of score function at $\hat{\theta}$. In terms of the plot of the score function $\lambda'(\theta)$ the first term of the product is a horizontal distance. The last two terms may be considered as an estimate of the 'height' of the triangle, found by projecting the slope of the score function at $\hat{\theta}$, $\lambda''(\hat{\theta})$, through a horizontal distance $(\theta_0 - \hat{\theta})$. Thus

$$W = - \text{distance} \times \text{slope} \times \text{distance} = - \text{distance} \times \text{'height'} \quad (35)$$

To make it comparable to the LR area, which must be doubled to give the test statistic, half of the Wald statistic has been represented in Figure 8(c) as the triangular area $\theta_0, s, \hat{\theta}$.

Another test is available which makes use of information about the shape of the log-likelihood function at θ_0 only. This *Lagrange Multiplier* (LM) or *Score* test is given by

$$LM = -\lambda'(\theta_0) \frac{1}{\lambda''(\theta_0)} \lambda'(\theta_0) \quad (36)$$

It should be noted that to calculate this statistic we don't actually need to know where $\hat{\theta}$ is. For this statistic the first term is a height and the second two terms are, from re-arrangement of Equation (32) an estimate of the 'distance' $(\theta_0 - \hat{\theta})$ found by projecting the slope of the score function at θ_0 downwards through a height $\lambda'(\theta_0)$.

$$LM = - \text{height} \times \frac{\text{height}}{\text{slope}} = - \text{height} \times \text{'distance'}. \quad (37)$$

The LM test statistic is the area of the rectangle found by multiplying the height and the 'distance'. Half of this test statistic is represented in Figure 8(d) by the triangular area s, θ_0, t .

In Figure 8 it is clear that the three test statistics will give different values. If the log-likelihood function is a quadratic in the parameter θ , the score function will be linear in θ and the slope of the score function will be a constant. Under such conditions the curve of the score function in Figure 8 is a straight line and all three statistics will represent identical areas. With increasing sample size many log-likelihood functions approximate a quadratic in the region of θ . It is therefore of no surprise to learn that like the LR test, the Wald and LM tests give statistics which are asymptotically distributed as χ^2 and each test may be used to examine the plausibility of hypotheses about the parameters. The hypotheses are usually considered in terms of restrictions imposed upon one or several parameters; that they might be equal to a particular value or that parameters must maintain some fixed relationship, such as equality, amongst themselves. In practice it has been found that each test is more suited to the testing of particular kinds of hypotheses.

The Wald test is particularly convenient for testing hypotheses about any single parameter in the model, especially testing if this parameter might be zero. We may reconsider the very first example examining the proportion of dwellings which are holiday homes, using the logit framework with parameter β in Section III(i). The probability of a dwelling i being a holiday home was given by

$$p_i = \exp(\beta) / [1 + \exp(\beta)]$$

The log-likelihood function was given in Equation (20) as

$$\lambda(\beta) = r\beta - n \log [1 + \exp(\beta)]$$

which for $r = 7$ and $n = 10$ gave the ML estimate of $\beta, \hat{\beta} = 0.847$. To test the hypothesis that $\beta_0 = 0$, the Wald statistic is given by

$$W = -(\beta_0 - \hat{\beta}) \lambda''(\hat{\beta}) (\beta_0 - \hat{\beta}) \quad (38)$$

Obtaining $\lambda''(\hat{\beta})$ involves differentiating the log-likelihood function twice. The first derivative was given in Equation (21) as

$$\lambda'(\beta) = r - n[\exp(\beta) / [1 + \exp(\beta)]]$$

which differentiated again gives

$$\begin{aligned} \lambda''(\beta) &= -n[\exp(\beta)/\{1 + \exp(\beta)\}]^2 \{1 - \exp(\beta)/\{1 + \exp(\beta)\}\} \\ &= -n[\exp(\beta)/\{1 - \exp(\beta)\}]^2 \end{aligned} \quad (39)$$

Substituting $\beta = 0.847$ and $n = 10$ gives $\lambda''(\beta) = -2.1$ from which with $\beta_0 = 0$,

$$W = -0.847 \times (-2.1) \times 0.847 = 1.51 \quad (40)$$

The value of $\beta = 0$ corresponds to $p = 0.5$ in the original parameter space. Earlier the value of the LR statistic was calculated, for the hypothesis that $p = 0.5$ to be 1.64. Thus as Figure 8 suggests, in practice the different test statistics give similar but not identical values. The advantage of the Wald test is that it only requires information about the likelihood function at $\hat{\beta}$, whilst the LR test requires the value of the log-likelihood function at both $\hat{\beta}$ and β_0 . For single parameter restrictions such as in the example, where the restriction $\beta = 0$ is being examined

$$\lambda''(\hat{\beta}) = 1/\sigma_{\hat{\beta}}^2 \quad (41)$$

and $\sigma_{\hat{\beta}}$ is the standard error of the parameter estimate, a quantity provided by many computer packages. This gives from Equation (39) for $\beta_0 = 0$

$$W = \beta^2/\sigma_{\hat{\beta}}^2 \quad \text{or} \quad \sqrt{W} = \beta/\sigma_{\hat{\beta}} \quad (42)$$

The statistic \sqrt{W} (the positive root) will be distributed as $\sqrt{\chi_{1,\alpha}^2}$ which is the z or normal distribution. Thus, selecting $\alpha = 0.05$, a value of zero for a parameter would be considered implausible if the magnitude of its maximum likelihood estimate divided by its standard error was larger than $\sqrt{\chi_{1,0.05}^2} = \sqrt{3.84} = 1.96$.

whereas the W test uses information only from the maximum of the log-likelihood function (or under the unrestricted or alternative hypothesis) the LM or score test uses information from the log-likelihood function at the restricted parameter value only (or under the null hypothesis). This may be very useful if the unrestricted model is very much more complex than the restricted model.

The LM test statistic for $\beta_0 = 0$ in the example is given by

$$LM = -\lambda'(\beta_0)\{1/\lambda''(\beta_0)\}\lambda'(\beta_0) \quad (43)$$

Evaluating Equation (21) for $\beta_0 = 0$, $r = 7$ and $n = 10$ gives

$$\lambda'(\beta_0) = 2$$

and Equation (39) gives

$$\lambda''(\beta_0) = -2.5$$

Hence

$$LM = -2 \times (1/-2.5) \times 2 = 1.60,$$

a test statistic with a value between that of the Wald and LR statistics. In general the ordering of the values of the W, LR and LM test statistics varies and is a topic of current research.

(iii) The LM Test, Newton Search Methods and Convexity

For many simple problems the ML parameter estimates are given by simple functions of the data (e.g. $p = r/n$), and inference is made from examination of the likelihood function around this point. The LM test, requiring information about the likelihood function around some other point is, therefore, not currently widely used. However, in more complex problems the ML parameter estimates cannot be directly calculated but must be found by searching over the likelihood function to find those parameter values which make it a maximum. One such method, the Newton Method, is closely related to the LM test. The Newton method of search begins with an initial guess θ_0 as to the ML estimates $\hat{\theta}$, and calculates an improved guess, θ_1 according to the following equation

$$\theta_1 = \theta_0 - \lambda'(\theta_0)/\lambda''(\theta_0) \quad (44)$$

This is repeated to obtain a still better value, θ_2 given by $\theta_2 = \theta_1 - \lambda'(\theta_1)/\lambda''(\theta_1)$, and so on. As θ_i approaches the unknown value $\hat{\theta}$ so the slope $\lambda'(\theta_i)$ falls to zero with the result that θ_{i+1} becomes increasingly little different from θ_i .

This is illustrated graphically in Figure 9 and the similarity between this diagram and that of Figure 8(d) for the LM test should be immediately apparent. If a Newton algorithm (Newton-Raphson or Gauss-Newton) is started at the restricted value of the parameter θ_0 , then the LM test statistic is given by $-\lambda'(\theta_0)(\theta_0 - \theta_1)$, where θ_1 is the estimate of $\hat{\theta}$ obtained after a single cycle of the Newton procedure.

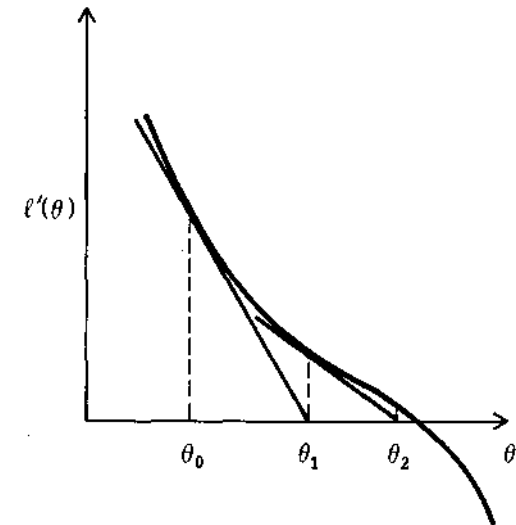


Figure 9 Newton Search Method

It is straightforward to draw more complex forms of the score function for which the Newton procedure may not always yield a solution at 6. For example, if the score function passes through zero for several different settings of the parameters, representing a log-likelihood function with several peaks, the Newton method may become trapped in a local peak (or optimum) and the analyst may remain unaware of the existence of a higher peak elsewhere. Certain conditions concerning the convexity and regularity of the likelihood function must be met before we can be sure that we can estimate the true ML parameters, those associated with the highest peak (or global optimum). This requirement for a well behaved likelihood function is in fact a general requirement for much likelihood theory and method.

V LIKELIHOOD METHODS FOR CONTINUOUS DATA

The application of likelihood methods to continuous variables such as time, height and so on requires no major modification to the procedures outlined so far for discrete data. As before, we begin by the construction of a probability model.

(i) The Exponential Model - Times Between Earthquakes

Some progress has already been made towards the analysis of duration in the example of time to sale of holiday homes in which time was 'discretised' into a series of year long intervals. Such discretisation involves some loss of information as durations which may differ by up to two years are grouped together. Where more precise duration data is available a continuous formulation may be superior.

Table 5 gives the duration in days between major earthquakes in Southern California, obvious after-shocks having been removed. In the simplest continuous time probability model, we consider that there is a constant risk of earthquake λ . Consider the probability of the next earthquake occurring in the short interval t to $t+\Delta t$ since the preceding one. This probability must be equal to the product of two elements. The first is the probability of there having been no earthquake up to time t , which we will denote by $S(t)$. The second is the probability of earthquakes in the following short interval Δt , which can be thought to be composed of a component $\lambda\Delta t$, the risk or instantaneous hazard rate of an earthquake times the length of the interval, and a component $O(\Delta t)$, representing the probability of two or more earthquakes and which can be ignored as the length of the interval becomes small. Together these give an expression for the probability of the next earthquake being in the interval t to $t+\Delta t$ as $[\lambda\Delta t + O(\Delta t)]S(t)$, which for very short intervals reduces to $\lambda\Delta tS(t)$.

An alternative viewpoint leads to another expression for the probability. The probability of no earthquakes to time t is $S(t)$. The probability of no earthquakes to $t+\Delta t$ is $S(t+\Delta t)$. Any difference between the quantities $S(t)$ and $S(t+\Delta t)$ must represent the probability of one or more earthquakes between t and $t+\Delta t$. Equating the two expressions

$$\lim_{\Delta t \rightarrow 0} \lambda \Delta t S(t) = S(t) - S(t+\Delta t) \quad (45)$$

Dividing by Δt gives

Table 5: Earthquakes of Magnitude 6.0 or Greater in Southern California Region 1912-1956

Date (d/m/y)	Interval (days)	Magnitude	Location
21/11/15	151	6.25	Callexico
23/10/16	343	7.10	Colorado Delta
21/4/18	552	6.00	Tejon Pass
23/7/23	1922	6.80	San Jacinto
29/6/25	707	6.25	Riverside
18/9/27	811	6.30	Santa Barbara
11/3/33	2001	6.00	Long Valley
30/12/34	659	6.30	Long Beach
31/12/34	1	6.50	Colorado Delta
24/2/35	55	7.10	Colorado Delta
25/3/37	760	6.00	Colorado Delta
19/5/40	1151	6.00	Terwilliger Valley
8/12/40	203	7.10	Imperial Valley
1/7/41	205	6.00	Colorado Delta
21/10/42	477	6.00	Santa Barbara
15/3/46	1242	6.50	Lower Borrego Valley
10/4/47	391	6.30	Walker Pass
4/12/48	604	6.40	Manix
21/7/52	1325	6.50	Desert Hot Springs
19/3/54	606	7.70	Ken County
24/10/54	219	6.20	Santa Rosa Mountains
9/2/56	454	6.00	Agua Blanca
		6.80	San Miguel

$$\lim_{\Delta t \rightarrow 0} \lambda S(t) = \frac{S(t) - S(t+\Delta t)}{\Delta t} \quad (46)$$

The right hand side is a simple example of obtaining a derivative by taking limits to give

$$\lambda S(t) = \frac{-dS(t)}{dt} \quad (47)$$

The solution to this differential equation is $S(t) = \exp(-\lambda t)$. Thus the probability density of an earthquake at time t in the presence of a constant earthquake hazard λ is an exponential density function $f(t)$ where

$$f(t) = \lambda S(t) = \lambda \exp(-\lambda t) \quad t > 0, \lambda > 0 \quad (48)$$

The likelihood function for analysing data, such as that of Table 5, is obtained from the probability (density) of the particular observed intervals. Thus the likelihood of a single observed interval t_1 is

$$L_1(\lambda) = \lambda \exp(-\lambda t_1) \quad (49)$$

and for a sample of n independent intervals

$$L(\lambda) = \prod_{i=1}^n L_i(\lambda) = \lambda^n \exp(-\lambda T) \quad (50)$$

where $T = \sum_{i=1}^n t_i$

The log-likelihood for the sample data is given by

$$\ell(\lambda) = n \log(\lambda) - \lambda T \quad (51)$$

and has been plotted for the data of Table 5 in Figure 10.

The ML estimate of λ , the earthquake rate, is found in the usual way by determining the value of λ which makes the derivative of the log-likelihood function zero.

$$\ell'(\lambda) = \frac{n}{\lambda} - T = 0 \quad \text{or} \quad \lambda = \frac{n}{T} \quad (52)$$

The ML estimate of the risk λ is simply the number of earthquakes divided by the total time on record.

As with the binomial model, the parameter λ of the exponential model may be made a function of variables. If earthquakes are viewed as a mechanism for the release of stress in the earth's surface, stress which builds up progressively until release, then we might suggest a model in which λ , the earthquake risk, is low immediately after a quake, but increases progressively as the time since the last quake increases. We might, therefore, examine a model in which λ is made a function of time. A suitable function, known as a *Weibull hazard*, is

$$\lambda(t) = \lambda \beta t^{\beta-1} \quad \lambda > 0, \beta > 0, t > 0 \quad (53)$$

in which the parameter β measures the rate of increase in the risk of an earthquake with time since the previous one. If $\beta = 1$, then the risk is constant, if $\beta > 1$ the risk increases with time and if $\beta < 1$ it decreases.

Figure 11 gives a contour plot of this new log-likelihood function. Our interest in plausible values of the β parameter can be explored by an examination of the -15 contour, as 2 below the maximum gives approximate 95% confidence limits. A moment's examination is sufficient to see that there is no evidence that the earthquake risk increases with duration since the preceding earthquake. It should be noted, however, that the data examined included only major earthquakes and earthquakes from several fault lines. A more thorough examination would take size, depth and fault line location all into account, within the model.

(ii) The Normal Model: Urbanisation and Per Capita Income in the U.S.

The form of any likelihood function from which all subsequent analysis proceeds, is determined by the particular form of probability density function chosen. In the previous examples this selection was based upon a relatively detailed consideration of the probabilistic aspects of the process being

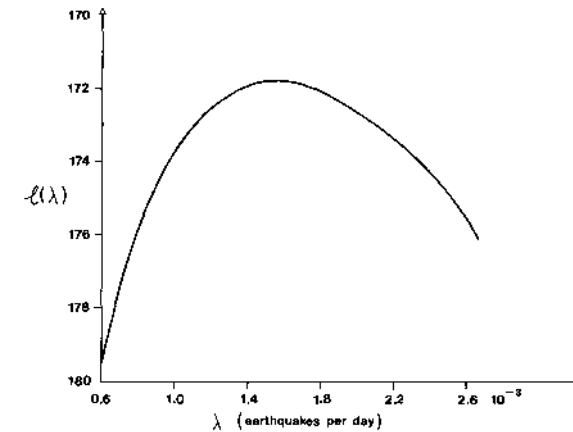


Figure 10 Earthquake Log-likelihood for Exponential Model

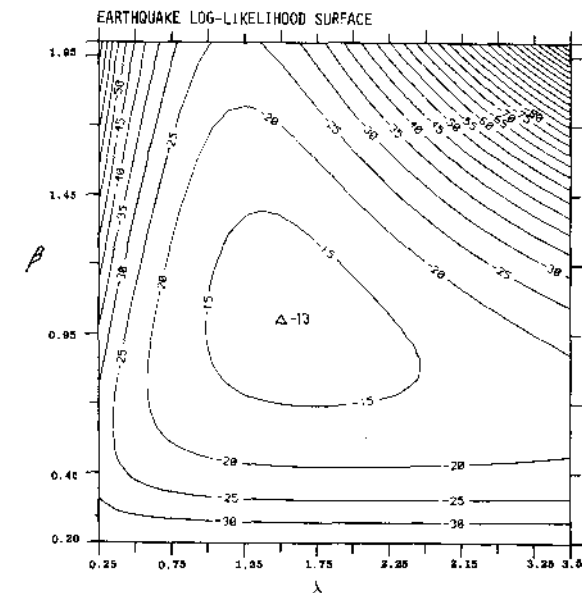


Figure 11 Earthquake log-likelihood for weibull Model

examined, which allowed us to derive appropriate density functions. Derivation of the *Normal density function* is more complex and so is not presented here. It was originally derived in the 19th century as a function to describe the effects of errors on astronomical measurements. Errors remain a problem of hardly less importance to 20th century geographers.

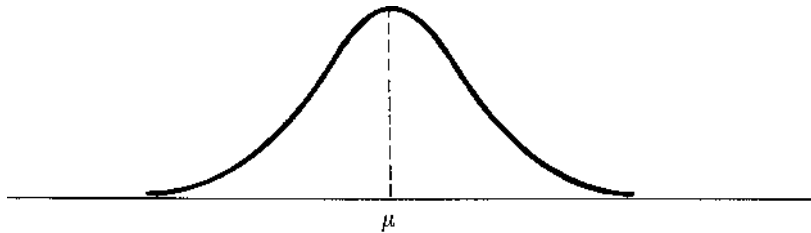


Figure 12 Normal Density Function

A normal density function $f(y)$ is sketched in Figure 12, and is defined by the following equation

$$f(y) = (2\pi\sigma^2)^{-1/2} \exp[-(y-\mu)^2/2\sigma^2] \quad \sigma > 0 \quad (54)$$

It has two parameters: μ the mean, and σ^2 the variance. Whereas the range of exponential and Weibull densities were restricted to positive numbers, the range of the Normal density is in theory $(-\infty, +\infty)$ the parameter μ locating the centre of the distribution on this range and σ^2 determining the width or dispersion. Although this density function may look more complicated than those previously encountered it in fact has some desirable mathematical features.

A typical example of its use is an analysis of the relationship between level of urbanism and per capita income. Table 6 gives the percentage of the population urbanised and the 1981 per capita income for each of the states of the coterminous U.S.. We could examine the data using a Normal density function in which the mean was made a linear function of the explanatory variables:

Expected state per capita income = $\alpha + \beta \times$ state level of urbanisation

$$\text{or} \quad \mu_i = \alpha + \beta x_i \quad (55)$$

where i refers to the i 'th state. This is an example of a Normal linear regression model. As with all the previous models we will not be able to exactly predict the value of the response variable for each state. We should, therefore, carefully distinguish between y_i , the actual value, from μ_i , the value expected from the model. The contribution to the likelihood from the data for each state is given by

$$L_i(\alpha, \beta, \sigma) = \frac{1}{\sigma} \exp[-(y_i - \mu_i)^2/2\sigma^2] \quad (56)$$

and the log-likelihood for the sample is given by

$$\begin{aligned} \ell(\alpha, \beta, \sigma) &= \log \left\{ \prod_{i=1}^n L_i(\alpha, \beta, \sigma) \right\} \\ &= -n \log(\sigma) - \sum_{i=1}^n [(y_i - \alpha - \beta x_i)^2 / 2\sigma^2] \end{aligned} \quad (57)$$

It can be seen from the second term of this equation that maximising the log-likelihood involves the minimisation of $\sum_{i=1}^n [(y_i - \alpha - \beta x_i)^2]$, or the

minimisation of the sum of squared differences between the actual and the model predicted values of the response variable y . It is relatively straightforward to derive estimates of the parameters using such a 'least squares' criterion (see Wilson and Kirby, pp 159-62).

The likelihood function itself is a four dimensional surface over the parameters, α , β , and σ . However, our interest is likely to be almost entirely with the parameter β . The plot of the profile log-likelihood $\max_{\alpha, \sigma} [\ell(\alpha, \beta, \sigma)]$

in Figure 13 provides the essential information as to the plausible values of this parameter. The data strongly suggests a positive relationship between urbanisation and income levels with a single percentage increase in the level of urbanisation being associated with approximately 65 dollars of additional income.

The desirable characteristics of the Normal density function are in part due to the fact that it gives rise to a log-likelihood function that is a quadratic in the parameters α and β . As noted in Section IV(ii) this implies that the derivative of the log-likelihood function, the score function, will be a straight line or flat plane in the α and β dimensions, and that the test statistics obtained from the LR, W or LM approaches should be identical. In more complex models using the Normal density this characteristic can be lost. Aitkin (1982) provides a more extensive discussion of likelihood inference for the Normal density.

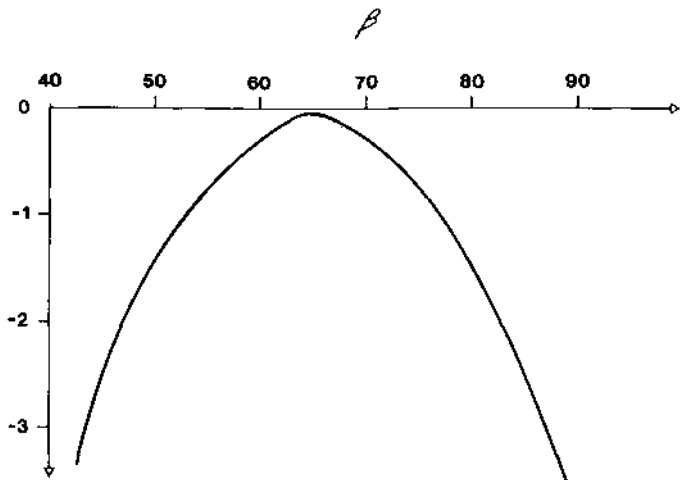


Figure 13 Profile Log-likelihood for Normal Model

VI ANALYSIS OF COUNTS

(i) The Gravity Model

The usual form of the gravity model, which attempts to explain the level of interaction between places according to their size and distance between them, may be written

$$F_{ij} = \frac{\alpha M_i^{\beta_1} M_j^{\beta_2}}{d_{ij}^{\beta_3}} \quad (58)$$

or

$$\log(F_{ij}) = \beta_0 + \beta_1 \log(M_i) + \beta_2 \log(M_j) - \beta_3 \log(d_{ij}) \quad (59)$$

where

F_{ij} is the frequency or count of the interaction between places i and j

M_i is the size of the origin i

M_j is the size of the destination j

d_{ij} is the distance i to j

and $\beta_0 = \log \alpha$

The β 's are parameters to be estimated. Such estimation has been undertaken using Normal regression methods using $\log(F_{ij})$ as the response variable and $\log(M_i)$, $\log(M_j)$ and $\log(d_{ij})$ as independent variables. However, Flowerdew and Aitkin (1982) have noted various problems with this approach. For example, there is a systematic tendency for the 'predicted' average flow to be under-estimated.

An alternative approach is to make use of a *Poisson distribution*, a distribution well suited to the analysis of counts and given by the equation

$$f(y) = \frac{\mu^y \exp(-\mu)}{y!} \quad y = 1, 2, 3, \dots \quad (60)$$

A Poisson distribution with the mean, μ , equal to 2 has been drawn below in Figure 14.

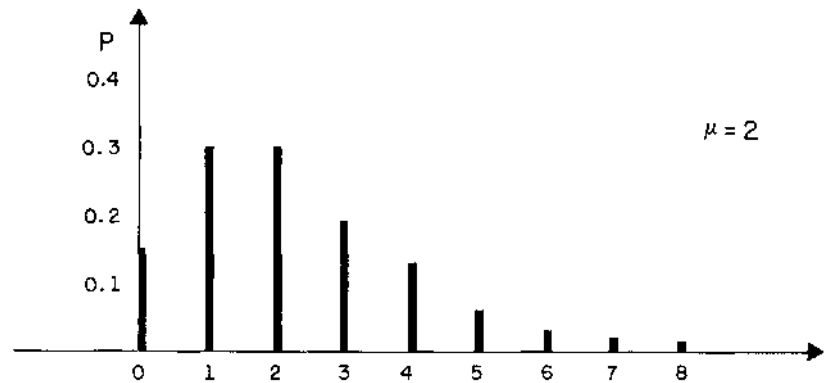


Figure 14 Poisson Density Function

Here again the mean of the distribution for each origin and destination pair μ_{ij} , may be made a function of their characteristics. As the gravity model of Equation (59) suggests

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \log(M_i) + \beta_2 \log(M_j) - \beta_3 \log(d_{ij}) \quad (61)$$

or

$$\mu_{ij} = \exp [\beta_0 + \beta_1 \log(M_i) + \beta_2 \log(M_j) - \beta_3 \log(d_{ij})] \quad (62)$$

The likelihood function for a single origin destination pair is found by substituting for μ and y in Equation (60), and ignoring the factorial.

$$L(\beta_0, \beta_1, \beta_2, \beta_3) = \frac{\exp[\beta_0 + \beta_1 \log(M_i) + \beta_2 \log(M_j) - \beta_3 \log(d_{ij})]}{\sum_{i,j} \exp[\beta_0 + \beta_1 \log(M_i) + \beta_2 \log(M_j) - \beta_3 \log(d_{ij})]} F_{ij} \quad (63)$$

The log-likelihood for flows from n origins to a single destination may be simplified, since with only one value of M_j the parameter β_2 cannot be estimated but is instead incorporated into the constant β_0 .

$$\ln(\beta_0, \beta_1, \beta_3) = \sum_{i=1}^n \{F_i [\beta_0 + \beta_1 \log(M_i) - \beta_3 \log(d_i)] - \exp[\beta_0 + \beta_1 \log(M_i) - \beta_3 \log(d_i)]\} \quad (64)$$

This corresponds to a model of the form

$$F_i = \alpha M_i^{\beta_1} d_i^{-\beta_3} \quad (65)$$

The maximum likelihood estimates of the β 's cannot be found by algebraic solution of the first order conditions for the maximum of the log-likelihood function. Computer routines must be used to search for their values.

(ii) Distance Decay and University Attendance

Table 6 gives data on the number of students from each U.S. state attending a 1983 Human Geography class at Northwestern University near Chicago, Illinois. Data is also provided on the 1980 populations of each state and the approximate distance from the population centre of each state to Chicago. How important are the population size and distance away of each state in determining the number of students from each state? Appropriate instructions are given in Appendix 1 for the fitting of a model of the form of Equation (65) using the GLIM computer package (Baker and Nelder, 1974), together with the output obtained.

The program selects some initial or trial values of the parameters but then improves upon them in much the same way as the Newton-Raphson procedure described earlier. After several cycles the peak of the likelihood function is reached and the maximum likelihood parameter estimates obtained are

$$\alpha = 3.16 \quad \beta_1 = 1.16 \quad \beta_3 = -0.73$$

The flows increase with population size but decline with distance. The plausibility of other values of these parameters may be examined in a variety of ways. To test the hypothesis that $\beta_1 = 1$, that the numbers of students are simply in proportion to the state's populations, a Wald test is convenient. Writing the null or restricted value of β_1 as $\hat{\beta}_1$

$$W = -(\hat{\beta}_1 - \hat{\beta}_1) \ln(\hat{\beta}_1) (\hat{\beta}_1 - \hat{\beta}_1) = -(1 - \hat{\beta}_1)^2 \ln(\hat{\beta}_1) \quad (66)$$

and remembering that $\ln(\hat{\beta}_1)$ is given by $-1/(\text{standard error})^2$, given by the computer output as

$$\text{Standard Error } \beta_0 = 1.06 \quad \text{S.E. } \beta_1 = 0.17 \quad \text{S.E. } \beta_3 = 0.17 \\ W = (0.16)^2 / (0.17)^2 = 0.89$$

Table 6 Data for U.S. States

State	Percent Urban Population	1981 p.c. Income Dollars	1980 Population (millions)	Distance (miles)	Students
Alabama	60.0	8,219	3.9	650	0
Arizona	83.8	9,754	2.7	1,456	0
Arkansas	51.6	8,044	2.3	560	0
California	91.3	11,923	23.7	1,848	4
Colorado	80.6	11,215	2.9	907	3
Connecticut	78.8	12,816	3.1	773	2
Delaware	70.7	11,095	0.6	706	0
Washington, D.C.	100.0	13,539	0.6	616	0
Florida	84.3	10,165	9.7	1,008	3
Georgia	62.3	8,934	5.5	661	0
Idaho	54.0	8,937	0.9	1,344	0
Indiana	64.2	9,720	5.5	180	1
Iowa	58.6	10,470	2.9	302	4
Kansas	66.7	10,813	2.4	538	1
Kentucky	50.8	8,420	3.7	325	0
Louisiana	68.6	9,518	4.2	795	0
Maine	47.5	8,535	1.1	974	0
Maryland	80.3	11,477	4.2	616	2
Massachusetts	83.8	11,128	5.7	829	1
Michigan	70.7	10,790	9.3	224	4
Minnesota	66.8	10,768	4.1	336	4
Mississippi	47.3	7,408	2.5	694	0
Missouri	68.1	9,651	4.9	336	2
Montana	52.9	9,410	0.8	1,064	1
Nebraska	62.7	10,366	1.6	470	1
Nevada	85.3	11,576	0.8	1,602	0
New Hampshire	52.2	9,974	0.9	829	0
New Jersey	89.0	12,127	7.4	706	6
New Mexico	72.2	8,529	1.3	1,120	0
New York	84.6	11,466	17.6	690	7
North Carolina	48.0	8,649	5.8	661	0
North Dakota	48.8	10,213	0.7	728	0
Ohio	73.3	10,313	10.8	280	9
Oklahoma	67.3	10,247	3.0	672	1
Oregon	67.9	10,008	2.6	1,758	0
Pennsylvania	69.3	10,370	11.9	560	3
Rhode Island	87.0	10,153	0.9	851	0
South Carolina	54.1	8,039	3.1	650	1
South Dakota	46.4	8,833	0.7	661	0
Tennessee	60.4	8,447	4.6	437	1
Texas	79.6	10,729	14.2	896	1
Utah	84.4	8,313	1.5	1,266	0
Vermont	33.8	8,723	0.5	784	0
Virginia	66.0	10,349	5.3	639	0
Washington	73.6	11,277	4.1	1,714	0
West Virginia	36.2	8,377	1.9	448	0
Wisconsin	64.2	10,035	4.7	112	0
Wyoming	62.8	11,665	0.5	1,030	0
Illinois	83.0	11,576	11.4		2

This is well below the critical value of 3.84 suggesting that $\beta_1 = 1$ is a plausible hypothesis.

Alternatively the relative likelihood function could have been examined. With three parameters the relative likelihood function for this surface is now four-dimensional. Figure 15 illustrates the surface $\max_{\alpha} \log LR(\alpha, \beta_1, \beta_2)$ obtained by selecting that value of α which gives the highest relative likelihood at each point in a grid of points of β_1 and β_2 values. The row of values of $\beta_1 = 1$ passes through the area of highly plausible values with relative log-likelihood values larger than -1.

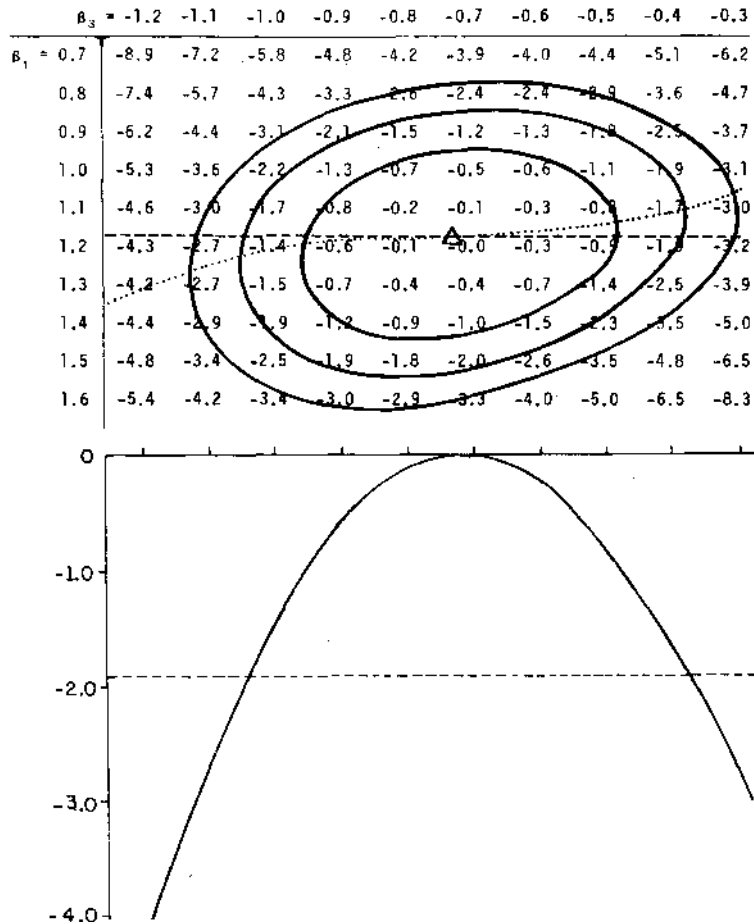


Figure 15 Contour and Profile Log-relative Likelihood Functions for the Poisson Regression Gravity Model

To examine the value of the distance decay parameter b , the profile contour plot could again be used. Alternatively, because our interest is now focussed on just β_3 both α and β_1 are nuisance parameters, and both may be eliminated by using the profile $\max_{\alpha, \beta_1} \log LR(\alpha, \beta_1, \beta_3)$. This curve is shown beneath the contour plot. It suggests a 95% confidence interval for β_3 of -1.05 to -0.38.

In all the examples so far, the model has been examined piecemeal, parameter by parameter. However in the travel mode example the idea of a perfectly fitting model was introduced. A perfectly fitting model can be constructed by allowing each observation a separate parameter, the parameter value being such as to make the predicted value from the model (e.g. the number of students from each state) exactly equal to the observed value. Overall performance or 'goodness-of-fit' may be examined using a likelihood ratio test of the fitted gravity model, considered as a model with only three unrestricted parameters (a , c and 133), against such a perfectly fitting model, with as many unrestricted parameters as observations. This statistic is available directly from the GLIM package output as the *scaled deviance*, with 48 observations the perfectly fitting model has 48 parameters, the fitted gravity model just 3, and so this likelihood ratio test statistic should be approximately distributed as χ^2 with $48-3 = 45$ degrees of freedom. Comparison of the value of the scaled deviance for the gravity model, given by GLIM as 56.4 with the appropriate χ^2 statistical table ($\chi^2_{45, 0.05} = 61.4$) suggests the model is plausible. Some caution is desirable, however, in view of the relatively small sample size. Other measures of goodness-of-fit based on the W or LM procedures, rather than the LR, might be checked.

The addition of a variable expressing the relative income of a state, I_i , ($I_i =$ per capita income of state i /mean per capita income of all states), might be considered. Northwestern University, as a private institution, may be expected to draw more strongly from the wealthy. The model is now given by

$$F_i = \alpha M_i^{\beta_1} d_i^{\beta_2} I_i^{\beta_3} \beta_4 \quad (68)$$

and the likelihood function for this larger model has an additional dimension, β_4 , with the previous model being represented within this space by those parts for which $\beta_4 = 0$. A test of the hypothesis that $\beta_4 = 0$ may be carried out by examining the improvement in the fit of the model that is possible as this constraint is removed. The reduction in the quantity named %X2 in GLIM the generalized Pearson χ^2 (a LM based measure of goodness of fit), during 1 cycle of GLIM's iterative fitting of the model with the income variable added, gives the appropriate LM statistic (see Section IV (i) and (ii) and Pregibon, 1982). The appropriate GLIM commands are given in Appendix I. The value obtained, 23.1 is much larger than the critical value for the addition of just one variable, 3.84, showing that $\beta_4 = 0$ is implausible, and that per capita income is an important variable within the model. Further cycles of the fitting procedure to reach the peak of the likelihood function in the extended parameter space (having originally started at the peak in the reduced parameter space in which $\beta_4 = 0$) allows a LR statistic of $\beta_4 = 0$ to be obtained from the scaled deviance values ($56.4 - 38.0 = 18.4$) together with output giving the parameter estimates and standard errors for the new fitted model with three explanatory variables.

(iii) Contingency Tables

The general structure of the statistical model used to estimate the gravity model is now often referred to as a Poisson regression model to emphasize its similarities with Normal regression. Where all the explanatory variables are dummy or nominal categorical variables a Normal regression model gives rise to analysis of variance and a Poisson regression model to contingency table analysis. For example, if population and distance were each classified as high (1) and low (0), the flows from each state would belong to one of four categories (high population - high distance, high population - low distance, low population - high distance and low population - low distance). Summing up the flows within each category to obtain N_{11} , N_{10} , N_{01} and N_{00} would give rise to the usual contingency table format

		Population	
		1	0
Distance	1	N_{11}	N_{01}
	0	N_{10}	N_{00}

VII SCIENTIFIC METHOD AND STATISTICAL INFERENCE

The Statistician cannot excuse himself from the duty of getting his head clear on the principles of scientific inference, but equally no other thinking man can avoid a like obligation. (Fisher 1922, quoted in Edwards, 1972)

(i) Alternative Statistical Frameworks.

Unfortunately, statisticians themselves have not been able to offer an agreed set of principles from which thinking men and women may gain clear guidance. There continues vigorous, sometimes vitriolic, debate amongst them. At risk of oversimplification we may identify three major schools of thought: Bayesian, likelihood and frequentist. Fisher himself made major contributions to more than one of these schools at different times in his career. The main problem seems to be one which is common to other areas of philosophy, for example social welfare theory. We can list a set of rules which we would like a statistical method to meet, but find that even though each rule appears both fundamentally necessary and innocuous, no method can fulfill each and every rule in all circumstances. Proponents of each school can point to persuasive failings in the others.

The strength of the Bayesian approach is that, quite rightly, they view the analysis of data as a means of adding to, updating or refining, our knowledge about a particular topic. We begin the analysis with certain prior views as to what are plausible values for a parameter and we use the data to improve these notions to derive a posterior view. The obvious criticism is that whereas the data is in some sense objective, the prior views are subjective and different analysts will hold different prior views and will obtain different results as a consequence. The Bayesians reply that to begin an analysis assuming complete ignorance as to plausible parameter values, as the likelihood and frequentist schools often do, may be woefully inefficient since it neglects much accumulated but imprecise knowledge that is available. The Bayesian approach does hold considerable appeal for some specific tasks,

such as the updating of an old input-output table as new but incomplete information arrives. However, a Bayesian approach is often unmanageably complex, both mathematically and computationally.

It is also possible to incorporate prior views within a likelihood type approach. In Section II(iv) it was shown how data from two samples could be combined within an analysis. Our prior views could be represented by a hypothetical set of data, specially constructed to be consistent with our expectations of how the process works. This hypothetical data could then be combined with the objective sample data for analysis.

The frequentist or classical school argue that sampling theory must form the basis of inference. As explained in the introduction, the sample data collected may be thought of as just one of a large set of possible samples that could have been obtained. A particular inferential decision, say to accept or reject the relevance of a particular explanatory variable, should be based on the probability that such a decision would be supported by these other potential data samples. This emphasis on hypothetical 'long-run results stands in contrast to the emphasis of likelihood-theory solely on the data at hand. Consider the very first example of a sample of 10 dwellings classified according to holiday home status. The probability of the data was given by Equation (2) as

$$P(p, r, n) = p^7(1-p)^3 \frac{10!}{(7! \times 3!)}$$

However, the same data could have been collected by a different procedure, for example continuing sampling until three permanent homes had been sampled. Under this procedure, although the sample size is not known before hand, the last dwelling is known to be a permanent home (the 3rd). Thus the possible permutations of dwellings within the potential samples is therefore restricted and the sample probability is now

$$P(p, r, n) = p^7(1-p)^3 \frac{9!}{(7! \times 2!)}$$

Calculation of confidence intervals for p using standard classical procedures are different in each case (see Pfanzagl, 1972). For the first case the lower bound on a 95% confidence interval is known to be larger than 0.4, whilst in the second it is known to be less than 0.25. The samples are considered as providing different information. But in likelihood terms both samples are equivalent since they give rise to the identical likelihood function (see Equation (3)). Likelihood protagonists point out that if some quite external event halted all survey work after the 10th dwelling had been surveyed (such as a nuclear strike vapourizing the survey site), then data from each sampling procedure must be equivalent. The debate remains unresolved.

In addition to the above disagreement, the classical school have constructed several hypothetical examples in which a likelihood approach appears to lead to patently incorrect conclusions. These examples have been recently reviewed (Hinde and Aitkin, 1984) and shown to be unpersuasive, their incorrectness being far from proven.

"The first clear strength of the likelihood framework is that the likelihood function and the likelihood ratio provide a coherent measure of the 'evidential information' within a sample. Such a fundamental and desirable

measure is missing in the classical approach and those measures which have been proposed have been later shown to be equivalent to the likelihood measure, and thus incompatible with the main body of classical theory. Another appeal of a likelihood approach is the ability to make inference purely from the data at hand without the need to consider other usually hypothetical data.

Whilst no-one doubts the value of a single comprehensive and entirely consistent framework, blind dogmatism cannot in itself provide one. This monograph has used the notion of hypothetical sample data to introduce the idea of random variation within actual sample data and it has linked likelihood intervals, LR, W and LM measures to significance levels. None of these would be eligible within a purist's likelihood approach.

(ii) Likelihood and Entropy

The relationship of likelihood to entropy has received passing attention from a number of writers, though no one seems to have provided a definitive statement (e.g. Wilson, 1981, pp. 75-6; Rao, 1973, pp. 172-5). The best explored link appears to be that which shows the equivalence of the logit model and entropy maximizing gravity specifications (Theil, 1972; Baxter 1982). Further understanding would be gained from a mutual exchange of insights from each approach.

VIII BIBLIOGRAPHY

a) Introduction

Aitkin, M. 1982. Direct Likelihood Inference in GLIM 82: *Proceedings of the International Conference on Generalised Linear Models*. Ed. Robert Gilchrist. Springer-Verlag, New York

Edwards, A.W.F. 1972. *Likelihood*. C.U.P. Cambridge

Kalbfleisch, J.G. 1979. *Probability and Statistical Inference*, Vol. II. Springer-Verlag, New York

Vincent, P. & Haworth, J. 1984. Statistical Inference: the use of the likelihood function, *Area*, 16, 131-146

b) Advanced References

Cox, D.R. & Hinkley, D.V. 1974. *Theoretical Statistics*. Chapman Hall, London

McCullagh, P. & Nelder, J.A. 1983. *Generalized Linear Models*. Chapman Hall London

Nelder, J.A. & Wedderburn, R.W. 1972 Generalised Linear Models. *Journal of the Royal Statistical Society A*, 135, 370-384

Rao, C.R. 1973. *Linear Statistical Inference and its Applications*. Wiley, New York

Silvey, S.D. 1970. *Statistical Inference*. Penguin, Harmondsworth

c) Methods and Applications within Geography

Examples of the use of some methods, like Normal regression, can be found in the majority of empirical texts and quantitative journals, whilst other methods, such as Poisson regression, are as yet a novelty and are therefore more difficult to find. For this reason the following selective list of references gives much greater emphasis to the less common methods.

Normal Regression Model

Ferguson, R. 1976. Linear Regression in Geography. *CATMOG 15*. Geo-Abstracts Norwich

Contingency Tables/Log-Linear Models

Upton, G.J.G. & Fingleton, B. 1979. Log-linear Models in Geography. *Transactions of the Institute of British Geographers*, 4, 103-15

Wrigley, N. 1981. Categorical data analysis. In: *Quantitative Geography: a British view*. Eds. N. Wrigley & R.J. Bennett, Routledge and Kegan Paul, London

Gravity Model/Poisson Regression

Baxter, M. 1982. Similarities in Methods for Estimating Spatial Interaction Models. *Geographical Analysis* . 14, 267-272

Flowerdew, R. and Aitkin, M. 1982. A Method of Fitting the Gravity Model Based on the Poisson Distribution. *Journal of Regional Science*, 22, 191-202

Logit Models

Hensher, D.A. and Johnson, L.W. -1981 *Applied Discrete choice Modelling* Halsted Press/Wiley, London

Wrigley, N. 1976. Introduction to the Use of Logit Models in Geography *CATMOG 10*, Geo-Abstracts, Norwich

Models for Discrete Time Duration Data

Crouchley, R., Pickles, A.R. & Davies, R.B. 1982. Dynamic Models of Shopping Behaviour: the linear learning model and some alternatives. *Geografiska Annaler B*, 27-33

Davies, R.B. and Pickles, A.R. 1983. Estimation of Duration of Residence Effects: a stochastic modelling approach. *Geographical Analysis*, 15, 305-17

Models for Continuous Time Duration Data

Davies, R.B. 1983. Destination Dependence: a re-evaluation of the competing risk approach. *Environment and Planning A*. 15, 1057-65

Pickles, A.R. 1983. The Analysis of Residence Histories and Other Longitudinal Data: a continuous time mixed Markov renewal model incorporating exogeneous variables. *Regional Science and Urban Economics* 12, 305-311

Other Applications in Geography

Burridge, P. 1980. On the Cliff-Ord Test for Spatial Correlation. Journal of the Royal Statistical Society B, 42, 107-8

d) *Computational Aspects of Likelihood Methods*

Baker, R.J. and Nelder, J.A. 1978. *The GLIM System: Release 3. Manual*. Numerical Algorithms Group, Oxford

O'Brien, L.G. 1983. Generalized Linear Modelling using the GLIM System. *Area*, 15, 327-36

Pregibon, D. 1982. Score Tests in GLIM with Applications in GLIM 82. *Proceedings of the International Conference on Generalised Linear Models*. Ed. Robert Gilchrist. Springer-Verlag, New York

Richardson, M.G. 1982. Learning to Use GLIM-3. *Introductory Notes on the GLIM-3 System*. Numerical Algorithms Group, Oxford

Sampson, R.J. 1976. *Surface 11 Computer Package*. Kansas Geological Survey

Wilson, A.G. and Kirby, M. 1982. *Mathematics for Geographers and Planners*, Routledge and Kegan Paul, London

e) *Likelihood and Other Approaches*

Birnbaum, A. 1972. Likelihood. In *Encyclopedia of the Social Sciences*, ed. D.L. Sills, 9, 299-301. Macmillan, London

Edwards, A.W.F. 1969. Statistical Methods in Scientific Inference. *Nature*, 222, 1233-7

Edwards, A.W.F. 1970. Likelihood. *Nature*, 227, 92

Hinde, J. and Aitkin, M. 1984. Nuisance Parameters, Canonical Likelihoods and Direct Likelihood Inference. *Mimeo*, Centre for Applied Statistics, University of Lancaster

Pfanzagl, J. 1972. Estimation: confidence intervals and regions. In: *International Encyclopedia of the Social Science*, ed. D.L. Sills, 506, 150-6, Macmillan, London

Theil, H. 1972. *Statistical Decomposition Analysis*, North Holland, Amsterdam

Wilson, A.G. 1981. *Geography and the Environment: systems analytical methods*. Wiley, London

APPENDIX 1. GRAVITY MODEL EXAMPLE USING GLIM

The following represents an interactive computer session using the GLIM package. User input has been distinguished from package output by the underlining of the latter; the underlining would not normally appear. Additional comments have been written in italics and would also not normally appear. The session would begin with a series of machine dependent commands to enable the user to login and call up the GLIM package. These would result in the following header being displayed and the ? prompt.

GLIM 3 (C) 1977 ROYAL STATISTICAL SOCIETY, LONDON

We now define the data vectors of length 48, beginning each statement with a \$ sign, and finishing each line with a RETURN key.

```
? $UNITS 48
? $DATA INC POP DIST STUD
? $READ 8219 3.9 650 0
? 9754 2.7 1456 0
? 8044 2.3 560 0
? 11923 23.7 1848 4
.
.
.
? 11665 0.5 1030 0
```

The explanatory variables are used in their log form

```
? $CALC LINC = %LOG(INC)
? $CALC LPOP = %LOG(POP)
? $CALC LDIST = %LOG(DIST)
```

To specify an appropriate Poisson regression model

```
? $YVAR STUD $ERROR P $LINK L
```

To fit the simple gravity model and display the parameter estimates

```
? $FIT LPOP + LDIST $DISPLAY E $
```

<u>SCALED</u>		
<u>CYCLE</u>	<u>DEVIANCE</u>	<u>DF</u>
<u>4</u>	<u>56.36</u>	<u>45</u>

	<u>ESTIMATE</u>	<u>S.E.</u>	<u>PARAMETER</u>
<u>1</u>	<u>3.157</u>	<u>1.061</u>	<u>%GM</u>
<u>2</u>	<u>1.162</u>	<u>0.1732</u>	<u>LPOP</u>
<u>3</u>	<u>-0.7315</u>	<u>0.1736</u>	<u>LDIS</u>

SCALE PARAMETER TAKEN AS 1.000

To display and store the generalised Pearson χ^2

? \$LOOK %X2 \$CALC %A = %X2
1 59.20

To perform one iteration with the addition of the income variable

? \$CYCLE 1
 ? \$FIT + LINC \$

	<u>SCALED</u>	
<u>CYCLE</u>	<u>DEVIANCE</u>	<u>DF</u>
<u>1</u>	<u>43.55</u>	<u>44</u>

----- NO CONVERGENCE BY CYCLE 1

To show the improvement in the generalised Pearson χ^2

? SCALC %A - %X2
23.05

To continue further cycles to obtain ML estimates

? \$CYCLE
 ? \$FIT LPOP + LDIS + LINC \$ DISPLAY E \$

	<u>SCALED</u>	
<u>CYCLE</u>	<u>DEVIANCE</u>	<u>DF</u>
<u>4</u>	<u>37.95</u>	<u>44</u>

	<u>ESTIMATE</u>	<u>S.E.</u>	<u>PARAMETER</u>
<u>1</u>	<u>-51.20</u>	<u>13.31</u>	<u>%GM</u>
<u>2</u>	<u>0.9166</u>	<u>0.1599</u>	<u>LPOP</u>
<u>3</u>	<u>-0.9607</u>	<u>0.1861</u>	<u>LDIS</u>
<u>4</u>	<u>6.073</u>	<u>1.486</u>	<u>LINC</u>

SCALE PARAMETER TAKEN AS 1.000

? \$STOP