01/28/99

Chapter 1:
The Likelihood Principle

"What has now appeared is that the mathematical concept of probability is ... inadequate to express our mental confidence or diffidence in making ... inferences, and that the mathematical quantity which usually appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability. To distinguish it from probability, I have used the term *'Likelihood'* to designate this quantity; since both the words 'likelihood' and 'probability' are loosely used in common speech to cover both kinds of relationship."

R. A. Fisher, *Statistical Methods for Research Workers*, 1925.

"What we can find from a sample is the *likelihood* of any particular value of ρ [a parameter], if we define the likelihood as a quantity proportional to the probability that, from a particular population having that particular value of ρ, a sample having the observed value *r* [a statistic] should be obtained. So defined, probability and likelihood are quantities of an entirely different nature."

R. A. Fisher, "On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample," *Metron, 1*:3-32, 1921.

Introduction

The likelihood principle as stated by Edwards (1972, p. 30) is that

> Within the framework of a statistical model, *all* the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data. ...For a continuum of hypotheses, this principle asserts that the likelihood function contains all the necessary information.

Although clearly implied in what Fisher wrote in the 1920's (see the epigraphs with which this chapter begins), the likelihood principle, which essentially holds that the likelihood function is the sole basis for inference, did not come into prominence until the 1950's and 1960's, principally through the work of Barnard, Birnbaum, and Edwards (see the references cited below) written largely in reaction to both the classical Neyman-Pearson (frequentist) and the Bayesian approaches to inference (see Press, 1989, and especially Jeffreys, 1934 and 1961).

The likelihood principle is very closely associated with the problem of parametric inference (Lindsey, 1996). Indeed, one hardly finds any discussion of it outside of the context of traditional parametric statistical models and their use, and I would be hard put even to give a clear definition of the concept in a general nonparametric context. It is also closely associated with the method of maximum likelihood estimation, both historically and conceptually, and, while the maximum or maxima of the likelihood function are, in some sense, its most interesting or significant point(s), it or they are not the only point(s). Taken seriously, the likelihood principle suggests that one might want to consider other points, for example, in the vicinity of a maximum particularly with respect to curvature, but farther away as well.

Writing in 1932, Fisher said this

> "...when a *sufficient* statistic exists, that is one which in itself supplies the whole of the information contained in the sample, that statistic is the solution of the equation of maximum

> likelihood: that when no sufficient statistic exists, the solution of the equation of maximum likelihood is *efficient* in the sense that the quantity of information lost tends in large samples to a zero fraction of the whole, and that this solution contains more information than other efficient statistics. Further, setting aside, for simplicity of statement, cases involving discontinuities, the limiting value of the amount of information lost may be reduced to zero by calculating, again from the likelihood, a second statistic ancillary to the primary estimate, and indeed may be reduced to a zero of any order by a series of such ancillary statistics. These latter properties are of interest in showing that, though the primary problem of estimation is solved by one feature only, namely the *maximum* of the likelihood, yet when we press for the fullest information obtainable from a finite body of data, it is the whole course of the function which has to be used." (pp. 259-260)

The concepts of sufficiency and ancillarity are discussed and illustrated in this Chapter below. Efficiency as defined by Fisher is an asymptotic property and discussion is postponed to the Chapter dealing with the properties of maximum likelihood estimates. As a closing caveat, it should be noted that the likelihood function is determined only within the context of a particular model. If we really believed that the likelihood function should serve as the sole basis for inference, we would never, for example, analyze the residuals from a regression. "Nearby" models may have quite different likelihood functions. In other words, inferences based solely on likelihood may not be robust. Properly supplemented by a variety of diagnostics, however, the likelihood principle can serve as the primary basis for statistical inference.

In the remainder of this chapter, I give a brief history of the likelihood principle and its relation to Bayes' Theorem and the method of inverse probability generally associated with the name of Laplace, and latterly with Jeffreys. Next I deal with the concepts of sufficiency and ancillarity and give a more formal definition of the likelihood principle and relate it to the now-popular methods of Bayesian inference, the renaissance of which is due primarily to L. J. Savage (1954) and, in econometrics, his apostle Arnold Zellner (1971). Finally, the problem of extending the likelihood principle to a generally applicable method of statistical inference, especially in the multi-parameter case, is introduced by means of the ideas of "concentrating" and "slicing" the likelihood function. These methods are discussed and applied in detail throughout the remainder of the book. In the course of the present discussion, I illustrate the concepts and methods for a variety of simple econometric examples.

1. From Bayes and Laplace to Fisher and Jeffreys[1]

> "Those who cannot remember the past are condemned to repeat it."
> George Santayana, *The Life of Reason, Vol. 1,*1905.

The history of statistical ideas is both fascinating and useful. The problem of scientific inference is to draw conclusions about the process which has (may have) generated a body of data (data generating process = DGP). Often the same person who seeks such conclusions is himself a part of the process which has generated the data, as in the experimental sciences or in the collection of survey data; more often in econometrics the analyst has not been part of the data generation process, which may involve data collected for regulatory purposes or interpolated from censuses. The question of what constitutes a satisfactory method of inference has had a long and somewhat tortured history from its early beginnings in the eighteenth century to the present day, and is still not fully resolved. It is useful to review this fascinating story briefly both because older ideas about inference continue to offer insight to present-day practitioners and because the confusions and groping towards solutions of the past are a lesson to those who think they now know all the answers.

In the beginning. Our story begins in the eighteenth century. By the end of the seventeenth century the mathematics of permutations and combinations had been extensively developed in conjunction with the

[1] Fisher's own version of this story (Fisher, 1956, revised augmented edition,1973, Chapters 1 and 2, pp.8-78) is quite self-serving and inaccurate and has been trenchantly criticized by Zabell (1989).

analysis of games of chance by such mathematical luminaries as Fermat, Pascal, Huygens, Leibnitz and Jacob Bernoulli (one of twelve of that illustrious mathematical family). The origins of the theory of probability in the study of games of chance is of some significance in two respects: First, it is natural to construct the "events" or outcomes as composed of equally probable cases. Second, focus on the stochastic nature of the outcomes themselves tends to obscure, or at least divert attention from, the mechanism generating the data or observations, since, in the case of games of chance, such mechanisms are perfectly known in general. With respect to the latter, Jacob Bernoulli's 1713 treatise, *Ars Conjectandi,* changed the focus from observed outcomes to the underlying "causes" or DGP.[2] The former, composition of complex outcomes in terms of equally likely elementary events, translates into equally likely or uniformly distributed causes in application of Bayes' Theorem (1764, discussed further below), and Bayes' Theorem with a uniform prior is nothing more or less than Laplace's (1774) independent development of the principle of inference by *inverse probability*[3].[4] The assumption of uniform priors in Bayesian analysis was later to be called the *principle of insufficient reason* and leads to serious conceptual difficulties and paradoxes, in large part responsible for Fisher's (1922, 1932) rejection of inverse probability inference.

Bayes in 1763 and Laplace in 1774. Laplace stated his "principe" as follows:

> "If an event can be produced by a number *n* of different causes, then the probabilities of these causes given the event are to each other as the probabilities of the event given the causes, and the probability of the existence of each of these is equal to the probability of the event given that cause, divided by the sum of all of the probabilities of the event given each of these causes." (Laplace, 1774, in Stigler's translation, pp. 364-365.)

[2] J. Bernoulli attempted to give a formal basis for the commonplace that the greater the accumulation of evidence about an unknown proportion of cases ("fertile" to total as he described it), the closer we are to knowledge of the "true" proportion, but that there is no natural bound in general which permits total elimination of all residual uncertainty. In his "limit" theorem, Bernoulli clearly shifts the emphasis from the observations themselves to inference about the underlying stochastic process which generates them.

[3] The term *inverse probability* is not used by Laplace but appears, according to Edwards (1997, p. 178), for the first time in de Morgan (1838, Chapter 3).

[4] Suppose we have two events A and B with joint probability of occurrence P[A, B]. It is elementary that P[A|B]P[B] = P[A, B], the joint probability equals the product of the marginal and the conditional. Thus far we have considered only a single "experiment" and induction from it, but we may have prior information about the parameters about which we seek to draw inferences from previous experiments or from theory; indeed, the parametric statistical model underlying our analysis is a prior belief. How should such prior information or belief be incorporated in the way in which we draw inferences? One possibility is to cast the prior information in the form of a probability distribution and make use of the above result.

Consider two hypotheses H1 and H2 and an "experiment" having the outcome x. The probability that we observe x give that H1 is true is P[x|H1], what Fisher (1921, p. 24) later called the *likelihood*, and if our prior belief about H1 can also be phrased in terms of a probability statement, P[H1], then

P[H1|x] = kP[x|H1]P[H1],

where k = 1/(P[x|H1]P[H1] + P[x|H2]P[H2]). This result is known as ***Bayes' Theorem***. P[H1] is the *prior distribution,* P[x|H1] is the *likelihood* of observing x if H1 holds, and P[H1|x] is the *posterior distribution* of H1. The proof follows from the fact that a similar relation holds for P[H2|x] for given priors and data

P[H1|x] = kP[x|H1]P[H1],

for the same k and then applying P[A|B]P[B] = P[A, B] twice. In general,

Bayes' Theorem: Given a likelihood L(θ|x) and a prior probability density defined on θ, p(θ), the posterior density for θ is

$$p(q|x) = cp(q)\,L(q|x) \text{ where } c^{-1} = \int_{\Theta} p(q)\,L(q|x)dq \text{ when } q \text{ is continuous.}$$

But as Stigler (1986, p. 103) remarks: "In this context an assumption of equally likely causes would have been understood by Laplace to mean that the problem is specified --the 'causes' enumerated and defined-- in such a way that they are equally likely, not that any specification or list of causes may be taken a priori equally likely. The *principle of insufficient reason* (to use a later name for the assumption that causes not known to have different a priori probabilities should be assumed a priori equally likely) would therefore be less a metaphysical axiom than a simplifying and approximative assumption invoked to permit calculation. Because Laplace considered a priori probabilities to be statements relative to our a priori knowledge of the causes of events, he supposed that if the causes under one specification were known not to be equally likely, we would respecify them in such a way that they were all equally likely, say, by subdividing the more likely causes."[5] True Bayesians give the principle a great deal more weight, of course, and make use of it in a far wider context.[6] The question of how to interpret the likelihood itself, indeed its very name, belongs to a later part of this story, when Fisher's ideas are discussed.

Combination of observations; Daniel Bernoull in 1777. Until Fisher (1922) there was not a sharp distinction between the likelihood of a series of observations and the posterior probability of those observations with noninformative prior, i.e., their inverse probability, but the question of how to use the likelihood or inverse probability to draw an inference about their DGP was already a subject of investigation in the eighteenth century. The idea of combining several measurements (plausibly independent) of the same quantity by taking their arithmetic mean had appeared at the end of the seventeenth century. But was this the best way of combining the information in the several observations to obtain a more accurate measurement of the quantity in question than that in any one of them? Good historical accounts of the problem of combining observations and its relation to the development of least squares by Legendre in 1805 and Gauss in 1809 are given in Plackett (1972) and Stigler (1986, Chapter 1).

[5] This is quite reminiscent of the generation of all sorts of pseudo random numbers in a digital computer from numbers uniformly distributed on the interval [0,1). See Section 1 on random number generation in Chapter 3, "Statistical Fantasies: Monte Carlo and Bootstrapping," below. Uniformly distributed random numbers may be the "building blocks" of more general random variables (RVs), but they are not prior in any metaphysical sense.

[6] Here's what Harold Jeffreys (1939, 1961, pp. 117-118) has to say:

> "Our first problem is to find a way of saying that the magnitude of a parameter is unknown, when none of the possible values need special attention. Two rules appear to cover the commonest cases. If the parameter may have any value in a finite range, or from $-\infty$ to $+\infty$, its prior probability should be taken as uniformly distributed. If it arises in such a way that it may conceivably have any value from 0 to ∞, the prior probability of its logarithm should be taken as uniformly distributed. There are cases of estimation where a law can be equally well expressed in terms of several different sets of parameters, and it is desirable to have a rule that will lead to the same results whichever sets we choose. Otherwise we shall again be in danger of using different rules arbitrarily to suit our taste. It is now known that a rule with this invariance exists, and is capable of very wide, though not universal, application.
>
> "The essential function of these rules is to provide a formal way of expressing ignorance of the value of the parameter over the range permitted. They make no statement of how frequently that parameter, or other analogous parameters, occur within different ranges. Their function is simply to give formal rules, as impersonal as possible, that will enable the theory to begin. Starting with any distribution of prior probability and taking account of successive batches of data by the principle of inverse probability, we shall in any case be able to develop an account of the corresponding probability at any assigned state of knowledge. There is no logical problem about the intermediate steps that has not already been considered. But there is one at the beginning: how can we assign the prior probability when we know nothing about the value of the parameter, except the very vague knowledge just indicated? *The answer is really clear enough when it is recognized that a probability is merely a number associated with a degree of reasonable confidence and has no purpose except to give it a formal expression.* [Italics added.] If we have no information relevant to the actual value of a parameter, the probability must be chosen so as to express the fact that we have none. It must say nothing about the value of the parameter, except the bare fact that it may possibly, by its very nature, be restricted to lie within, certain definite limits."

Savage(1954, pp. 64-66) has a good deal to say about the objections to such "noninformative" priors; Zellner (1971, pp. 41-53) gives a wealth of examples of priors designed to represent "knowing little." I deal further with the relation between Bayesian inference and likelihood inference in Section 5 below. Laplace appears to have been blissfully insensible to these latter day objections. Jeffreys stays reasonably clear of such objections by adopting a personalistic, as opposed to objectivist, view of probability.

Simpson (1755) took the first crucial step forward by focusing not on the observations themselves but in the errors of measurement. That freed Laplace (1774) and Daniel Bernoulli (1778) --another of the famous Bernoulli clan-- to concentrate on the properties that the distribution of the errors of measurement ought to have.[7] In an unpublished paper, Stigler (1998, p. 2) describes D. Bernoulli's contribution:

> "In brief outline, Bernoulli's argument runs like this: The common practice of taking the arithmetic mean of a set of observations cannot be correct, for it weights all observations equally, while they are surely not all of equal validity. Indeed, he noted, astronomers seem to acknowledge this by discarding extreme observations -- those too far from the rest to be considered as plausible measurements -- before taking the mean. Bernoulli did not claim that the probabilities of errors of various magnitudes could be specified precisely, but he felt that some of their characteristics could be stated qualitatively. In this, he seems to follow Laplace's 1774 practice of starting with a list of properties an error curve should have, but he cited neither Laplace nor any other writer.
> Bernoulli regarded it as axiomatic that errors above and below the true point may be taken as equally possible, so the scale of the probabilities of the errors will be symmetrical about that point. Furthermore, observations nearer the true point will be more probable than those more distant, and errors beyond some limit of maximum error will not be possible."

Bernoulli then argued that the "true" value of the magnitude being measured ought to be taken as the value which maximizes the probability of the whole set of observations with such a distribution of errors. This is clearly what we call today the method of maximum likelihood. Unfortunately, he made a poor, one might even say truly awful, choice of error distribution and came up not with the mean but something quite different even for two or three observations.[8] And solving the problem with his choice of error distribution with more than three observations is horrendous. Still, Daniel Bernoulli usually gets the credit for having invented the method of maximum likelihood.[9]

Gauss in 1809. Even the great Laplace stumbled over the form of the error distribution (see Stigler, 1986, pp. 105-122). Resolution of the matter had to wait until Gauss reinvented the normal distribution in 1809.[10] The principal subject of Gauss's book was an investigation of the mathematics of planetary orbits, but at the very end he added a section on the combination of observations. Gauss considered a somewhat more general problem than the estimation of one "true" value from a series of measurements subject to error: The estimation of a linear equation with constant coefficients, known independent variables, and observations on the dependent variable subject to error, ε_i. Assuming the errors to be independent of one another and distributed with unknown distribution, $\varphi(\varepsilon_i)$, Gauss proposed to estimate the unknown coefficients by maximizing $L = \Pi\varphi(\varepsilon_i)$. So far nothing beyond what Laplace and Bernoulli had done in the case of simple measurements, but now Gauss took a new and inspired direction: Rather than starting with the principle of insufficient reason, imposing some properties the distribution of errors ought to have (such as symmetry about zero and lesser probabilities for large errors than for small

[7] Noting that the binomial, or multinomial, distribution was used in the study of games of chance, Stigler (1986, p. 91) remarks:

> "For all its attractive properties, the binomial distribution has the unfortunate feature that the distribution of the difference between an empirical relative frequency and the unknown true proportion depends upon the unknown true proportion. Thus binomial 'error' distributions are not fixed; they cannot be taken as known (even by hypothesis) unless the true proportion is taken as known. [This is true] even for large numbers of trials ... [so a limiting argument cannot be used]"

[8] In his commentary on Bernoulli's memoir, Euler correctly points out that the maximum likelihood principle is quite arbitrary in the sense that there is no reason to believe that the DGP of the observations is that which gives them the greatest probability, which is, of course, why one should in general look at the whole of the likelihood function. (Fortunately, in those days the referee's comments were merely appended to the published paper instead of being sent to the author with a letter of rejection!)

[9] Edwards (1974, p. 10) credits one J. H. Lambert, *Photometria* ,London: Augustae Vindelicorum, 1760, with priority.

[10] De Moivre, of course, had it as early as 1733 in an unpublished memorandum later published as part of de Moivre (1738). See Hald (1998, pp.17-25) and Daws and Pearson (1972).

ones), then arriving at a suitable form for φ, Gauss reversed the procedure. Instead of imposing further conditions directly as Laplace and others had done, he assumed the conclusion! He reasoned that in the case of a single value measured numerous times with error, dozens of eminent and not so eminent astronomers couldn't be wrong: the arithmetic mean must be the correct answer. He then proved in this case that the arithmetic mean of the observations maximizes L only when

$$j\,(e) = \frac{h}{\sqrt{p}}\, e^{-h^2 e^2}$$

for some positive constant h, which could be interpreted as a measure of the precision of observation. Extending this result to the more general case of a linear equation with unknown coefficients and known explanatory variables, dependent variable subject to errors of measurement, Gauss then showed that maximizing L yields the least squares estimates of the unknown coefficients. In this way, both the normal distribution and least squares were born. The circularity and boldness of this argument are breath-taking. And why should a distribution which solves one special case, a single value measured several times over with error, generalize to a much more complex problem?

The Laplace-Gauss synthesis. One can imagine Laplace sputtering at such nonsense when he encountered Gauss's book, probably, according to Plackett (1972) and Stigler (1986, p. 143), sometime after April 1810. Laplace's major contribution to probability theory is now called the central limit theorem; he read his memoir (1810) reporting this result to the French Academy on April 9, 1810. It was a major generalization of de Moivre's limit theorem (1733, 1738) for the binomial distribution: Any sum or mean, not merely the total number of successes in a series of trials, will, if the number of terms be large enough (under certain regularity conditions and except in exceptional cases), be approximately normally distributed. He proved this more general result using what we now call Fourier transforms or characteristic functions (probably as early as 1785, Stigler, 1986, p. 137) but had no real use for it until he saw Gauss's nonsensical derivation. One can further imagine Laplace saying to himself, "Carl Friedrich really laid a big one this time! But -- par Dieu-- I can fix it." Laplace had just the right justification for the distribution of errors which Gauss had shown led to least squares: if the errors were caused by numerous insignificant causes, each negligible in effect, but summed together, then they ought to be distributed normally more or less. He rushed into print with a supplement to his memoir. Laplace (1812, 1820) went even further. He showed that all estimates of the coefficients which are linear functions of the independent variables are approximately normally distributed and that, within this class, the ordinary least squares (OLS) estimates have the smallest expected squared error. He further derived the multivariate normal as the limiting distribution of two or more least squares estimates. Perhaps piqued by Laplace, Gauss (1823) had second thoughts concerning his "derivation" of OLS as the "maximum likelihood" solution from a normal distribution of errors.[11] He noted that the analysis really depended on second moments and that if one was content to measure the accuracy of estimates which were linear functions of the observations by their expected squared error then his conclusion held without regard to the distribution of the errors (as long , of course, as they were iid with zero mean and common variance). He freed the method of least squares from the assumption of normally distributed errors and thus from Laplace's asymptotic justification, producing what we call today the Gauss-Markov Theorem (see Chapter 7 below for a detailed discussion and derivation of this result).[12]

[11] Although I must say that I personally find his earlier "upside down" derivation most appealing; it is so like what economists are wont to do!

[12] Plackett (1972, pp. 245-246) quotes a letter from Gauss to Wilhelm Olbers, an astronomer friend, dated February 22, 1819, in which Gauss reports his progress in freeing least squares from the normality assumption:

> "I am also occupied at present with a new basis for the so-called method of least squares. In my first basis I supposed that the probability of an observational error x was represented by e^{-hhxx}, in which event that method gives the most probable result with complete rigour in all cases. When the law of error is unknown, it is *impossible* to state the most probable results from observations *already made*. Laplace has considered the matter from a different angle and chosen a principle, which leads to the method of least squares, and which is quite independent of the law of error, when the number of observations is indefinitely large.

Writing of the Gauss-Laplace synthesis, Stigler (1986, p. 158) says:

> "The Gauss - Laplace synthesis brought together two well-developed lines ---one the combination of observations through the aggregation of linearized equations of condition, the other the use of mathematical probability to assess uncertainty and make inferences--- into a coherent whole. In many respects it was one of the major success stories in the history of science. Yet it also poses a puzzle, for the applications of this marvelous statistical technology were widespread only geographically; to an amazing degree they remained confined to the narrow set of disciplines that spawned them. They became commonplace in astronomical and geodetic work while remaining almost unknown in the social sciences, even though their potential usefulness in the social sciences was known from the time of Bernoulli and even though some early social scientists (for example, Adolphe Quetelet) were also astronomers."

The method of inverse probability, however, was to remain the paradigm until well into the next century. Nineteenth and early twentieth century statistics was dominated by Laplace (1820), *Théorie analytique des probabilités,* and maximizing the posterior distribution, usually with noninformative prior, until Fisher's papers "On the Mathematical Foundations of Theoretical Statistics" (1922), "Theory of Statistical Estimation" (1925), and his book, *Statistical Methods for Research Workers* (1925), and Neyman's series of papers with Egon Pearson (1928, 1933a and 1933b). During the nineteenth century, the inverse probability paradigm gradually penetrated other disciplines (Dale, 1991, pp. 257-438; Stigler, 1986, pp. 161-361). By the end of the nineteenth century, for example, inference by means of inverse probability and least squares regression are to be found in the biological sciences (Karl Pearson), physical anthropology and human genetics (Francis Galton), and economics (F. Y. Edgeworth). For the most part the interpretation of probability as a measure of subjective belief happily co-existed with the objective interpretation as a limiting frequency, set out explicitly in Laplace's *Théorie analytique des probabilité,* but certainly implicit in the origins of probability in the study of games of chance. Karl Pearson (1892, pp. 168-174) is quite explicitly on the frequentist side.[13] No one until Fisher (1921, 1922), however, seems to have noticed the contradiction between inverse probability inference and "objective" probability as opposed to a subjective, personal view; and Fisher himself was not aware of it in 1912. Writing much later, Jeffreys (1934, 1961) and Savage (1954) are very sensitive to the issue and insist on subjective probability as a basis for Bayesian inference.

Fisher's rejection of the inverse probability paradigm and the invention of likelihood. Fisher is notoriously difficult to interpret; people are inclined to read into him what they want to hear, and I am probably no exception. His view of probability is obscure. Savage (1976, pp. 461-462) writes: "Fisher, as everybody knows, was a frequentist, yet I -- who profess to [take] an interest in such things -- was somewhat taken aback in my rereading to find how vehemently he denies that probability is a limiting frequency in an indefinite sequence of repeated [actual] trials, which is the position that frequentists ordinarily take....For Fisher, a probability is the fraction of a set, having no distinguishable subsets, that satisfies a given

> "With a moderate number of observations, however, one remains quite in the dark if the law of error is unknown, and Laplace himself has also nothing better to say in this case, than that the method of least squares may also be applied here because it affords convenient calculation. I have now found that, by the choice of a principle somewhat different from that of Laplace (and indeed, as cannot be denied, one such that its assumption can be justified at least as well as that of Laplace, and which, in my opinion, must strike *anyone without a previous predilection* as more natural than Laplace's) - all those advantages axe retained, namely that in all cases for every error-law the method of least squares will be the most advantageous, and the comparison of the precision of the results with that of the observations, which I had based in my *Theoria* on the error-law e^{-hhxx}, remains generally valid. At the same time there is the advantage, that everything can be demonstrated and worked out by very clear, simple, analytical developments, which is by no means the case with Laplace's principle and treatment, thus, for instance, the generalization of his conclusions from two unknown parameters to any number does not yet appear to have the justification necessary....."

[13] For some historical insights see also Neyman (1977). Pearson's most famous paper (1900), in which he derives the chi-square distribution, implies the frequentist or sampling view of probability quite clearly. I will explore the implications of this approach for econometrics in the next chapter dealing with Neyman and Pearson (Egon, that is) and Haavelmo.

condition....Such a notion is hard to formulate mathematically, and indeed Fisher's concept of probability remained very unclear...."[14]

Whatever Fisher's view of probability was, it was not consistent with the subjective and personalistic concept compatible with inverse probability, or Bayesian statistical inference more generally. Aldrich (1997) has carefully reviewed Fisher's early papers with the purpose of better understanding how Fisher came to the concept of likelihood as distinguished from posterior probability (with noninformative prior) and maximum likelihood estimate as distinguished from most probable posterior value. It's a nightmarish Fisherian tangle, to be sure, but with Aldrich as my guide, my conclusion is that Fisher finally came to the view that what clung to the prior was not a probability distribution but a statistic and he called it the ***likelihood.*** In his 1921 paper, Fisher (p. 24) noted the confusion (his or ours?) between Bayes' Rule and what he (Fisher) called maximum likelihood in 1922:

> "Bayes attempted to find, by observing a sample, the actual probability that the population value [ρ] lay in any given range. ... Such a problem is indeterminate without knowing the statistical mechanism under which different values of ρ come into existence; it cannot be solved from the data supplied by a sample, or any number of samples, of the population."

And in 1922 he wrote (Fisher, p. 326):

> "...probability is a ratio of frequencies, and about the frequencies of such [parameter] values we can know nothing whatever."

Again in (1921, pp. 4 and 25):

> "...two radically distinct concepts have been confused under the name of 'probability' and only by sharply distinguishing between these can we state accurately what information a sample does give us respecting the population from which it was drawn....We may discuss the probability of occurrence of quantities which can be observed... in relation to any hypotheses which may be suggested to explain these observations. We can know nothing of the probability of hypotheses... [We] may ascertain the likelihood of hypotheses... by calculation from observations: ... to speak of the likelihood ... of an observable quantity has no meaning."

One further problem that Fisher (1912, 1922) noted was the lack of invariance of the parametrization of the prior: for example, a uniform prior for h in the normal of Gauss (1809) is inconsistent with a uniform prior for c = 1/h. Edgeworth (1908-1909, p. 392) also noted this problem but it does not seem to have troubled him. The likelihood and its maximum are invariant to the parametrization chosen.

At first, Fisher (1921) thought in terms of a single statistic, the sample correlation r, in relation to the corresponding population parameter ρ (see the second epigraph with which this Chapter begins) and considered the sampling distribution of r, but by 1922 (p. 310) he was beginning to think of the entire likelihood function as a statistic the properties of which might be investigated --enter the key idea of sufficiency, which is explored in the next section (Fisher, 1920; but he only named the concept in 1922) .[15]

[14] John Pratt, who edited Savage's notes for his Fisher Lecture, which became the basis for the posthumous 1976 article cited above, added in footnote:

> " 'Actual' has been inserted, at the risk of misrepresenting frequentists, because in early papers Fisher defines probability as a proportion in an 'infinite hypothetical population' of what seem to be repeated trials under the original conditions, where 'the word infinite is to be taken in its proper mathematical sense as denoting the limiting conditions approached by increasing a finite number indefinitely.' (Fisher,1925, p. 700; see also 1922 p. 312.) Later he says, 'An imagined process of sampling ... may be used to illustrate ... Rather unsatisfactory attempts have been made to define the probability by reference to the supposed limit of such a random sampling process.... The clarity of the subject has suffered from attempts to conceive of the 'limit' of some physical process to be repeated indefinitely in time, instead of the ordinary mathematical limit of an expression of which some element is to be made increasingly great.' (*Statistical Methods and Scientific Inference,* p.110.)"

The "measure theoretic" view quoted by Savage is actually quite modern; see Kingman and Taylor (1966) and Billingsley (1995). But I, and apparently Savage too, doubt that this is what Fisher really had in mind.

[15] In this same paper, Fisher (1922, p. 311) also introduced the terms "statistic" and "parameter" and later wrote: "I was quite deliberate in choosing unlike words for these ideas which it was important to distinguish

But how to investigate these properties and with what to compare them? In the period leading up to the characterization of the likelihood, that which clung to the prior in the Bayesian formulation, Fisher was working on comparing the sampling properties of estimators (sample statistics for the purpose of drawing an inference about a parameter or parameters) of the standard deviation and correlation parameters. You will recall that Gauss (1809) gave an inverse probability argument for least squares, but, piqued by Laplace (Gauss, 1823, and letter to Olbers, 1819, *loc. cit.*), he later gave what amounted to a justification in terms of the mean-square error of OLS compared with any other estimator linear in the observations on the dependent variable, that is, essentially an argument about the sample properties of the estimates. The likelihood, while it may be a sufficient statistic, is a whole function of all the parameters of the DGP conditional on the observed data and thus makes a rather awkward creature with which to compare anything. Fisher took a leaf from Laplace and the inverse probabilists who followed Laplace throughout the century: he considered the maximum of the likelihood and called it the "maximum-likelihood estimate" (Fisher, 1922, p. 323). Now sufficiency, which I consider in the next section, is a finite sample property, but maximum-likelihood estimators (when they exist, and existence did not trouble Fisher) do not generally have particularly good finite sample properties (such as unbiasedness or minimum variance) within a broader class of estimators, although, when a unique maximum likelihood estimate (MLE) exists, it is always a function (obviously in some trivial sense) of sufficient statistics.

In Fisher (1922, p. 316), two desiderata are added to sufficiency: consistency and efficiency. These are both large-sample or asymptotic properties and will be discussed in detail in Chapters 4 and 6 below. For now the following must suffice: While sufficiency is the criterion that a statistic (or statistics) should summarize all of the relevant information contained in the sample of observations, efficiency refers to the "efficiency" with which they (it) do(es) so in terms of sampling variance for large samples within a class of estimators having a particular large sample distribution (usually taken to be normal following Laplace). Consistency, which nowadays we also treat in terms of the asymptotic distribution of the estimator(s), was linked by Fisher rather to the method of moments (Pearson, 1902-1903), and he took for granted that the ML estimator(s) satisfied the criterion.[16] Finally, by 1932 (see the quote from Fisher, 1932, pp. 259-260, in the introduction to this Chapter, above), Fisher gave some thought about what to do if the ML estimate by itself was not sufficient and introduced the idea of ancilliary statistics, which I will discuss in Section 3 below.[17]

Did Edgeworth have priority? Before leaving this section on the history of the likelihood principle and maximum likelihood, I have to dispose of one very important question: Did Edgeworth (1908-1909) have priority? In a paper following Savage (1974), Pratt (1974, p. 501) considers the following questions:

> "1.To what extent did Edgeworth derive the method of estimation he advocates theory (which coincides with maximum likelihood) via sampling theory (direct probability) as well as via inverse probability?
> 2.However he derived it, did he advance the idea that it was a general method with desirable sampling-theory properties, especially asymptotic efficiency ?
> 3. How far did he go toward proving such sampling-theory properties?"

Pratt's conclusions (pp. 511-512) are:

> "In fact, Edgeworth did not bind his method on the theory of inverse probability.
> ...
> "To answer the questions raised [above]..., in [one]...case, Edgeworth actually derived the method of maximum likelihood (without the name or its connotation) via direct as well as inverse probability .... He was convinced of its sampling-theory asymptotic efficiency in general (like Fisher in 1922). He adduced enough evidence for it and made enough progress toward proving it to deserve very considerable credit.

as clearly as possible." (From Fisher's unpublished correspondence collected in Bennett,1990, p. 81, and cited by Aldrich, 1997, p.174.) It is difficult now for us to realize how difficult this distinction was to comprehend before Fisher introduced the terms -- and how much confusion failure to do so engendered.

[16] Savage (1976, pp. 454 and 459) has a fairly detailed discussion of what Fisher meant by consistency in relation to the modern concept.

[17] Recall Euler's criticism of the 1713 paper by Jacob Bernoulli!

> "Thus I believe that Edgeworth anticipated Fisher more than most commentators suggest (and the neglected 1909 Addendum is an important part of the evidence). If Edgeworth's contribution is minor, what of Fisher in 1922, thirteen years later ? Fisher's view of the problem was clearer and grander, but his conjecture was the same and his 1922 "proof" was only an invalid gesture.
>
> "Fisher's great advance was to give a general proof of efficiency (in 1925, for one parameter, and in the senses of 'general' and 'proof ' relevant at the time), and of course to introduce and explore very fruitful related concepts. Fisher's 'derivation' by maximizing the likelihood is also valuable to whatever extent 'likelihood' is a free-standing concept, meaningful in itself. But once good sampling properties are proved, the mode of derivation should be unimportant anyway (from a 'classical' point of view). If Edgeworth or Fisher had proved the sampling-theory asymptotic efficiency of all posterior modes for smooth, nonvanishing prior densities, not merely maximum likelihood estimates, would not that have been still better? And of course Edgeworth's treatment of inverse probability does not diminish his contribution on the direct side, but only its visibility."

Still, it's Fisher's influence, not Edgeworth's that matters now.

2. The Concept of Sufficiency

> "By 1922, Fisher had named sufficiency, related it to maximum likelihood, stated the factorization theorem, and applied the concept in a variety of situations. The theory of estimation had taken a gigantic leap forward; the concept of sufficiency, one of Fisher's most original contributions had been born." (Stigler,1973)

A statistic is a function of the observations alone which does not depend on any of the parameters of the model about which we are trying to draw an inference. A *sufficient statistic* is a statistic such that the conditional distribution of the original observations, *given the statistic,* also does not depend on any unknown parameters. The concept of sufficiency was first defined by Fisher (1921), who later wrote (1925) that a sufficient statistic "... is equivalent, for all subsequent purposes of estimation, to the original data from which it was derived." Fisher therefore concluded that any inference should be based only on sufficient statistics, a conclusion which is called the *sufficiency principle.*[18]

A parametric statistical model consists of a random vector $x\varepsilon X$ of observations having a joint distribution function $F(x;\theta\ )$, with corresponding density $f(x;\theta)$, depending on the unknown parameters $\theta\varepsilon\Theta$. It is assumed that F is known. A *statistic* is a possibly vector-valued function of x which does not depend on θ; a *sufficient statistic* is a statistic T such that the conditional distribution of x given T does not depend on θ. The factorization theorem (Fisher, 1922) states:

Factorization Theorem: Let T(x) be a statistic having the density $g(t;\theta)$. T is sufficient if and only if $f(x;\theta) = k(x)h[T(x);\ \theta]$ for some functions k and h; or, equivalently, if and only if $f(x;\theta) = k(x)g[T(x);\ \theta]$ for some function k(x).

It is clear that any invertible function of a sufficient statistic is also sufficient.

The most common example given of sufficient statistics is the sample mean and sample variance for the mean and variance of a normal population from which the sample is presumed to be drawn.

[18] Hogg and Craig, 1978, Chapter 10, pp. 341-369, have a very good discussion of the concept of sufficiency. A more recent and more detailed discussion is Azzalini, 1996, pp.29-47. A more advanced discussion of sufficiency and related topics is contained in Barndorff-Nielsen, 1978.

Example 1. Let x = (x_1, x_2, ..., x_n) be n independent drawings from N(μ, $σ^2$), θ = (μ, $σ^2$). Consider T(x) = (T_1, T_2) = $(\sum x_t^2, \sum x_t)$. f(x;θ) can be written

$$\frac{1}{(2ps^2)^{\frac{n}{2}}} e^{(-\frac{T_1}{2s^2}+\frac{T_2 m}{s^2}-\frac{nm^2}{2s^2})} = \frac{1}{(2ps^2)^{\frac{n}{2}}} e^{(-1/2s^2)\sum\{(x_i-T_2/n)^2+(T_2-nm)^2\}} .$$

Therefore T is sufficient. Moreover,

$$\bar{x} = T_2 / n \text{ and } s^2 = \frac{T_1 - T_2^2 / n}{n-1}$$

together are an invertible function of T, so the sample mean and sample variance, taken together are jointly sufficient for (μ, $σ^2$). If we know $σ^2$, the sample mean is itself sufficient for μ , but even if μ is known, T_1 by itself is not sufficient for $σ^2$ since the term multiplying

$$e^{(-\frac{T_1}{2s^2})}$$

contains $σ^2$.

The original data are always jointly sufficient for the parameters of the model; the trick is to find a smaller number. The following result often helps (see Hogg and Craig, 4th edition,1978, section 10.4): Suppose we can write

$$f(x;q) = k(x)p(q)\exp\{c(q)'T(x)\} ,$$

then T(x) is a *minimal sufficient statistic*.

Example 1 continued. In the above example f(x;θ) can be so written with

$$\begin{aligned} k(x) &= (2p)^{-n/2} \\ p(q) &= s^{-n} e^{-\frac{nm^2}{2s^2}} \\ c(q) &= (-\frac{1}{2s^2}, \frac{m}{s^2})' \\ T(x) &= (\sum x_t^2, \sum x_t)' \end{aligned} .$$

So T(x) is minimally sufficient.

Example 2. Consider a Poisson distribution with parameter θ, 0 < θ < ∞, with p.d.f.

$$\begin{aligned} f(x;q) &= \exp\{(\ln q)x - \ln x! - q\},\ x = 0,1,2,... \\ &= 0,\ \text{elsewhere.} \end{aligned}$$

$T(x_1, x_2, ..., x_n) = \sum x_t$ is a minimal sufficient statistic for θ.

The concept of *completeness* (Hogg and Craig, 1978, section 10.3) is also important. A statistic T is said to have a complete family of distributions if Eh(T) = 0 for all θ implies Prob{h(T) = 0} = 1. A *complete sufficient statistic* is a sufficient statistic with a complete family of distributions. Any invertible function of a complete sufficient statistic is also a complete sufficient statistic. A complete sufficient statistic is a minimal sufficient statistic but not all minimal sufficient statistics are complete.

Suppose

$$f(x;q) = k(x)p(q)\exp\{c(q)'T(x)\} .$$

If {c(θ); θεΘ} contains an open interval or set, T(x) is a complete sufficient statistic for θ.

Example 1 continued. Since the set defined by { $c(q) = (-\frac{1}{2s^2}, \frac{m}{s^2})'$; θεΘ} is the set of all vectors whose first coordinate is negative, it contains an open set and T is a complete sufficient statistic for (μ, σ^2). So $(\bar{x}, s^2)$ is also a complete sufficient statistic for (μ, σ^2) as well.

Caveat. Fisher's sufficiency principle, that inference should be based only on sufficient , presumably minimally sufficient, statistics, is a compelling one, but if we really believed this we would never, for example, analyze the residuals from a regression. The problem is that the sufficiency or insufficiency is determined only within the context of a particular model. "Nearby" models may have quite different sufficient statistics. In other words, sufficiency may not be robust in practice.

3. Ancillary Statistics and the Conditionality Principle

> "Since the original data cannot be replaced by a single statistic, without loss of accuracy, it is of interest to see what can be done by calculating, in addition to our estimate, an ancillary statistic which shall be available in combination with our estimate in future calculations."
>
> Fisher (1925, p. 724.)

Ancillary statistics were first defined by Fisher (1925) -- who else?--and were important in his development of a general theory of estimation in *Statistical Methods and Scientific Inference* (1956, Chap. 6).

Continuing the discussion of the preceding section: We are given a parametric statistical model, i.e., a random vector xεX of observations having a joint distribution function F(x;θ), with corresponding density f(x;θ), depending on the unknown parameters θεΘ. F is known. Suppose a statistic *a* = *a*(x) which does not depend on θ. *a* is said to be ancillary for θ if and only if the joint distribution of x and a can be factored as

$$f(x;q) = f_{x|a}(x;q|a)\, f_a(a)$$

An ancillary *statistic* is a possibly vector-valued function of x, *a*, which does not depend on θ, such that the *marginal* distribution of a does not depend on θ. In contrast, a *sufficient statistic* is a statistic T such that the *conditional* distribution of x given T does not depend on θ. If all you are concerned about is θ, you might as well base your inferences on the conditional distribution of the sample observations given the ancilliary statistic *a*. This is called the *conditionality principle* and it is an important basis of the *likelihood principle*, which is discussed in the next section .

The easiest way to see what the meaning of ancillarity is, is to look at a couple of examples:

Example 2, ancillarity: The sample size can often be considered an ancillary statistic. For example, suppose you conduct an experiment to measure the mean of a normal distribution by first tossing a fair coin to decide whether to draw a sample of 10 or a sample of 10,000; then, based on whatever size sample was drawn, we compute the mean (and the variance) of the distribution from the jointly sufficient statistics, which are the sum and sum of squares of the sample observations in this case. Clearly, any inference you might want to draw from the sample about the mean doesn't depend on whether you drew a sample of 10 or a sample of 10,000 because the sample size depends in no way on the mean about which you are making an inference. As usual, you might "estimate" the unknown mean of the normal distribution by the sample mean (this would be the maximum likelihood estimate) and it would make no difference whether the sample on which you based your estimate was 10 or 10,000. But if we now imagine repeating this experiment over and over again (taking the frequentist point of view), the sample mean would vary from experiment to experiment; it would have a distribution centered about the population mean, but *conditional on the sample size* the variance of the sample mean would be either $\sigma^2/10$ or $\sigma^2/10{,}000$, rather different. The variance of the sample mean in the *unconditional* experiment is still different, for sometimes the sample size would be 10 and sometimes 10,000 and it would be necessary to take this variation into account in deriving the result. I will return to this example in Chapter 3 below.[19] The important point is that ancillary statistics are often associated with the *precision* of the estimation of another parameter from a frquentist point of view. The question of how "comfortable" you might feel about the precision of the sample mean as an estimate of the population mean, however, can only be answered *conditional* on the number of observations which turned up in the particular experiment you in fact ran.

Example 3, ancillarity continued: This example is due to Cox (1958). Suppose we are going to measure a certain quantity using one of two instruments: Instrument A is very precise; our measurement error variance will be only 10; but it is also expensive to operate. Instrument B is cheap to operate but it is very imprecise; its measurement error variance is 100. If one of the two instruments, A or B, is chosen at random and a series of measurements are made, the mean of which is taken as an estimate of the quantity in question, 10 observations in the case of instrument A and 100 in the case of instrument B. What is the precision of the estimate as measured by the variance of the mean error? Clearly the same in both cases if the error distribution is assumed to be normal. The random variable the value of which indexes the instrument chosen is ancillary; inferences should be conditional upon its value even though the sample space for the complete experiment contains possible results from both experiments. The point is that measurements were actually made with only one of the two instruments.

The *conditionality principle*, in the sense that inference should always be conditional on any ancillary statistics, is not entirely uncontroversial: First, there are no general constructive techniques for finding ancillary statistics. Second, ancillary statistics sometimes exist which are not functions of the minimally sufficient statistics so that conditioning upon the observed values may conflict with the sufficiency principle discussed in the preceding section. Nonetheless, my own view is that the two principles of sufficiency and conditionality are convincing in an econometric context: For the most part, we are not engaged in running replicable experiments; the data are what they are; it takes a truly heroic imagination to think in terms of repeated sampling or really taking ever larger samples repeatedly, which is the basis for asymptotic statistical theory. In as much as repeated sampling is the prevailing paradigm, I will discuss this point of view as well as the likelihood view throughout this book, beginning in the next Chapter with a thorough discussion of the Neyman-Pearson theory as implemented in Haavelmo's famous 1944 supplement to *Econometrica*, continuing in Chapter 3 with what I characterize as statistical fantasies, namely Monte Carlo experimentation and bootstrapping, and in Chapter 4 with some discussion of asymptotic statistical theory.

3. The Likelihood Principle

To repeat Edwards' formulation (1972, p. 30):

> "Within the framework of a statistical model, *all* the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data. ...For a continuum of hypotheses, this principle asserts that the likelihood function contains all the necessary information." (Edwards, 1972, p. 30.)

Although clearly implied in what Fisher wrote in the 1920's (see the epigraphs with which this Chapter begins), the likelihood principle, which essentially holds that the likelihood function is the sole

[19] Note that this experiment is quite different than the more often discussed mixture of distributions experiment in which individual observations are sampled randomly from one of two different distributions.

basis for inference, did not come into prominence until the 1950's and 1960's, principally through the work of Barnard, Birnbaum, and Edwards (see the references cited below) written largely in reaction to both the classical Neyman-Pearson (frequentist) and the Bayesian approaches to inference (see Press, 1989, and especially Jeffereys, 1934 and 1961).[20]

Returning to the discussion at the beginning of Section 2: A statistical model consists of a random vector $x \varepsilon X$ of observations having a joint distribution function $F(x;\theta)$, with corresponding density $f(x;\theta)$, depending on the unknown parameters $\theta \varepsilon \Theta$. It is assumed that F is known. The *likelihood function* determined by any given outcome x is defined as the function on Θ equal to $cf(x;\theta)$, $L(\theta,x)$, where c is an arbitrary positive constant which may depend on x but does not depend on θ. Two likelihood functions, $L(\theta,x; E_1)$ and $L(\theta,y; E_2)$ defined on the same parameter space Θ, whether arising from the same "experiment" or from different "experiments," E_1 and E_2, are *equivalent* if their ratio is positive and independent of Θ for all $\theta \varepsilon \Theta$ except possibly at points at which both functions are zero (so that the ratio is undefined). In terms of evidential meaning, a statistic t(x) is sufficient if and only if, for two possible outcomes of E, $t(x) = t(y) \Leftrightarrow$ the evidential meaning of the observations x and y are the same.

The *likelihood principle* asserts that for a given experiment E, the evidential meaning of any outcome x, for inference regarding θ is contained entirely in the likelihood function determined by x. All other aspects of how the data may have been generated are irrelevant, e.g., the sample space, provided, of course, that the sample space itself doesn't depend on θ. It follows that if two "experiments," E_1 and E_2, have pdf's $f(x,\theta)$ and $g(y,\theta)$, respectively, and if for some particular outcomes, x^* of E_1 and y^* of E_2,

$$f(x^*,\theta) = h(x^*, y^*)g(y^*,\theta),\ h(x^*,y^*) > 0, \text{ for all } \theta \varepsilon \Theta,$$

then these outcomes must result in the same inference about θ.

Birnbaum (1962) derives the likelihood principle from the sufficiency principle, discussed in Section 2, and a still more basic assumption, the *conditionality principle*, which was discussed in the preceding section . This principle can be restated: if an "experiment" involving θ is chosen from a collection of possible experiments, *independently of q*, then any experiment not chosen is irrelevant to the statistical analysis.[21] The conditionality principle makes clear the implication of the likelihood principle that any inference should depend only on the outcome observed and not on any other outcome we *might* have observed and thus sharply contrasts with the method of likelihood inference from the Neyman-Pearson, or more generally from a frequentist, approach, in which inference does depend crucially on a hypothetical sequence of experiments, the outcome of but one of which is observed. In particular, questions of unbiasedness, minimum variance, consistency and the like and the whole apparatus of confidence intervals, significance levels, and power of tests, are ruled out of bounds. While maximum-likelihood estimation does satisfy the likelihood principle (and thus sufficiency and conditionality), the frequentist assessment in terms of asymptotic properties is irrelevant. Later in this Chapter, I will try to make clear the role of the maximum of the likelihood function and its Hessian evaluated at the maximum in approximating the whole of the likelihood function for purposes of inference. As, however, suggested in Section 1, to the extent Fisher was

[20] Royall (1997, pp. 24-31) gives an excellent discussion of the likelihood principle and the controversy surrounding it. Lindsey (1996, pp. 74-94) gives a somewhat more discursive discussion including the concepts of sufficiency, ancillarity, and conditionality.

[21] The gist of Birnbaum's argument is as follows: Let S denote the sufficiency principle, C the conditionality principle, and ‹ the likelihood principle. Clearly ‹ implies C; ‹ also implies S since, by the factorization theorem two outcomes giving the same value of the sufficient statistic yield proportional likelihoods. To show that C and S together imply ‹ , consider two experiments E_1 and E_2 with outcomes x_1 and x_2, respectively, such that $L(\theta,x_1; E_1) \propto L(\theta,x_2; E_2)$. For example, consider the mixture experiment E with components E_1 and E_2 with outcomes x_1 and x_2, respectively. Since the likelihood function itself is sufficient, S implies that the evidential meaning of E for the two outcomes is the same. Moreover C implies that the evidential meaning of E_1 with outcome x_1 and E_2 with outcome x_2 are each the same as the evidential meaning of E with corresponding outcome. Hence, C and S imply ‹ .

a frequentist, he regarded the likelihood function itself and particular points on it such as the maximum as statistics susceptible to discussion in terms of sampling properties. At least in this respect, he was quite in accord with the ideas of Neyman and Pearson discussed in the next Chapter. Sampling properties, especially large sample asymptotic, properties of the maximum are discussed in Chapter 6.

Example 4. The following example is given by Barnard, *et al.* (1962) to illustrate incompatibility of the likelihood principle with the Neyman-Pearson theory: Suppose we are tossing a coin in order to find out the probability θ of heads; we toss 20 times and observe 3 heads in 20 tosses: there are 21 possible results, number of heads = 0, 1, 2, ..., 20 with probabilities =

$$\frac{20!}{x!(20-x)!} q^{x}(1-q)^{20-x}, \text{ x = 0, 1, 2, ..., 20.}$$

Inserting the number of heads observed, we obtain the likelihood function

$$f(3,\theta) = 1140\theta^{3}(1-\theta)^{17}.$$

According to the likelihood principle, f(3,θ) contains all the information about θ there is in the observation of 3 heads. (Note that f(3,θ) is maximized for $\overset{\exists}{q}$ = 0.15; the inverse of the negative of the second derivative of the log is $1/\{3/\theta^{2} + 17/(1-\theta)^{2}\}$,at θ = 0.15, = 51/8000 =$(0.08)^{2}$; so the ML inference is θ = 0.15 with a standard error of 0.08.) Now, say Barnard, *et al.*, suppose we had performed another experiment: We toss until 3 heads are observed. It turns out that it takes 20 tosses to achieve this outcome. the probability of y throws to achieve 3 heads, y = 3, 4, 5, ... is

$$g(y,\theta) = [(y-1)!/2!(y-3)!]\theta^{3}(1-\theta)^{17}.$$

For y = 20, therefore, the likelihood of y = 20 is

$$g(20,\theta) = 171\theta^{3}(1-\theta)^{17},$$

which is not the same, but which is proportional to f(3,θ) with the factor of proportionality 20/3 and gives exactly the same ML estimate and standard error as before. Yet as Barnard, et al., show, the Neyman-Pearson theory rejects the null θ = 0.5 at the 0.0013 level in the first case and the 0.0004 level in the second. (Exercise: How would you find the probabilities of no more than 3 heads in 20 tosses in the first case, and of more than 20 tosses to achieve 3 heads in the second?) So, for the Neyman-Pearson theory it matters not only what the outcome is, but also what we intended to do when we started the experiment. (That is to say, the classical N-P approach leads to a result that depends on which of a set of possible experiments was chosen.) What we intended is quite irrelevant from the likelihood point of view. To define the N-P significance level, we need to know the totality of all possible outcomes with which to compare the given outcome; but from a likelihood point of view all that matters is the ratio
f(3, 0.15 or any other value)/f(3,0.5) or g(20,0.15 or any other value)/g(20,0.5) and these are the same regardless of our intentions when we began.

However, Barnard, *et al.*, (1962, p. 323) make the following interesting comment concerning this example:

> "In advocating the likelihood principle as a basic proposition in the theory of inference, it is necessary to emphasize that it is not proposed that it should be used uncritically unless the model is known very precisely. Suppose, for example, that the results from two independent sequences each consisting of 100 Bernoulli trials indicated 50 successes in both cases. If, however, the successes occurred haphazardly in one sequence and in the form 01010101...0101 in the other, the inference for the two

> sequences would be very different despite the fact that the likelihoods were the same; one would investigate the systematic pattern of 0's and 1's in the second sequence.
>
> "In general, in order to write down the likelihood function, it is necessary to *assume* that the model is known apart from some parameters which are to be estimated. Some of these parameters may characterize lack of independence amongst the errors. If the model has to be guessed, then the nature of the inference becomes much less precise than is suggested by the formal statement of the likelihood principle. ...
>
> "When the distribution of the errors is not known, one would be interested in the *robustness* of the likelihood function with respect to changes in the form of the error distribution and in the corresponding robustness of any estimation procedures one would care to advocate."

The likelihood function, and in particular its maximum, have another desirable property, *invariance*, which was an important consideration for Fisher as early as 1912 in rejecting inverse probability (Aldrich, 1997, p. 165). I cite from Aldrich's quote: "...the probability that the true values lie within a region must be the same whether it is expressed in terms of θ or φ [where θ is a transformation of φ]." Of course relative probabilities will be unaffected by transformation, and *a fortiori*, likelihoods, but absolute probabilities not unless the Jacobian of the transformation is one; hence, Fisher later argued that inverse probability where inference is based on the posterior for noninformative prior ought to be rejected because conclusions drawn in this manner would not be invariant under parameter transformation.

4. Maximum Likelihood or Maximum Support

The likelihood principle is clearly incomplete from the standpoint of inference since it nowhere states how the evidential meaning of the likelihood function is to be determined. To the principle, therefore, "likelihoodists" generally append the *method of support* (a term coined by Jeffereys, 1934). The *support function* is defined as the natural logarithm of the likelihood function. Since the likelihood function incorporates an arbitrary constant, the support function is defined only up to the addition of an arbitrary constant. Conventionally this constant is often taken to be the value which makes support at the maximum equal zero. In multiplicative terms, this is equivalent to normalizing the likelihood function by dividing it by its value at the maximum. Only relative support for a particular parameter value over another can be interpreted in any case, so the constant disappears when looking at the difference between support values of different parameter values. The *method of maximum support* is thus the *method of maximum likelihood*. But the interpretation of the parameter value which yields this maximum and of the inverse of the negative of the Hessian at the point of maximum is different than in the frequentist interpretation in terms of asymptotic properties. The likelihood interpretation of these magnitudes is in terms of a quadratic approximation to the support function in the neighborhood of its maximum. I will return to this point below in section 5. In the remainder of this Chapter I will use the terms *support* and *log likelihood* interchangeably.

Example 5. The likelihood for the binomial parameter θ with n_1 successes and n_2 failures in n = $n_{1\,+}\,n_2$ trials is $\mathrm{L}(q;n_1,n_2)=kq^{n_1}(1-q)^{n_2}$ so that the support function is

$$\mathrm{S}(q;n_1,n_2)=const.+n_1\ln q+n_2\ln(1-q)\,.$$

The value of θ which maximizes this is n_1/n which suggests choosing the constant =

$$-\left\{n_1\ln(\frac{n_1}{n})+n_2\ln(\frac{n_2}{n})\right\}.$$

Thus, the normalized support is

$$\mathrm{S}(q;n_1,n_2)=-\left\{n_1\ln(\frac{n_1}{n})+n_2\ln(\frac{n_2}{n})\right\}+n_1\ln q+n_2\ln(1-q)\,.$$

In general for the multinomial case with k categories, the normalized support function for the parameters $\theta_1,\theta_2,\theta_3$, ...,θ, such that $\sum_1^k q_i = 1$ is

$$\mathrm{S}(q_1, q_2, q_3, \ldots q_k; n_1, n_2, \ldots n_k) = -\left\{\sum_i n_i \ln(\frac{n_i}{n})\right\} + \sum_i n_i \ln q_i \ ,$$

where the n_i, i = 1,2,..., k, are the numbers observed in each category such that $\sum_i n_i = n$, the total number of trials.

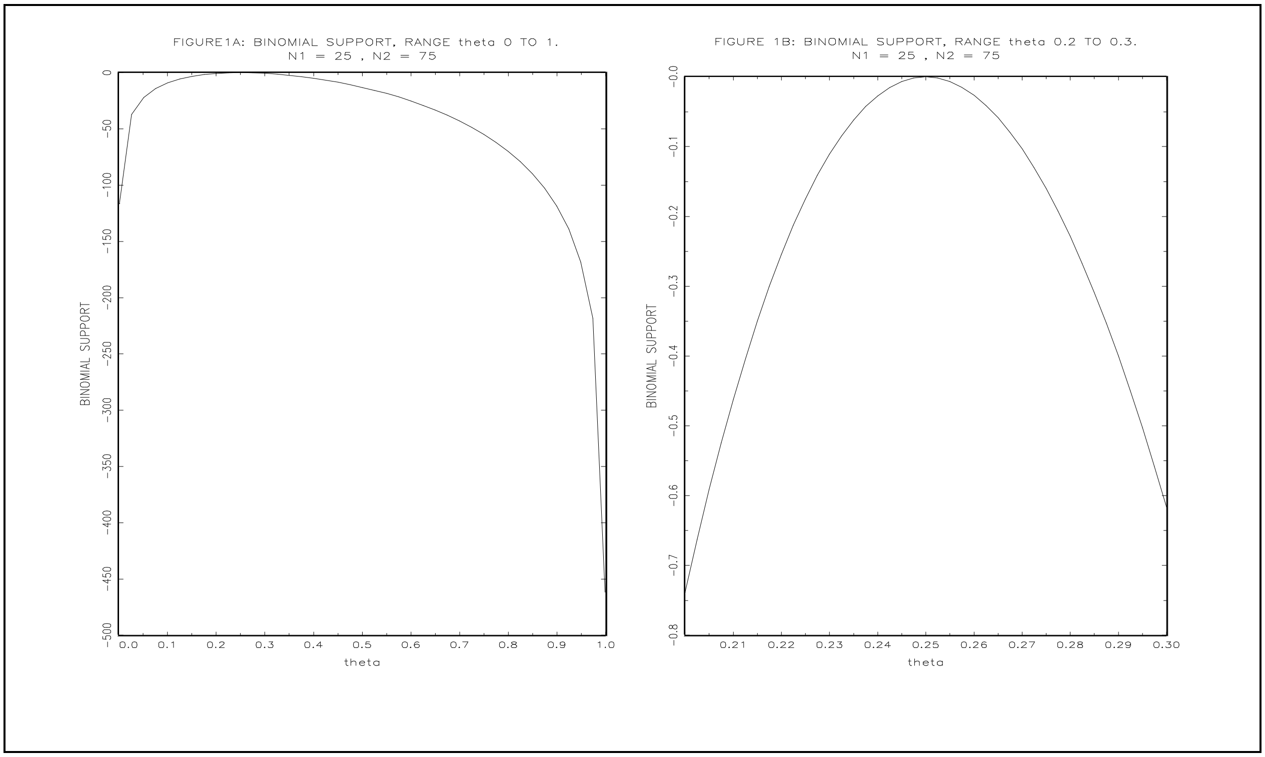

Figure 1: The Binomial Support Function for n_1 = 25, n_2 = 75

Note the support function is smooth and nearly quadratic in the vicinity of the maximum.

*Example 6, Univariate normal, n(m, s^2):*For the univariate normal distribution with mean μ and variance σ^2, the normalized support function for the N observations $x_1, x_2, \ldots, x_N$ is

(1) $$S(m, s^2) = \frac{N}{2}\ln s^2 + \frac{N}{2} - \frac{N}{2}\ln s^2 - \frac{N[s^2 + (\bar{x} - m)^2]}{2s^2} \ ,$$

where $$\bar{x} = (N^{-1})\sum_1^N x_i \ ,$$ $$s^2 = (N^{-1})\sum_1^N (x_i - \bar{x})^2 \ .$$

The values of μ and σ^2 which maximize the function S are $\hat{m} = \bar{x}$ and $\hat{s}^2 = s^2$.

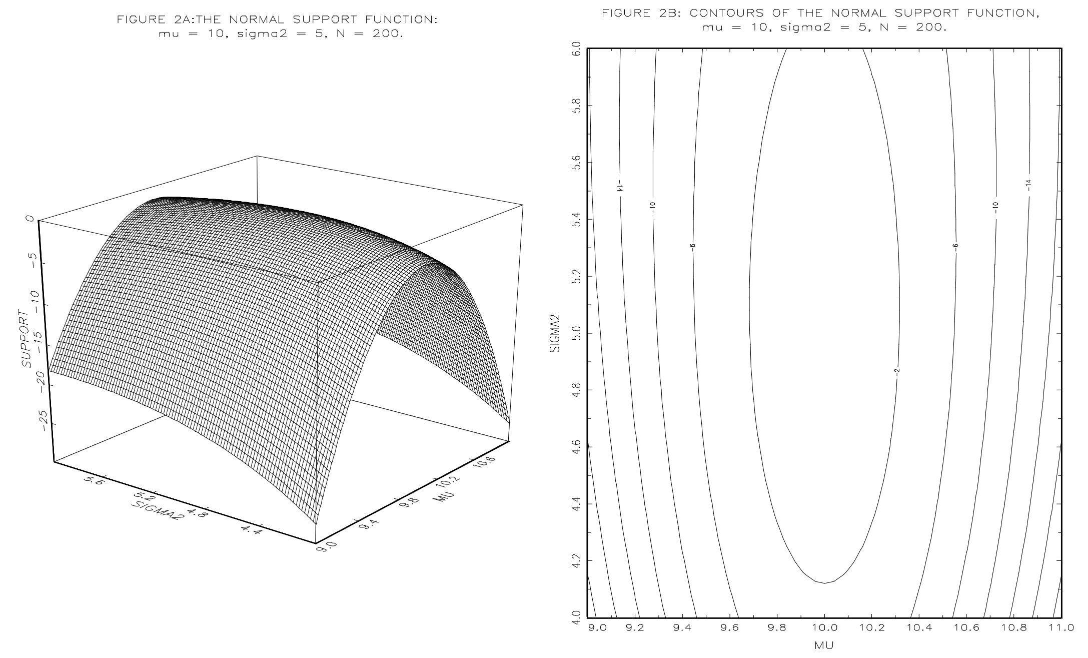
FIGURE 2A:THE NORMAL SUPPORT FUNCTION:
mu = 10, sigma2 = 5, N = 200.

FIGURE 2B: CONTOURS OF THE NORMAL SUPPORT FUNCTION,
mu = 10, sigma2 = 5, N = 200.

Figure 2: The Normal Support Function, m = 10, s^2 = 5, n =200

As in the previous example, note the extremely well-behaved support function. unfortunately such good behavior is the exception rather than the rule.

5. Incorporating Additional Evidence or Beliefs in a Likelihood Approach to Inference

Suppose we have two events A and B with joint probability of occurrence P[A, B]. It is elementary that P[A|B]P[B] = P[A, B], the joint probability equals the product of the marginal and the conditional. Thus far we have considered only a single "experiment" and induction from it, but we may have prior information about the parameters about which we seek to draw inferences from previous experiments or from theory; indeed, the parametric statistical model underlying our analysis is a prior belief. How should such prior information or belief be incorporated in the way in which we draw inferences? One possibility is to cast the prior information in the form of a probability distribution and make use of the above result. This is essentially the Bayesian approach in which the likelihood is interpreted as a probability and multiplied by the prior to obtain the posterior probability of the parameters from which inferences are to be drawn. To restate:

Bayes' Theorem: Given a likelihood L(θ|x) and a prior probability density defined on θ, p(θ), the posterior density for θ is

$$p(q|x) = cp(q)L(q|x) \text{ where } c^{-1} = \int_{\Theta} p(q)L(q|x)dq \text{ when } q \text{ is continuous.}$$

Example 6: Univariate Normal population with Known Variance and Unknown Mean. Consider an iid sample from n(μ,σ^2) where σ^2 is known and μ is unknown.

$$f(x|m) = \left(\frac{1}{\sqrt{2ps^2}}\right)^n \exp\left\{-\tfrac{1}{2}\frac{\sum\limits_i (x_i - m)^2}{s^2}\right\} \propto \exp\left\{-\tfrac{1}{2}\frac{\sum\limits_i (x_i - \bar{x})^2 + n(\bar{x} - m)^2}{s^2}\right\}$$

$$\propto \exp\left\{-\tfrac{1}{2}\frac{n(\bar{x} - m)^2}{s^2}\right\} \text{ when } s^2 \text{ is assumed known.}$$

Under the prior $p(m) = n(m_0, s_0^2)$ where m_0 and s_0^2 are known constants the posterior density is also normal with a mean which is a weighted average of the prior and posterior means with weights which depend on the known prior and posterior variances:

$$m_0\left(\frac{1/s_0^2}{1/s_0^2 + n/s^2}\right) + \bar{x}\left(\frac{n/s_0^2}{1/s_0^2 + n/s^2}\right).$$

As n→∞, the prior flattens out and we are left with an expression proportional to the likelihood, i.e., the data "swamp" the prior. The prior $p(m) = n(m_0, s_0^2),\ \ s_0^2 \to \infty,$ is said to be *noninformative.* In this case also the data dominate, however the limiting prior is not a proper distribution in this case. In general, however, truly noninformative priors are difficult to obtain, Jeffreys (1939,1961) notwithstanding.

In Bayesian inference, where the prior actually comes from is not generally discussed. Jeffreys (1939, 1961, pp. 117-125) suggests that, when prior beliefs are diffuse, we use a so-called noninformative prior, but, as Example 5 suggests, these are not so easy to obtain. Zellner (1971, pp. 41-53) gives a thoughtful and detailed discussion. Another useful discussion of "vague" priors is given by Press (1989, pp. 48-52), who reports the following suggestion by Savage: Since the posterior density depends only on the product of the prior and the likelihood, it suffices to be vague about θ, the parameter about which we seek to draw an inference, we need only take the prior on θ to be uniform over the range in which the likelihood function is not small. Outside of this range, the prior can be anything since it can't possibly affect the posterior by much. So, in this way good Baysians can, as it were, sneak up on the likelihood approach. But, it should be emphasized, both the frequentists and those following the likelihood approach would reject such an interpretation. For all its superficial resemblance, the likelihood function is *not* a probability; it is a statistic and as such we may or may not wish to study its properties in repeated samples depending on how far along the scale we are from the Bayesian to the strict sampling approach.

Nonetheless, the Bayesian approach suggests a way of combining results from more than one "experiment" in the context of likelihood: simply multiply the two likelihoods or, alternatively, sum the support functions. Thus, the Bayesian point of view offers a convenient way of incorporating additional evidence or beliefs. Note that, as an important special case of this result, the log likelihood or support for more than one independent observations is simply the sum of the values for each of the observations, a fact which may be useful in a computational context.

6. The Inferential Meaning of Support

It is clear that the difference in the value of the support function at two different values of a parameter has the significance that the value for which support is greater is more consistent with the observed data than the value of lesser support. What we have is essentially a likelihood ratio test without the frequentist apparatus of asymptotic chi-square. It is also clear that the values of parameters for which maximum support is obtained (that is, the maximum-likelihood estimates), especially if the maximum is unique, have a special significance in relation to other possible values. Moreover, how sharply defined such a maximum of the likelihood function is, if a unique maximum exists, is also clearly relevant to any inference we may wish to draw. On the negative side, a poorly behaved likelihood function, for example, one having ridges of equal likelihood, many local maxima, or a maximum on the boundary of an a priori admissible region of the parameter space, is generally indicative of an incompletely or ill-formulated underlying statistical model.

From a frequentist point of view what matters about the likelihood function is only its maximum and curvature in the neighborhood of the maximum, and all the desirable properties and the assessment of the reliability of the maximum-likelihood estimates are only asymptotic. Greene (1997, pp.133-140) gives a very brief discussion of these matters; Davidson and MacKinnon (1993, Chapter 8, pp.243-287) give a more complete and rigorous discussion; a more intuitive discussion with many econometric examples is given by Cramer (1986). That only the maximum and the Hessian at the maximum are all that matters from a frequentist point of view is perhaps not surprising in view of the fact that for the mean of a normal distribution the quadratic approximation is exact (see the discussion below) and because of the central limit theorem in its many forms many estimators, including ML estimators in regular cases, tend to normality in distribution.

Let S(θ|x) = logL(θ|x) be the log likelihood or support for a parameter vector θ given observations x and let $\hat{\theta}$ be a value of θ maximizing this function; expand the function in a Taylor's series about $\hat{\theta}$:

$$(2) \qquad S(\theta|x) = S(\hat{\theta}|x) + \frac{1}{2}(\theta - \hat{\theta})' \frac{\partial^2 S(\hat{\theta}|x)}{\partial\theta\partial\theta'}(\theta - \hat{\theta}) + \text{ additional terms},$$

since, at the maximum, the derivatives $\partial S(\theta|x)/\partial\theta$ vanish. $-\frac{\partial^2 S(\hat{\theta}|x)}{\partial\theta\partial\theta'}$, the negative of the Hessian at the maximum of the log likelihood function is called the *Fisher information matrix*, or sometimes just the *information matrix*. It measures the curvature of the likelihood function in the vicinity of the maximum.

Example 5, Continued, Binomial Support. For the normalized binomial support function discussed above, we have the following quadratic approximation (note that normalization makes $S(\hat{\theta}; n_1, n_2) = 0$[22]):

$$S(\theta; n_1, n_2) = -\frac{1}{2}(\theta - \frac{n_1}{n})^2 (\frac{n}{n_1} + \frac{n}{n_2}).$$

[22] The maximized value of S is $n_1 \ln(\frac{n_1}{n}) + n_2 \ln(\frac{n_2}{n})$; the constant of proportionality has been chosen as the negative of this value.

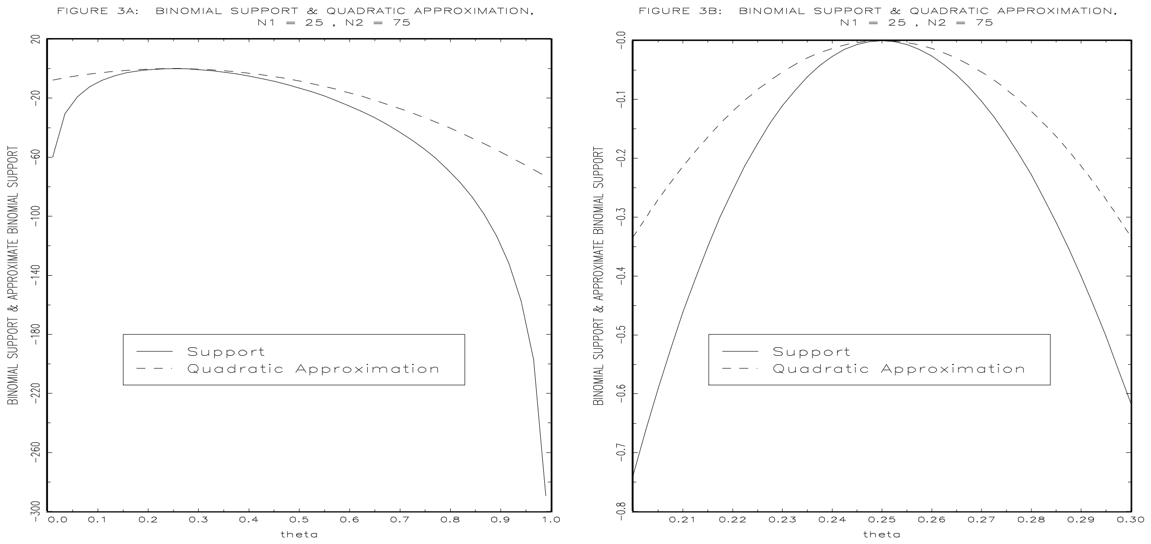

Figure 3: Binomial Support and Quadratic Approximation, n_1 = 25, n_2 = 75

Clearly, a quadratic approximation is good in the vicinity of the maximum, but increasingly poor as one deviates, particularly near the a priori bounds of the probability parameter at 0 and 1. Were the maximum to occur very near one of these bounds, it is clear that the common use of the quadratic approximation to estimate standard errors could lead to misleading results.

Example 6, Continued, Univariate Normal Support: The quadratic approximation to the normalized normal support function is

(3)
$$S(m, s^2) = -\frac{N}{2}\left\{\frac{(\bar{x}-m)^2}{s^2} + \left(\frac{s^2 - s^2}{s^2}\right)^2\right\}.$$

If we set $\sigma^2 = s^2$, its maximum likelihood estimate, the second term in (2) vanishes; comparing (1) and (3), we see that the quadratic approximation is exact for σ^2 at its maximizing value. However, this is not true for the variance when μ is set to its maximizing value. "Slices" through the support function, or so-called "likelihood profiles," for μ and σ^2 are shown in Figures 4A and 4B below; in 4A the quadratic approximation is indistinguishable from the likelihood profile itself. Note that, for the variance, the quadratic approximation is not in fact very good at even moderate distance from the maximizing value.

An alternative way of viewing the three-dimensional support function in two dimensions is to concentrate the likelihood function as explained in the next section. Figures 4C and 4D below exhibit these concentrated support functions and corresponding quadratic approximations:

For μ:
$$\bar{S}(m, \bar{s}^2(m)) = -\frac{N}{2}\frac{(m-\bar{x})^2}{s^2}.$$

The quadratic approximation is the same.

For σ^2
$$\bar{S}(s^2, \bar{m}(s^2)) = \frac{N}{2}\log(\frac{s^2}{s^2}) + \frac{N}{2}(1-\frac{s^2}{s^2}),$$

with quadratic approximation $-(N/2)(s^{2} - s^2)^2/(s^2)^2$. Note that for the mean the quadratic approximation is exact.

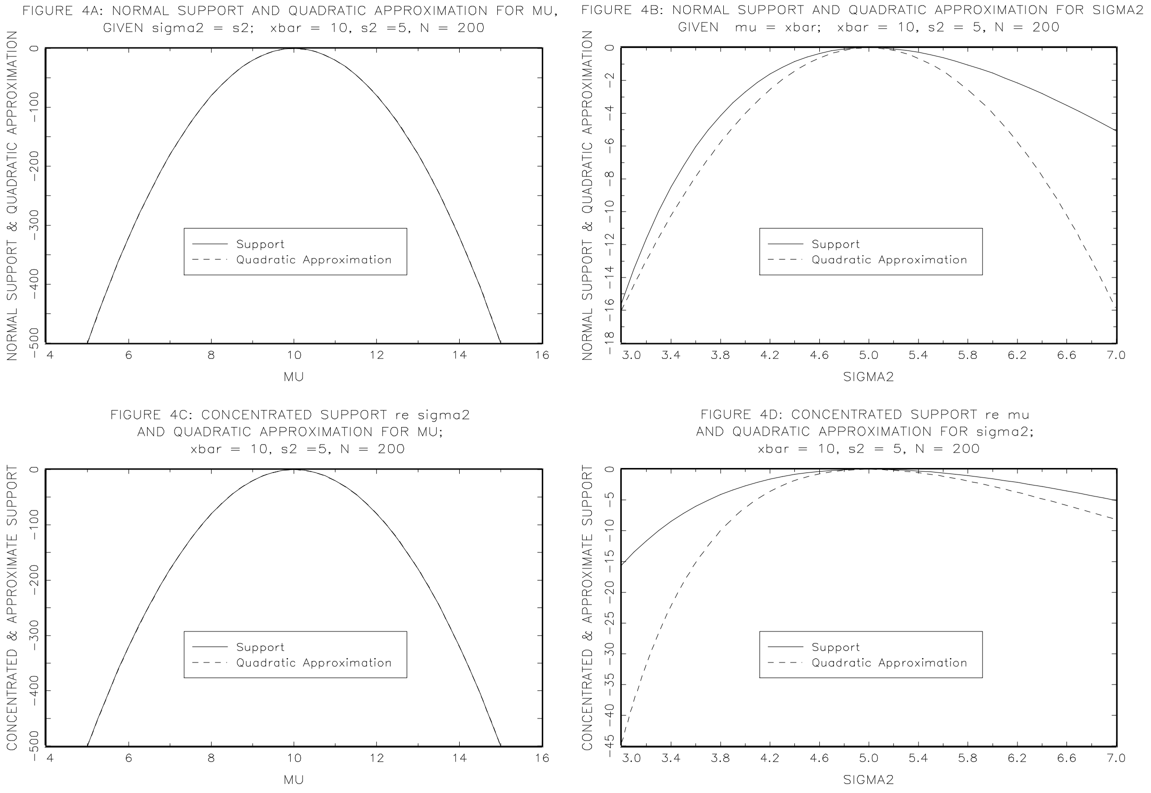

Figure 4: Likelihood Profiles and Concentrated Likelihoods for the n(m, s^2)

Example 6 illustrates that when we are dealing with only one or two parameters looking at the whole of the likelihood or support function is feasible, although some summary measures may be helpful. For three or more parameters, however, it is no longer possible to examine the whole of the support function. In this case, concentrating the likelihood function and corresponding support function may be helpful, and looking at a quadratic approximation to the support function in the neighborhood of the maximum may be revealing. In the next section, I look at partial maximization or concentration and slicing in greater depth. Quadratic approximation in the multidimensional parameter case is the general approach in maximum likelihood estimation. The point of maximum support, particularly if unique, obviously has considerable intuitive appeal. A quadratic approximation at that point is likely to be pretty good if we want to consider only points quite nearby and has the added advantage of being directly interpretable from a frequentist point of view in terms of the information matrix of asymptotic maximum-likelihood theory. The disadvantage is that except for cases, such as the mean or regression function associated with a normal distribution, for which the quadratic approximation is exact, the approximating function may be quite wide of the mark. I will discuss this interpretation in Chapter 6. How tightly focused the support function is around the maximum value is an indicator similar to the conventional confidence interval in the frequentist approach to inference.[23]

[23] Edwards (1972, pp. 71-72) suggests using what he calls the *m-unit support region* which he defines as "...that region in the parameter space bounded by the curve on which support is *m* units less than the

7. Parameter Reduction Techniques: Slicing and Concentrating

> You are living on a Plane. What you style Flatland is the vast level surface of what I may call a fluid, or in, the top of which you and your countrymen move about, without rising above or falling below it.
>
> I am not a plane Figure, but a Solid. You call me a Circle; but in reality I am not a Circle, but an infinite number of Circles, of size varying from a Point to a Circle of thirteen inches in diameter, one placed on the top of the other. When I cut through your plane as I am now doing, I make in your plane a section which you, very rightly, call a Circle. For even a Sphere--which is my proper name in my own country--if he manifest himself at all to an inhabitant of Flatland--must needs manifest himself as a Circle.
>
> E. A. Abbott, *Flatland,* 1884.

It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.
The First approached the Elephant,
And happening to fall
Against his broad and sturdy side,
At once began to bawl:
"God bless me! but the Elephant
Is very like a wall!"
The Second, feeling of the tusk,
Cried, "Ho! what have we here
So very round and smooth and sharp?
To me 'tis mighty clear
This wonder of an Elephant
Is very like a spear!"

The Third approached the animal,
And happening to take
The squirming trunk within his hands,
Thus boldly up and spake:
"I see," quoth he, "the Elephant
Is very like a snake!"
The Fourth reached out an eager hand,
And felt about the knee.
"What most this wondrous beast is like
Is mighty plain," quoth he;
" 'Tis clear enough the Elephant
Is very like a tree!"
The Fifth, who chanced to touch the ear,
Said: "E'en the blindest man
Can tell what this resembles most;
Deny the fact who can
This marvel of an Elephant
Is very like a fan!"

The Sixth no sooner had begun
About the beast to grope,
Than, seizing on the swinging tail
That fell within his scope,
"I see," quoth he, "the Elephant
Is very like a rope!"
And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong!
Moral:So oft in theologic wars,
The disputants, I ween,
Rail on in utter ignorance
Of what each other mean,
And prate about an Elephant
Not one of them has seen!

John Godfrey Saxe , "The Blind Men and the Elephant: A Hindoo Tale," 1880

In this section, I discuss the two techniques for looking at multidimensional likelihood function which were introduced in Example 6 for looking at the univariate, two-parameter normal support function: *slicing* and *concentrating.* Both slicing and concentrating are discussed in the literature of parametric inference under the heading of *likelihood profiles.* (See Lindsey, 1996, pp. 111-114.) But the important distinction between the two is not generally made clear.

maximum." In the case of one parameter, for which the support function has a unique maximum, the m-unit support region is easy to determine graphically. In the two parameter case, the support region can be determined directly from a contour plot: For example, in the normal case considered in example 6, the 2-unit support region is enclosed by the innermost contour curve plotted. (See Figure 2.) Support regions smaller than m = 2 would require a more detailed contour plot. The difficulty with such support regions in practice is that they may be extremely difficult to calculate for more than two parameters. Even for one parameter, algebraic, as opposed to graphical determination, is not trivial. In the binomial case, for example, we have to solve the nonlinear equation in θ:

$$n_1 \ln q + n_2 \ln(1-q) = m .$$

The suggestion made by Birnbaum (1962) of calculating *intrinsic confidence intervals* is even more problematic. Instead of normalizing the likelihood function by dividing by the maximum value, normalize it so that the area under it is one, that is, write it as the probability density from which it is derived. Then find an interval or region around the maximum in the parameter space which encloses an area or volume equal to some conventionally chosen value such as 90%. Of course, there will in general be many such regions depending on how we choose to center them. But that is not the principal problem; the chief difficulty is that in more than two dimensions the evaluation of the integrals necessary to determine the region may be extremely difficult.

a. Slicing

To slice, section or slice the support function along the plane of all but one or two of the parameters; in the case in which all but one of the parameters have been eliminated in this way, we are back to a two-dimensional plot; when we have done this for all but two parameters we can plot a three-dimensional surface and an associated contours of equal support. The latter is particularly useful if we want to examine how two of the parameters interact with one another; for example, the two transformation parameters in a double Box-Cox transform, or in the two-parameter transformation for heteroskedasticity, both of which are discussed below. It would be natural to choose the values of all but one or two of the parameters equal to the maximizing values. However, slicing in this way must be carefully distinguished from the technique of *concentrating* the likelihood function, which is also a useful technique in finding the maximum.

Slicing is essentially what one typically does in viewing a three-dimensional surface when we look at a contour map: We take a slice through the surface in the direction parallel to the plane of the two arguments. A slice can, of course, be thought of more generally as any lower dimensional hyperplane, whether parallel to the plane defined by the axes of a subset of arguments or in some other direction. In four dimensions, a slice in any two-dimensional plane, which eliminates all the arguments but two, yields a surface of the functional values in three dimensions. (If you have trouble visualizing this, try reading *Flatland* by Edwin A. Abbott, 1884.) Fixing, or conditioning on, the values of any subset of parameters is obviously a way of defining a particular hyperplane corresponding to the remaining parameters; in this instance, those values which maximize the overall support.[24]

b. Partial maximization. Concentration

On the other hand, if a point is chosen on the hyperplane, on which we want to view the support, and the values of the other parameters are chosen to maximize support *at that point*, we are dealing with a different way of looking at the likelihood. In discussions of maximum likelihood, *concentration of the likelihood function* with respect to a subset of parameters corresponds to selecting a hyperplane for the remaining parameters in just this way.[25] Sometimes we say that we are "maximizing out" the deselected parameters. In the method of maximum likelihood, for example, it frequently turns out that, *given* the values of one or two of the parameters, it is very easy to maximize with respect to the remaining ones (see the examples of the double Box-Cox transformation and the transformation for heteroskedasticity given below).

Two-step maximization is commonly employed to reduce the dimension of a maximum problem: Suppose we want to max f(θ_1, θ_2) re: θ_1 and θ_2. Under very general conditions, we can "hold θ_2 fixed" and

$$\max_{\theta_1} f(\theta_1, \theta_2) = f(\theta_1(\theta_2)) = f^*(\theta_2)$$

where $\theta_1(\theta_2)$ is the value of θ_1 which maximizes f(θ_1, $\theta_2 = \theta_2$). Then varying θ_2, find

$$\max_{\theta_2} f^*(\theta_2) = f^*(\theta_1(\theta_2)) = f^{**}$$

[24] Slicing is also suggested by Larkin and Kadane (1990, p. 459).

[25] The best discussion of the technique of concentrating the likelihood function that I know of is to be found in Koopmans and Hood (1953, pp. 156-158), in which they derive the *limited information maximum likelihood estimates* for the parameters of a single equation of a system of simultaneous structural equations; see Chapter 25 below. In the examples presented in the remainder of this Chapter likelihoods are concentrated analytically, but it is obvious that the technique can be carried out numerically as well. Some examples of such concentrated likelihood or support functions are give later in this book.

is the maximum maximorum.

The function $f^*(\theta_2)$ is said to be "concentrated with respect to θ_1." That is, θ_1 has been "maximized out," so that the concentrated function is a function of θ_2 only.

A common examples of this is concentrating the univariate normal likelihood function by maximizing out the mean, then maximizing the concentrated likelihood function re: σ^2. (Exercise: Try this the other way: first re: σ^2, then re: μ.)

Example 7: First-order residual autocorrelation.[26] The log likelihood function, or support function, for the case of first-order residual autoregression is

$$(4)\qquad S(b, r, s^2 | y_t, x_t, t = 1,\ldots,T) =$$
$$-\frac{T}{2}\log 2\pi - \frac{T}{2}\log\sigma^2 + \frac{1}{2}\log(1-\rho^2)$$
$$-\frac{1}{2\sigma^2}\left\{(1-\rho^2)(y_1 - x_1'\beta)^2 + \sum_{t=2}^{T}\left[(y_t - x_t'\beta) - \rho(y_{t-1} - x_{t-1}'\beta)\right]^2\right\}.$$

Holding ρ fixed at some value in the interval (-1, 1) shows that β and σ^2 are just the GLS estimates. Call these values $\beta(\rho)$ and $\sigma^2(\rho)$; *they are functions of r.* To concentrate the support function, substitute in

$$(5)\qquad S^*(r | y_t, x_t, t = 1,\ldots,T)$$
$$= -\frac{T}{2}\log 2\pi - \frac{T}{2}\log\sigma^2(\rho) + \frac{1}{2}\log(1-\rho^2) - \frac{1}{2\sigma^2(\rho)}\cdot T\cdot\sigma^2(\rho)$$
$$= -\frac{T}{2}(\log 2\pi + 1) - \frac{T}{2}\log\sigma^2(\rho) + \frac{1}{2}\log(1-\rho^2).$$

This shows that the ML estimates of ρ and of β and σ^2 are obtained by minimizing

$$(6)\qquad T\log\sigma^2(\rho) - \log(1 - \rho^2),$$

where $\sigma^2(\rho)$ is the GLS estimate of σ^2, *given r*. Of course, this function is determined numerically by the least-squares procedure.

Note, that since the term with T dominates for large T, this also shows that the *asymptotic* ML estimates can be obtained by iterating on ρ to obtain the smallest GLS estimate of σ^2. But for small samples the term $(1 - \rho^2)$ is more important and the iterated GLS (sometimes called Yule-Walker) will not be the same approximately as the ML estimates.

In this simple case, $S^*(\rho)$ is a function of only one parameter which is a priori constrained to lie in the interval (-1, 1), so a *grid search* is feasible and easy. I will take up the general problem of maximizing support, or likelihood, including grid search methods, in Chapter 5 below. The concentrated likelihood function for a numerical example is shown in Figure 5 below.

What happens if the maximum occurs on a boundary point, $\rho = -1$ or 1? At such a point any quadratic approximation is clearly invalid.

[26] Time series problems, of which estimation of a first-order autoregression is the simplest case, are considered in depth in Part 3.

The information matrix can be obtained by differentiating the original likelihood function with respect to β, σ^2 *and* ρ. It is *not* block diagonal (as in the standard regression case for β and σ^2) and therefore the correctly estimated variance-covariance matrix for β is *not the same as the GLS estimate for the maximizing value of r*.

Example 8: Heteroskedasticity. Consider a regression in which the residual variance is a function of an exogenous variable z_t. If z_t is a dummy variable, it simplifies matters considerably since, essentially, it permits us to divide the observations into *two* groups, each with a different but constant variance. But let us consider the more general problem here: Let

(7)
$$s_t^2 = \exp(g_0 + g_1 z_t) \ .$$

The problem is to find γ_0 and γ_1 <u>and</u> β and σ^2 in

(8)
$$\mathrm{y_t = x_t'\beta + u_t}$$
$$\mathrm{Eu_t = 0, \quad Eu_t u_t' = \sigma_t^2, \quad t = t'}$$
$$= 0, \text{ otherwise.}$$

The density of $(y_t\,|x_t\,)$ t = 1, ..., T is given by

(9)
$$f(y_1,\ldots,y_T|x_1,\ldots,x_T) = \left(\frac{1}{\sqrt{2p}}\right)^T |\Omega^{-1}|^{\frac{1}{2}} \; e^{-(1/2)(y-Xb)'\Omega^{-1}(y-Xb)}$$

where

$$\Omega = \begin{bmatrix} \exp(g_0 + g_1 z_1) & 0 & \Lambda & 0 \\ 0 & \exp(g_0 + g_1 z_2) & \Lambda & 0 \\ \mathrm{M} & \mathrm{M} & & \mathrm{M} \\ 0 & 0 & \Lambda & \exp(g_0 + g_1 z_T) \end{bmatrix}$$

so that

$$\left|\Omega^{-1}\right|^{1/2} = \prod_{t=1}^{T} \frac{1}{\sqrt{\exp(g_0 + g_1 z_t)}} \ .$$

Note: When maximizing numerically it pays to ensure that the variance s_t^2 (and standard deviation) will always be positive (real). This is always the case in this example because the exponential function is never ≤ 0, but other functions might be used in practice for which this could be a problem.

Thus, the log-likelihood or support function is

(10)
$$S(b, g_0, g_1|y, X) =$$
$$\frac{T}{2}\log 2p - \frac{1}{2}\sum_{t=0}^{T}(g_0 + g_1 z_t)$$
$$-\frac{1}{2}\sum_{t=1}^{T}\frac{(y_t - x_t'b)^2}{\exp(g_0 + g_1 z_t)} \ .$$

Things are not so simple now because the parameters γ_0 and γ_1 enter the two terms of S( ) *with the sequence* z_1, ..., z_T. However, the problem can be further simplified by setting $\sigma^2 = \exp(\gamma_0)$; then, the support function becomes

(11) $S(b, s^2, a \;|\; y, X) =$

$$\frac{T}{2}\log 2p - \frac{T}{2}\log s^2 - \frac{a}{2}\sum_{t=0}^{T}(z_t)$$
$$-\frac{1}{2s^2}\sum_{t=1}^{T}\frac{(y_t - x_t'b)^2}{\exp(a\;z_t)},$$

with $\alpha = \gamma_2$.

In this case the number of parameters combined with the individual observations is reduced from two to one, but the problem remains.

In either case, we can still proceed in a stepwise fashion if we understand the nature of a function properly.

What is a function? A function is a numerical recipe or *algorithm* which takes the arguments as "ingredients," or inputs, and outputs the final result which may be a whole meal or merely one dish, that is, multiple values or a single value. (In formal mathematical usage, the term function is usually reserved for algorithms which deliver a single-valued output. My usage here corresponds to what is often called a *procedure* or a *subroutine* in computer programming languages.) Most functions, with which we deal, output a unique value or vector of values for each vector of arguments, but for a given output there may be more than one possible input combination which gives rise to it.

When you type sin x, log x, or e^x in your computer, your machine or system is not generally "hard-wired" to return a value of the function to you for a specified value of x. Rather, your program calls up a recipe or algorithm which specifies a series of arithmetical operations which yield the correct functional value. These recipes are generally based on series expansions of the function in question.

Example 7, Continued. The GLS algorithm which delivers $s^2(r)$ is an example of such a recipe. It would be possible, in principle, to differentiate this function analytically, but difficult in practice.

Example 8 Continued: In the case of the concentrated support function based on (10) or (11), which you want to plot, and perhaps eventually to maximize, the recipe does not take an analytic form but rather contains, as part of the recipe, a GLS regression in much the same way as a real recipe might call for chicken stock or a roux.

If we knew γ_0 and γ_1, we could maximize S in (7) with respect to β (there is no σ , as such). In (8), we fix α and compute the adjustment factor $\exp(\alpha z_t)$ for each observation; then the problem becomes the standard ML linear regression problem: given α, we have to maximize $S(b, s^2, a \;|\; y, X)$ with respect to β and σ^2 .

To maximize (10) re: β, which occurs only in the last term, do a GLS regression and minimize the transformed residual sum of squares. This yields $\beta(\gamma_0, \gamma_1)$ and

$$RSS(g_0, g_1) = \sum_{t=0}^{T}\frac{[y_t - x_t'b(g_0, g_1)]^2}{\exp(g_0 + g_1 z_t)},$$

so that the concentrated support function is

(12)
$$S^*(\gamma_0,\gamma_1|y,X) = S(\beta(\gamma_0,\gamma_1),\gamma_0,\gamma_1|y,X) = \frac{1}{2}\log 2\pi - \sum_{t=0}^{T}\log\sqrt{\exp(\gamma_0+\gamma_1 z_t)} - \frac{1}{2}RSS(\gamma_0,\gamma_1)\ .$$

$S^*(\gamma_0,\gamma_1)$ is a function only of the two parameters γ_0 and γ_1. It is not given analytically. The ingredients in the recipe are (y, X, and z) and (γ_0, γ_1). If you have a given set of data y, X and z and give the algorithm values of γ_0 and γ_1, it first makes a "roux," RSS(γ_0, γ_1), and then cooks up S*(γ_0, γ_1) from (12).

When the support function is given by (11), we can easily find the maximizing values of β and σ^2 given γ as functions of γ: $\hat\beta(\gamma)$ and $\hat\sigma^2(\gamma)$ by regressing the transformed $y_t^* = \dfrac{y_t}{\exp \gamma z_t}$ on the transformed $x_t^* = \dfrac{x_t}{\exp \gamma z_t}$. The residual sum of squares from this regression, RSS(γ) is a function of γ. Substituting this in (11):

(13)
$$S(\hat\beta(\alpha),\hat\sigma^2(\alpha),\alpha\mid y,X) = S^*(\gamma|\,y,X) = -\frac{T}{2}\{\log 2\pi + 1 + \alpha\bar z + \log\frac{RSS(\gamma)}{T}\}$$

In either case, a numerical method of maximization is necessary to maximize such a support function since you cannot even differentiate your recipe. But you can graph the support function, in three dimensions in the two-parameter case, and in only two dimensions in the one parameter case, and do a grid search to find its maximum. The results should be identical.

Example 7 Continued. Judge, et al. (1988, pp. 405-409) consider an artificially constructed example of a regression with two independent variables and a first-order serially correlated disturbance (the true values of ρ and the other parameters are not reported):

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + u,\ \text{where } x_1 = (1,1,1,\ldots,1)' \text{ and } u_t = \rho\, u_{t-1} + \varepsilon_t,\ \varepsilon_t \overset{iid}{\sim} n(0,\sigma^2).$$

There are only 20 observations, which is relatively short for time-series analysis. The unnormalized support function, concentrated with respect to β and σ^2 is plotted in Figure 5A and 5B. In the third and fourth panels, Figure 5C and 5D, I plot the three dimensional sliced likelihood for $\hat\rho$ and $\hat\sigma^2$, holding the estimated β's at their support maximizing levels.

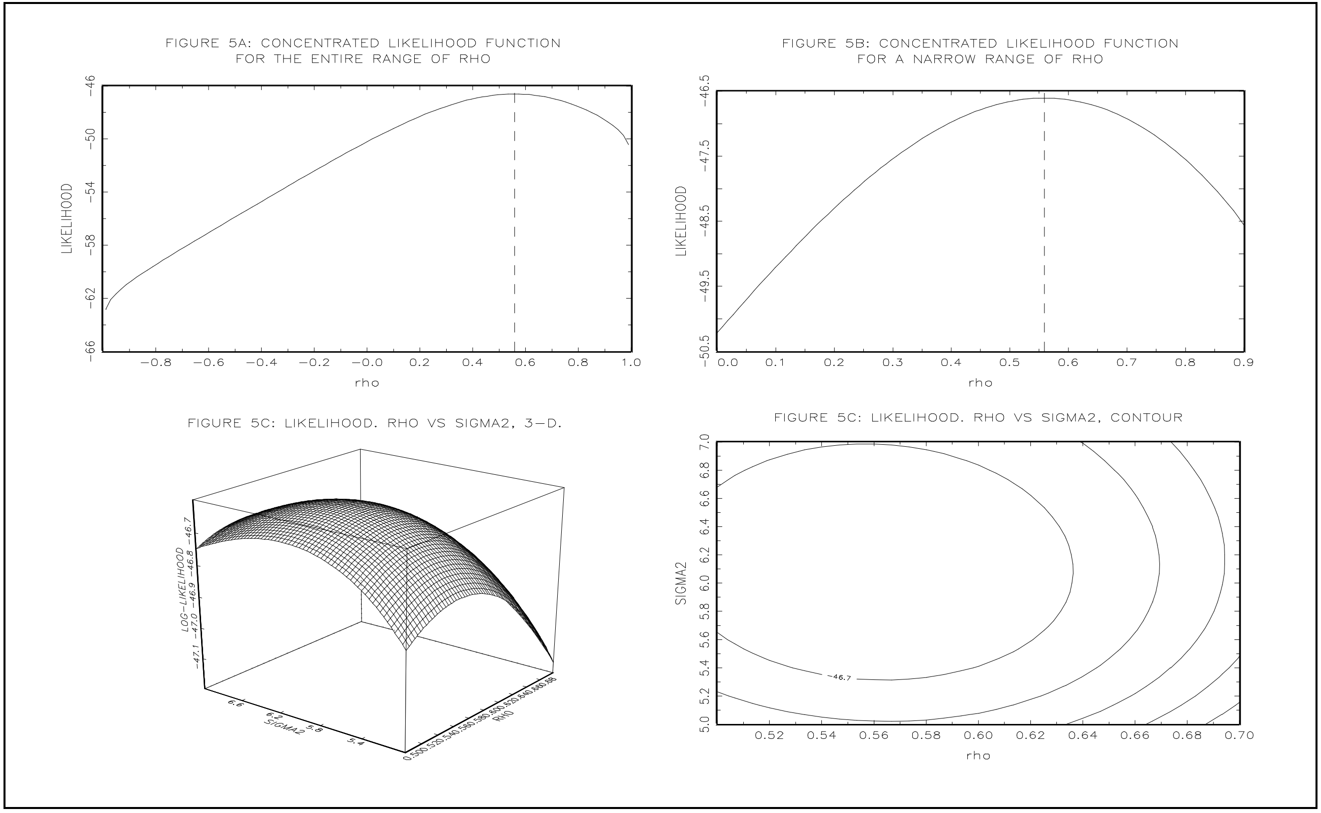

Figure 5: Concentrated and Sliced Support Function for a Regression with First-Order Autoregressive Disturbances, y_t =x_1b$_1$+x_2b$_2$+x_3b$_3$+u_t, u_t=ru_{t-1}+e$_t$, e$_t$ iid n(0, s^2)

The value of ρ which maximizes the concentrated support is 0.55887796; the value of the concentrated support at that value is -46.605855, with gradient = -5.09×10^{-5}, and Hessian = -31.060.[27]

Example 8 Continued. Judge, et al. (1988, pp. 374-378) consider an artificially constructed example of a regression with two independent variables and a heteroskedastic disturbance:

$$y = x_1 b_1 + x_2 b_2 + x_3 b_3 + u, \text{ where } x_1 = (1,1,1,...,1)' \text{ and } u_t \overset{iid}{\sim} n(0, s_t^2), \text{ where } s_t^2 = \exp\{g_1 + g_2 x_{2t}\}.$$

There are 20 observations, which have been artificially generated with β = (10, 1, 1)' and (γ$_1$, γ$_2$)' = (-3, 0.03)'. The concentrated support of the reparametrized model, with full support given by (13), with respect to the remaining parameter γ, is plotted in Figure 6A and B. In the third and fourth panels, Figure 6C and 6D, I plot the three dimensional sliced likelihood for $\bar{a}$ and $\bar{s}^2$, holding the estimated β's at their support maximizing levels.

[27] These results were obtained by the method of steepest ascent and numerical calculation of the gradient and Hessian as described in Chapter 5. The square root of the reciprocal of the Hessian is 0.179 and is not a good measure, even asymptotically, of the standard error of the estimates because it does not take account of possible variability in the estimates of the other parameters. To obtain the asymptotic standard errors (SEs) described in Chapter 6, the Hessian of the *full* support function at the maximum with respect to all the parameters must be obtained . The values of these parameters are est. (β$_1$, β$_2$, β$_3$, ρ, σ^2) = (4.055, 1.67, 0.7618, 0.559 6.109), with correct asymptotic standard errors = (7.685, 0.348, 0.146 , 0.180, 1.948).

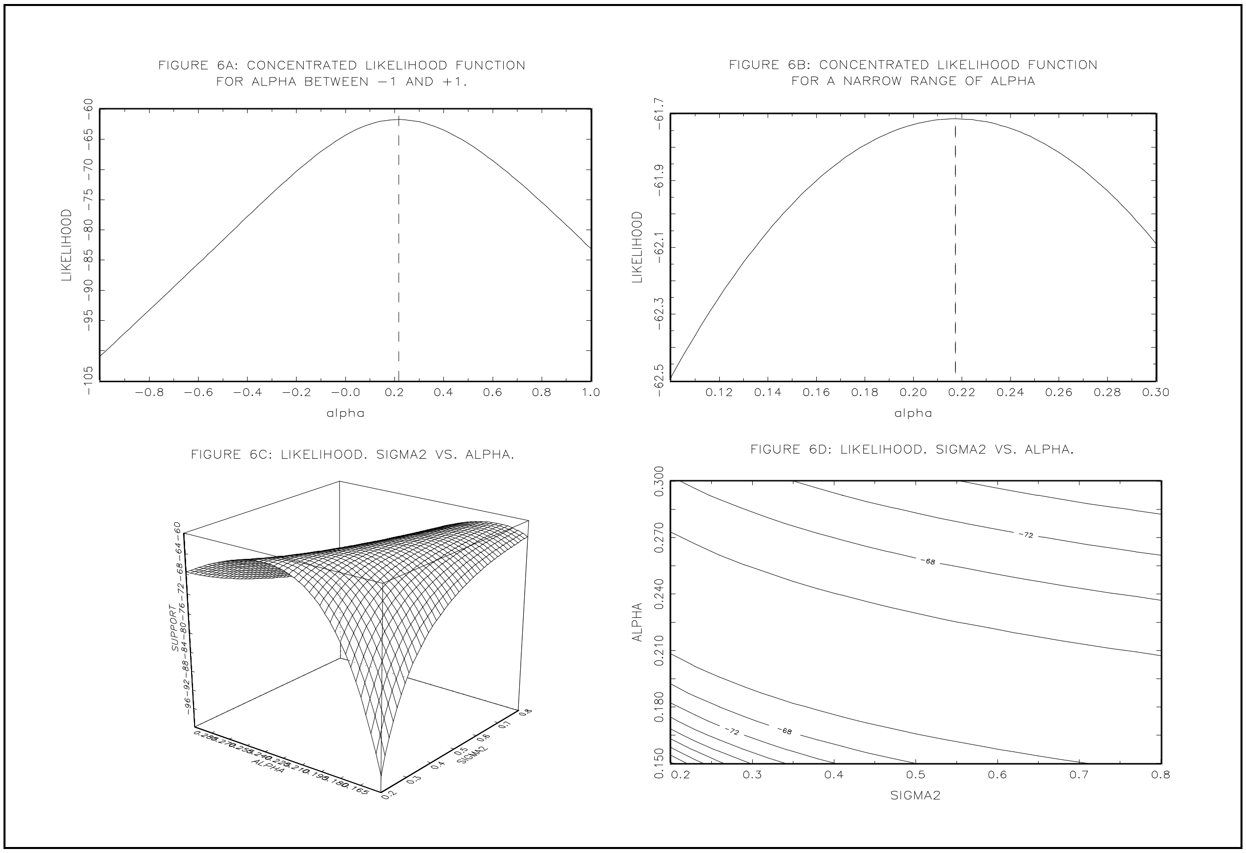

Figure 6: Heteroskedastistic Disturbances, $y_t = x_t'\beta + u_t$, u_t ~ n(0, s_t^2),
$s_t^2 = \exp(g_0 + q_1 z_t)$, **Concentrated and Sliced Support Functions**

It is apparent from panels 6C and 6D that the overall variance $s^2 = e^{g_0}$ is not well-determined. Although the maximizing value is not hard to find by first concentrating the likelihood function, many values of σ^2 yield very nearly the same likelihood for the data at hand. I consider the problem of heteroskedasticity in Chapter 9 in more detail; we will find that this indeterminacy of the overall variance is a general characteristic of econometric problems involving relationships with heteroskedastic disturbances. Nonetheless, the parameter α is well-determined; setting it equal to the value which maximizes the concentrated likelihood function converts the general problem into an ordinary regression problem. The value of α obtained by maximizing the concentrated likelihood function by the method of steepest ascent is 0.217 and at this value the hessian of the CLF is -111.832.[28]

The examples of this section serve to illustrate the power of the techniques of concentrating and slicing to reveal the econometric significance of the data for models involving more than one parameter. Moreover, concentration is a powerful tool for finding the maximum of the likelihood or support function, even under somewhat problematic circumstances. In the next and final section of this Chapter, I explore

[28] See the preceding footnote. The square root of the negative of the reciprocal of the hessian at $\hat{a}$ is 0.0089, which, however, for the reason stated above, is not a good measure of the uncertainty attached to the estimate of α. The full maximum likelihood estimates of the parameters (β_1, β_2, β_3, β_1, σ^2, α)' are (0.910, 1.603, 0.951, 0.303, 0.217) with conventionally estimated asymptotic SEs (6.95, 0.387, 0.342, 0.604, 0.0946)'. These clearly reveal the imprecision attached to the estimate of σ^2 and the relative precision with which the remaining parameters, except for the constant term, are estimated.

three additional examples of econometric significance: nonlinear regression (treated at length in Chapter 8); estimation of a two-parameter Box-Cox transformation (also discussed further in Chapter 8); and, finally, a classic problem in spatial econometrics due to Anselin (1988, discussed in greater depth in Chapter 11).

8. Further Examples of Likelihood inference in Econometrics

Example 9: Nonlinear Regression. This example, considered in Davidson and MacKinnon (1993, pp. 745-747), concerns the simplest form of nonlinear regression   I consider the likelihood of the parameters in, and the maximum-likelihood (identical to nonlinear least-squares) estimates of the coefficients in,

$$y = a_0 + a_1 x^b + e, \quad e \text{ iid } n(0, s^2) \,.$$

In this case, I generated the data: T = 20 observations, with $a_0 = 2.5$, $a_1 = 0.25$, b = 2.5, $\sigma^2 = 100$. The x series was also generated by $x_t = 0.5t + u_t$, u_t iid n(0,10), t = 1,...,T. The resulting series was then conditioned on in forming the log likelihood or support function:

(14)
$$S(a_0, a_1, b, s^2 | y, x) = -\frac{T}{2}\log 2p - \frac{T}{2}\log s^2 - \frac{1}{2s^2}\sum_{1}^{T}(y_t - a_0 - a_1 x_t{}^b)^2 \,,$$

where x = (3.5104, 1.7278, 3.1601, 4.0099,3.4259, 4.1231, 4.7934, 4.7501, 5.0277, 7.2197, 7.5301, 7.6453, 7.4881, 7.7315, 9.6487, 10.2315, 10.8605, 11.0478, 11.2747, 10.1609)' and
y = (32.2262, -3.2802, 17.2982, 1.4356, 8.6228, 17.2995, 10.5950, 15.9421, 20.0281, 37.2247, 30.8677,48.8092,56.2994,31.7057,86.5378,95.0347,95.1128,115.2190, 121.8168, 103.2862)'.

These data are graphed in Figure 7.

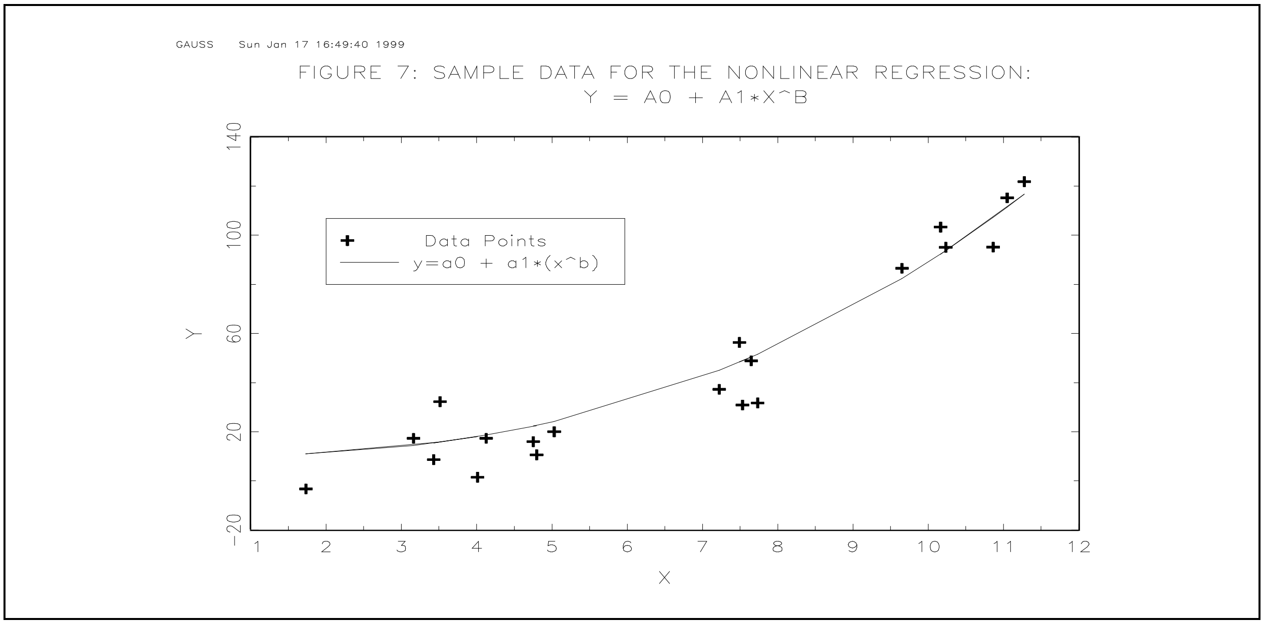

Figure 7: A Nonlinear Regression and a Sample of Data Points from It

Even without the benefit of the "true" regression line plotted in the figure, I think anyone looking at the data points would suspect a nonlinear relationship, although the exact form of such a relationship would be somewhat problematic. I will assume that the "correct" form is known.

The likelihood function can easily be concentrated in the parameter b so that the problem can be reduced to maximizing the concentrated likelihood function with respect to the single parameter b and then recovering estimates of a_0, a_1, and σ^2 by regressing y on a vector of ones and $x^{\hat{b}}$, where $\hat{b}$ is that value which maximizes the concentrated likelihood function. The concentrated likelihood function is graphed in Figure 8; it has a beautiful shape, but, as remarked above, 2-D appearances may be deceptive.

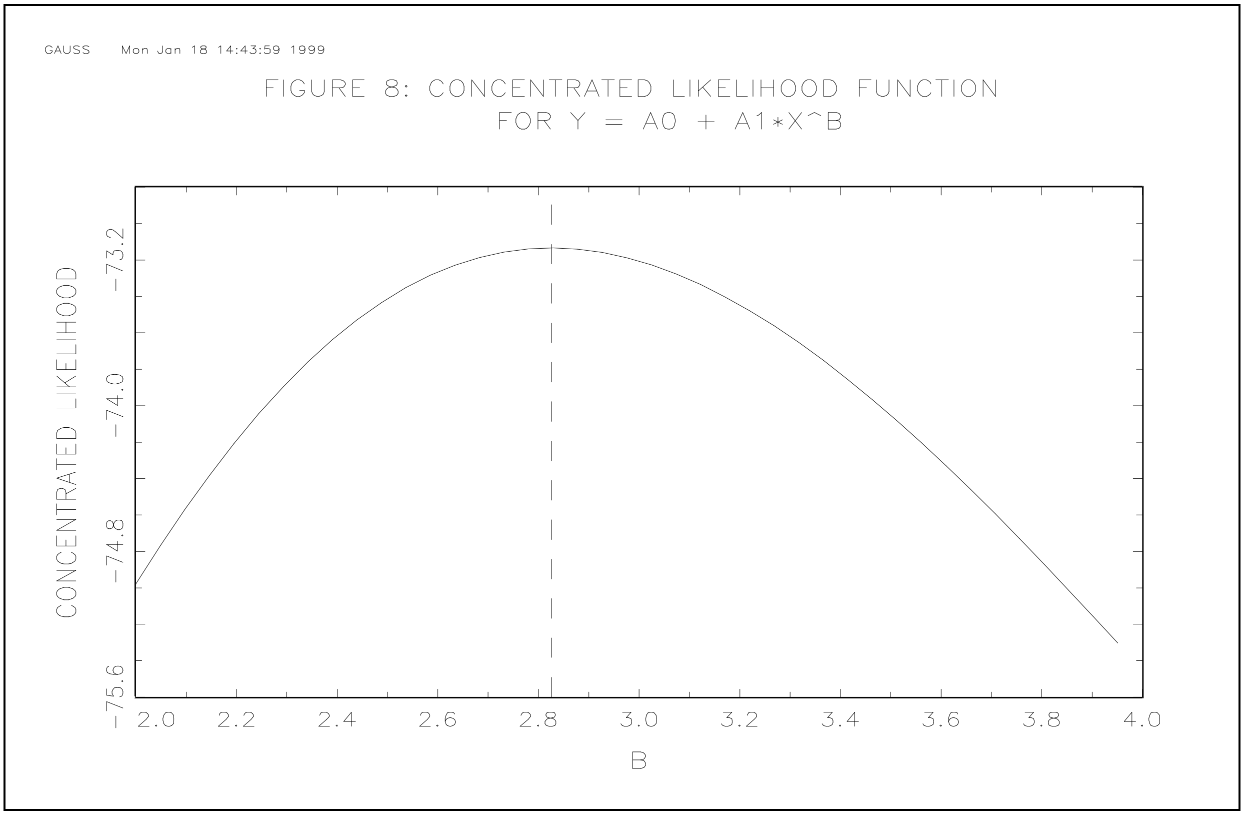

Figure 8: Concentrated Support function for b in $y = a_0 + a_1x^b + u$, $u \sim N(0, s^2I)$

Applying the method of steepest ascent, $\hat{b}$ = 2.8261 with likelihood -73.1343 and hessian = -4.9347. The remaining parameters are obtained by OLS of y on $x^{\hat{b}}$ and the asymptotic standard errors from the hessian of the full likelihood function evaluated at the maximizing values: est (a_0 , a_1, b, σ^2) = (6.811, 0.121, 2.826, 92.466) with SE = (5.148, 0.137, 0.465, 30.823). It appears from the asymptotic results that only a_1 is difficult to estimate accurately. 3-D slices of the full likelihood function in the direction of b vs. a_1 are plotted in Figures 9A and 9B and of b vs. σ^2.

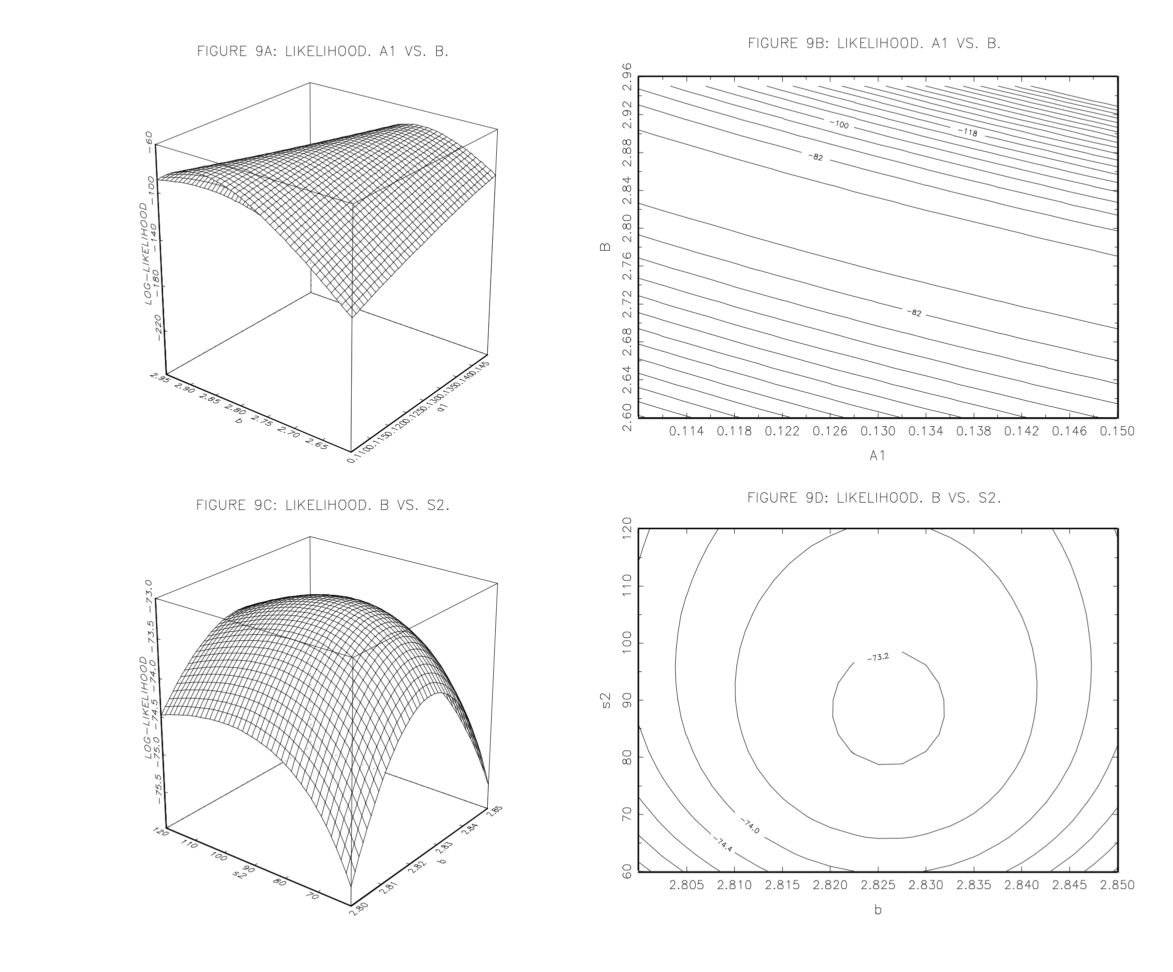

Figure 9: 3-D Likelihood Slices, b vs. a_1 and b vs. s^2, for the Nonlinear Regression Model, $y = a_0 + a_1x^b$

While it is apparent that b and σ^2 are jointly estimated with relative precision, even small variations in b are reflected in the inference one can draw concerning a_1.

Example 10: Regression with Box-Cox Transformations. This transformation was introduced by G.E.P. Box and D.R. Cox (1964). For a strictly positive variable x, the transform is

$$(15) \qquad x^{\lambda} \quad = (x^{\lambda} - 1)/\lambda, \qquad \lambda \neq 0$$
$$\qquad\qquad\quad = \log x, \qquad \lambda = 0.$$

It includes linearity, $\lambda = 1$, and log-linearity, $\lambda = 0$, as special cases. The case $\lambda = -1$ is the reciprocal transformation. In its most general form, a Box-Cox transformed regression is

$$(16) \qquad y_t^{(\lambda_0)} = \alpha + \beta_1 x_{1t}^{(\lambda_1)} + \beta_2 x_{2t}^{(\lambda_2)} + \ldots + \beta_k x_{k_t}^{(\lambda_k)} + \varepsilon_t .$$

But a form so general is rarely estimated. Rather λ is generally restricted to be either 1 or the same for every variable to which it is applied. The three leading cases are thus:

Case 1: $y = \alpha + \beta_1 x_{1t}^{(\lambda)} + \beta_2 x_{2t}^{(\lambda)} + \ldots + \beta_k x_{kt}^{(\lambda)} + \varepsilon_t .$

Case 2: $y^{(\lambda)} = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_k x_{kt} + \varepsilon_t.$

Case 3: $y^{(\lambda)} = \alpha + \beta_1 x_{1t}^{(\lambda)} + \beta_2 x_{2t}^{(\lambda)} + . . . + \beta_k x_{kt}^{(\lambda)} + \varepsilon_t$.

In many situations, however, we would want to transform the dependent variable differently from at least some of the independent variables. This can be done relatively easily provided there are only *two* different values for λ in addition to 1.0. There is a fundamental difference, as we shall see, between case 1, in which only the independent variables are transformed, and the others. It suffices to treat the following case to show the difference and to discuss estimation and inference issues:

(17) $$y_t^{(\theta)} = \alpha + \beta x_t^{(\lambda)} + \varepsilon_t ,$$

where $\theta \neq \lambda$, y_t, $x_t > 0$ all t and where $\varepsilon_t \overset{\text{i.i.d.}}{\sim} n(0, \sigma^2)$. Note the inconsistency between the assumption $y_t > 0$ all t and $\varepsilon_t \sim n(0, \sigma^2)$; this can cause convergence problems for several estimation methods.

Under the assumption that $\varepsilon_t \overset{\text{i.i.d.}}{\sim} n(0, \sigma^2)$ for the true value of λ and θ, the density for a sample of size T, $(\varepsilon_1, ..., \varepsilon_T)'$, is

(18) $$\varepsilon \sim \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left\{\frac{-1}{2\sigma^2} \sum_{t=1}^{T} \varepsilon_t^2\right\} .$$

Now the Jacobian of the transform from ε_t to y_t, given x_t is

(19) $$J(\varepsilon_t \to y_t) = |y^{\theta-1}| = y^{\theta-1},$$

if y > 0. Thus the joint density of $y = (y_1, ..., y_T)'$ conditional on
$x = (x_1, ..., x_T)'$ is

(20) $$f(y|x) = \left(\frac{1}{2\pi\sigma^2}\right)^{T/2} \exp\left\{\frac{-1}{2\sigma^2} \sum_{t=1}^{T} \left(y_t^{(\theta)} - \alpha - \beta x_t^{(\lambda)}\right)^2\right\} \prod_{t=1}^{T} y_t^{\theta-1} .$$

It follows that the support or log likelihood function is

(21) $$S(\alpha, \beta, \sigma^2, \theta, \lambda|y, x) = k - \frac{T}{2} \log \sigma^2 + (\theta - 1) \sum_{t=1}^{T} \log y_t - \frac{1}{2\sigma^2} \sum_{t=1}^{T} \left(y_t^{(\theta)} - \alpha - \beta x_t^{(\lambda)}\right)^2 .$$

Concentrate this likelihood function with respect to σ^2:

(22) $$\hat{\sigma}^2 = RSS(\alpha, \beta, \theta, \lambda|y, x) = \frac{1}{T} \sum_{t=1}^{T} (y^{(\theta)} - \alpha - \beta x_t^{(\lambda)})^2 ,$$

hence

(23) $$S^*(\alpha, \beta, \lambda, \theta|y, x) = k^* - \frac{T}{2} \log RSS(\alpha, \beta, \lambda, \theta) + (\theta - 1)T\ \overline{\log y} ,$$

where $\overline{\log y} = \frac{1}{T} \sum_{t=1}^{T} \log y_t$ is the geometric mean of the dependent variable.

If θ = 1, ML estimation of α, β, and λ is equivalent to minimizing the sum of squares

$$RSS(\alpha, \beta, \lambda|y, x) = \frac{1}{T} \sum_{t=1}^{T} \left(y_t - \alpha - \beta x_t^{(\lambda)}\right)^2 .$$

For a fixed value of λ this is equivalent to estimating α and β from the OLS regression of y_t on $x_t^{(\lambda)}$. Hence to find $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\lambda}$, we can iterate or do a grid search on λ using OLS to concentrate S with respect to α and β.

But for q ¹ 1, the likelihood function involves another term,

$$(\theta - 1)T\ \overline{\ln y}\ ,$$

so that ML is *not* equivalent to

$$\min_{\alpha,\beta,\theta,\lambda} S(\alpha, \beta, \theta, \lambda|y, x).$$

Moreover, $y_t^{(\theta)}$ may not exist for some possible values of ε_t, depending on the value of θ.

If λ = θ is assumed, a grid-search on λ, usually in the interval [-2, 2] is easy to implement. The criterion function in this case is

(24) $$(\lambda - 1)\ T\ \overline{\log y} - \frac{T}{2} \log S(\alpha, \beta, \lambda) = C(\lambda).$$

For given λ, regress

$$\frac{y_t^{\lambda} - 1}{\lambda} \text{ on } \frac{x_t^{\lambda} - 1}{\lambda}$$

to obtain $\hat{\alpha}_{(\lambda)}$ and $\hat{\beta}(\lambda)$. Calculate

$$C(\lambda) = (\lambda - 1)T\ \overline{\log y} - \frac{T}{2} \log S(\hat{\alpha}(\lambda), \hat{\beta}(\lambda), \lambda).$$

Change λ in such a way as to locate the maximum C.

In this example, I consider finding the maximum of the 2-parameter likelihood function for the example of estimation of a relationship involving the Box-Cox transformation The data, 20 observations, were generated from the following model:

$$y^{(q)} = b_0 + b_1 x^{(l)} + e,$$

$$\text{where } e \underset{\sim}{\overset{iid}{}} n(0,10.0) \text{ and } y^{(q)} = \frac{y^q - 1}{q},\ x^{(l)} = \frac{x^{l_i} - 1}{l},$$

$$q = 0.25,\ l = \text{-}0.50, b_0 = 100, b_1 = -10.$$

Note that the constant term is not transformed. The series used for x was the same as that in the previous example. The two series, y and x, were:

x = (3.5104, 1.7278, 3.1601, 4.0099, 3.4259, 4.1231, 4.7934, 4.7501, 5.0277, 7.2197,
7.5301, 7.6453, 7.4881, 7.7315, 9.6487, 10.2315, 10.8605, 11.0478, 11.2747, 10.1609)',
y = (39.2579, 24.0685, 31.1748, 29.5964, 21.7834, 24.2571, 20.9925, 31.6998, 22.4851, 22.7277,
17.5754, 19.8328, 14.9361, 19.8355, 17.2624, 19.2217, 18.1855, 20.7217, 15.0625, 19.3294)×10^4.

The data are graphed against the true relationship in Figure 10.

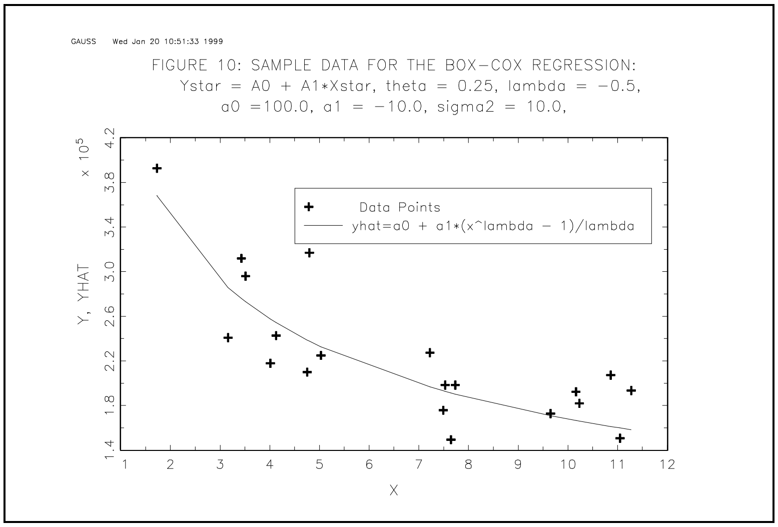

Figure 10: Sample Data for the Box-Cox Example and the Relationship from which They Were Generated

Notwithstanding the apparently reasonably appearance of these data, they are nonetheless associated with a likelihood function which is far from easy to analyze.

A 3-D plot of the concentrated likelihood function is given in Figure 11.

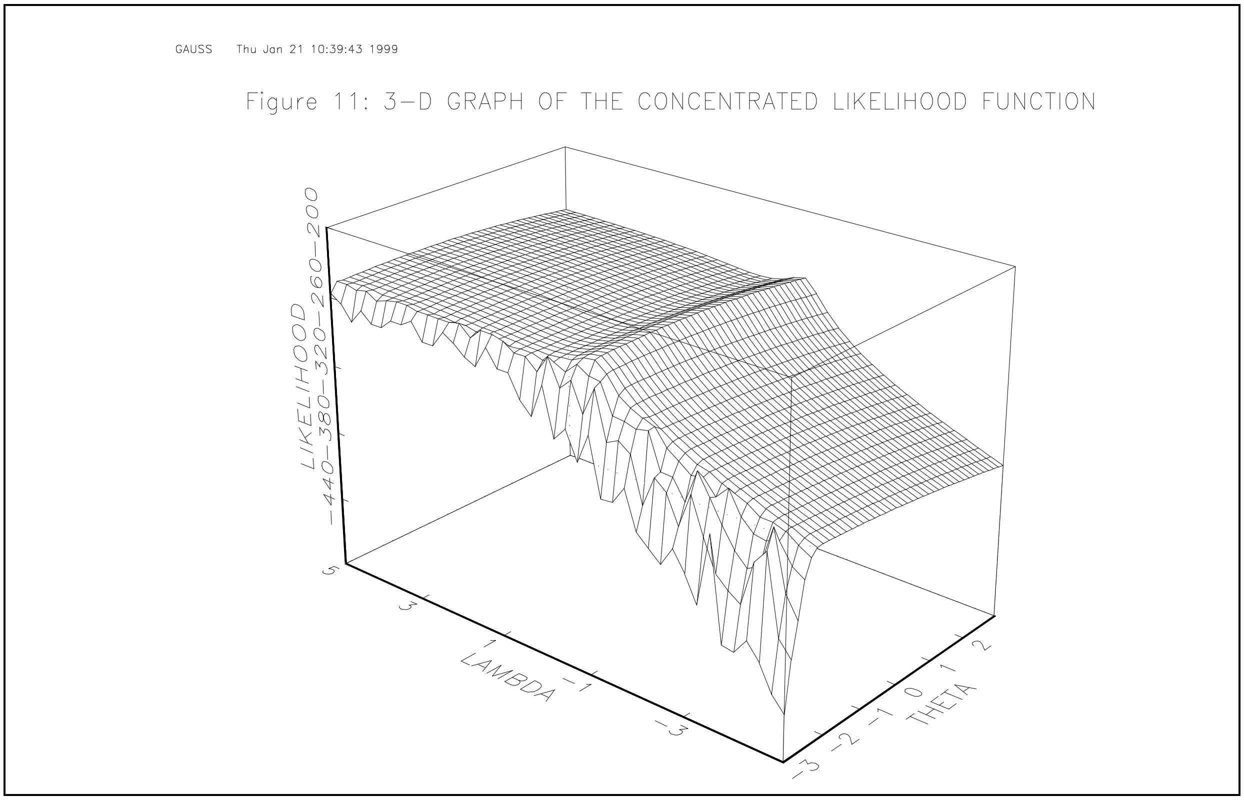

Figure 12: Support Function for the Two-Parameter Box-Cox Model, Concentrated in l and q

It is apparent that this likelihood function will be difficult to maximize: Around -1.5 there appears to be a deep crevice in the θ direction, but, apart from than anomaly, the LF is nearly level in that direction. The situation is even more peculiar in the λ direction: There is "welt" near zero. Beware of "welts" or crevices when graphing. Because plotting at a finite number of points may smooth and obscure, singularities may be smoothed over. To get a better idea of this elephant, I sliced through the concentrated likelihood holding first λ fixed at 0 , then θ at 1. The slices are plotted in Figure 12, for a wide range of θ and λ in panels 12A and 12B and a narrow range in panels 12C and 12D. I have also plotted a dashed line at the point in each slice where visual inspection suggests the maximum to be. (Of course, this won't be the true maximum because I am not allowing for simultaneous variation of the two parameters.) These points are $\theta = -1.5$ and $\lambda = -0.7$, respectively. Starting at these values, I then used the method of steepest ascent to obtain the values $\theta = -1.121$ and $\lambda = -0.692$. I will discuss the method of steepest ascent and other methods of maximization, including grid search, for problems involving two or more parameters at length in Chapter 5.

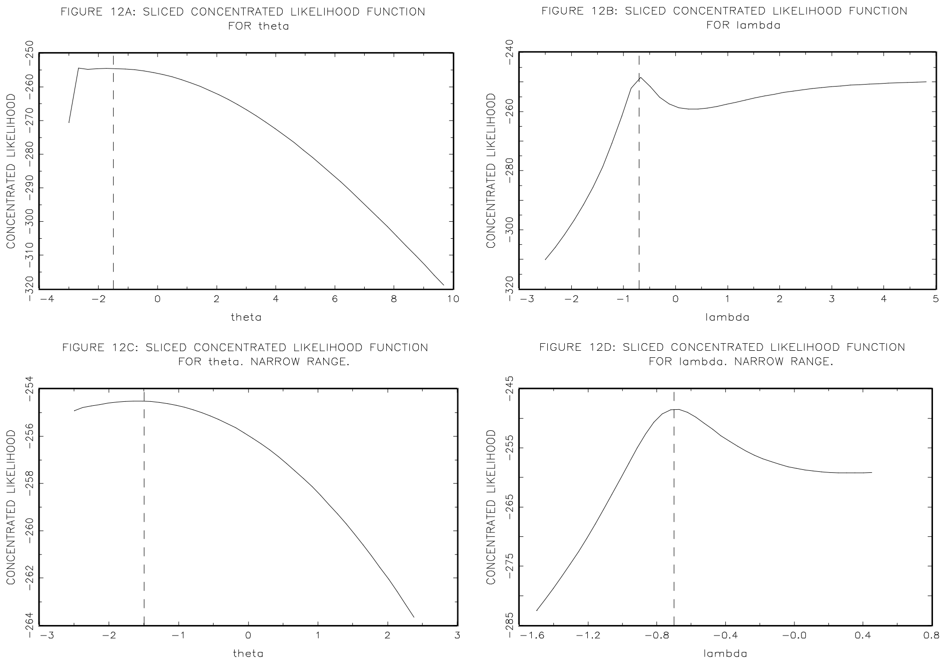

Figure 12: Slices through the Concentrated Likelihood Function in the Direction of q **and of** l

In this case, we have the good fortune to know the true values of the parameters which generated the data: $\theta = 0.25$ and $\lambda = -0.5$. While -0.69 is not far off from -0.5, $\theta = -1.12$ is very wide of the true value 0.25. On the other hand, until one gets close to the crevice, the concentrated likelihood does not vary much along the λ direction, so θ is not well-determined. However, the slope and intercept β_1 and β_0 are extremely sensitive to the transformation of the dependent variable in a Box-Cox regression, and *a foriori* the value of σ^2 is even more so; for the values of θ and λ which appear to maximize the concentrated likelihood these values are: $\beta_0 = 0.89$, $\beta_1 = -1.10\times10^{-6}$ and $\sigma^2 = 2.25\times10^{-14}$! The hessian at the supposed maximum of the concentrated support function is $\begin{pmatrix} 5.812 & 0.00199 \\ 0.00199 & -216.240 \end{pmatrix}$, which is not negative definite.[29] Under the circumstances, I do not bother to present the hessian for the full support function at $\theta = -1.121$,

[29] There is an approximation to the hessian in terms of the negative of the sum over observations of the outer product of the gradient vector evaluated at each individual observation (see Davidson and

λ = -0.692, β_0 = 0.89, β_1 = -1.10×10^{-6} and σ^2 = 2.25×10^{-14}. It is not negative definite either. The moral of this tale is that curvature is at best a highly problematic property to estimate, especially with only 20 observations. I will further discuss the Box-Cox transformation and related estimation problems in Chapter 8.

Example 11: Regression with Spatially Dependent Data.[30] Spatially dependent data are explored at length in Chapter 11; here I consider only a very simple example. Following the original suggestion by Whittle (1954) for the treatment of spatially dependent data, Cliff and Ord (1969) elaborated what has come to be called the spatial autoregressive model (see also Fisher, 1971, and Kelejian and Robinson, 1995). The most important exposition of the general problem is Anselin (1988). In a famous example, Anselin (1988, pp. 187-198) relates crime incidence in contiguous Columbus, Ohio, neighborhoods linearly to income and housing values. Consider the following regression in which spatial dependence among the disturbances is modeled by the so-called weight matrix W:

(25)
$$\begin{aligned} y &= Xb + e \\ e &= lWe + u, \quad u \sim N(0, s^2 I). \end{aligned}$$

y, ε, and u are N×1 vectors, β is k×1, X is N×k, and λ is a scalar, analogous to an autoregressive coefficient, which expresses the dependence among the N observations. W, the so-called "weight" matrix, is N×N, expresses which disturbances are related to one another, and how. Here, I assume that the matrix W consists of zeros and ones and is not necessarily of full rank.[31] The standard approach to nonspherical disturbances is to transform the data to a form in which the disturbances are spherical (Fisher, 1971, p.21). If λ were known, the appropriate transformation would be:

(26)
$$\begin{aligned} y^* &= (I - lW)y = By \\ &= BXb + Be = X^*b + u \end{aligned}.$$

The Jacobian of this transformation is $J = |BB'|^{\frac{1}{2}}$. Consequently, the joint density of y, given X, W, β, σ^2, and λ, is

$$\left(\frac{1}{2ps^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2s^2}(y^* - X^*b)'(y^* - X^*b)} |BB'|^{\frac{1}{2}}.$$

It follows that the log likelihood or support is

(27)
$$\begin{aligned} &S(b, s^2, l|y, X, W) \\ &\quad = -\frac{N}{2}\left\{\ln 2p + \ln s^2 + \frac{(y^* - X^*b)'(y^* - X^*b)/N}{2s^2}\right\} + \frac{1}{2}\ln BB' \end{aligned}.$$

MacKinnon, 1993, pp.265-267). This is $\begin{bmatrix} -51.131 & -788.089 \\ -788.089 & -12245.748 \end{bmatrix}$, which is negative definite, but not likely in present circumstances to be a very good approximation. I discuss this approximation in detail in Chapter 6.

[30] I am indebted to Mark Fleming for introducing me to the subject of econometric estimation with spatially dependent disturbances, calling my attention to key references to the literature, and finding an error in some of my computations.

[31] In a more general context these weights might express the nature of the relation between spatial data points, for example, inversely related to distance (Anselin, 1988, p. 28).

Holding λ fixed and maximizing the support with respect to β and σ^2, we obtain

$$\hat{\sigma}^2(\lambda) = \frac{(y^*-X^*b(\lambda))'(y^*-X^*b(\lambda))}{N}, \text{ where}$$

$$b(\lambda) = (X^{*\prime}X^*)^{-1}(X^{*\prime}y^*)$$

Thus the concentrated support as a function of the spatial dependence parameter λ is

(28) $$S^*(\lambda) = -\frac{N}{2}\left[1+\ln 2\pi+\ln\hat{\sigma}^2(\lambda)\right]+\frac{1}{2}\ln|BB'|.$$

The final term assumes a special significance. Clearly when λ is equal to the reciprocal of any characteristic value of W, |BB'| = 0 so that this term tends to minus infinity.[32] If W is nonsingular, there will be N distinct characteristic values; between each pair, S*(λ) has a local maximum. This does not, however affect the existence of a global maximum. The concentrated log likelihood is graphed in Figure 13A for a wide range of λ-values and in Figure 13B for a narrow range.

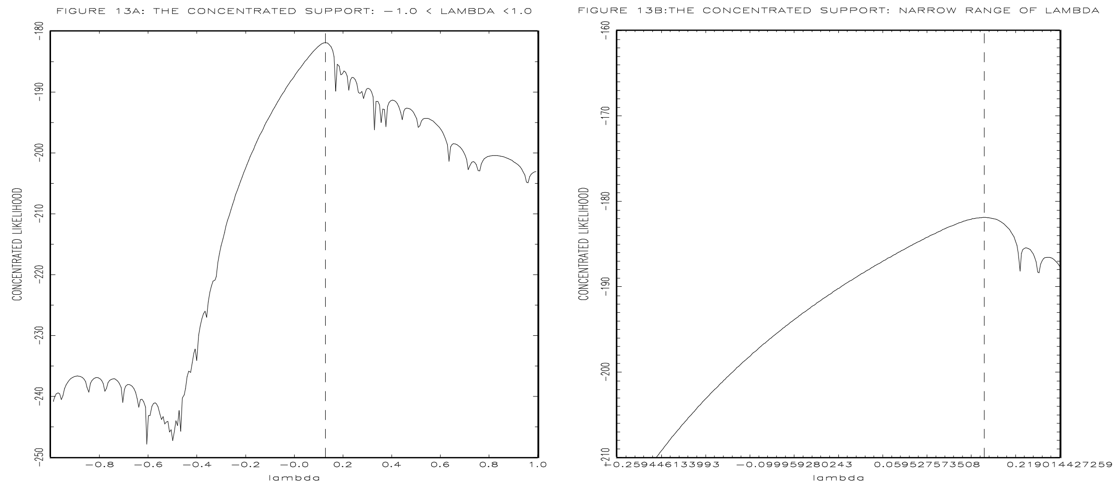

Figure 13: Concentrated Support for the Anselin Data, Spatial Dependence Data

The jagged appearance is the result of the singularities present for values of λ equal the reciprocals of the characteristic values of W; graphing the concentrated support at a finite number of points smoothes these singularities into sharp dips. But clearly the global maximum is well-defined and S* is smooth, convex and well-approximated by a quadratic in the vicinity of the global maximum, which I find by the method of steepest ascent to be at λ = 0.12838043. At these values, the remaining parameters for the Anselin data are: b_0 = 55.325,b_1 = -0.935, b_2 = -0.301, and $\hat{\sigma}^2(\lambda)$ = 87.896, with asymptotic standard errors derived from the hessian of the full support function (27) evaluated at the maximizing values of SE = λ: 0.0222, b_0: 5 .8740, b_1: 0.3298, b_2: 0.0882, σ^2: 18.3044, respectively.[33]

[32] This fact has been noted, *inter alia*, by Kelejian and Robinson (1995) who conclude, erroneously in my opinion, that maximum likelihood is defective.

[33] By contrast, Anselin (1988, p.195), using a method which normalizes the row sums to 1, finds: λ = 0.562, b_0 = 59.893,b_1 = -0.941, b_2 = -0.302, and $\hat{\sigma}^2(\lambda)$ = 95.575, with asymptotic standard errors of SE = λ:

9. Suggestions for Further Reading

Of course, one should begin with Fisher's profound and deep treatment (1956), *Statistical Methods and Scientific Inference.* Also very accessible, albeit somewhat idiosyncratic, is A. W. F. Edwards (1972, 1992) *Likelihood*. Recent treatments of the subject at roughly the same level include Azzalini (1996), *Statistical Inference Based on the Likelihood* and Royall (1997), *Statistical Evidence: A Likelihood Paradigm.* Lindsey (1996), *Parametric Statistical Inference* and Tanner (1996), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, are somewhat more advanced. A more comprehensive treatment of the likelihood principle than Birnbaum (1962) is Berger and Wolpert (1984),*The Likelihood Principle.*

For focus on a variety of econometric application of maximum likelihood, nothing beats J. S. Cramer's rather neglected *Econometric Applications of Maximum Likelihood* (1986).

On history -- and one should not neglect history, Aldrich's (1997) paper, "R. A. Fisher and the Making of Maximum Likelihood, 1912-1922," is both careful and thoughtful. Jimmie Savage's posthumously published Fisher Lecture (1976), "On Rereading R. A. Fisher," is difficult but worth the effort. Finally, Steve Stigler's *The History of Statistics: The Measurement of Uncertainty before 1900,* is both delightful and instructive. Anders Hald's recent (1998), *A History of Mathematical Statistics from 1750 to 1930,* is a bit ponderous but frightenly complete and as accurate and detailed as they come.

REFERENCES FOR CHAPTER 1

Abbott, E. A., *Flatland.* New York: Dover, 1884, reprinted 1946.

Aldrich, J., "R. A. Fisher and the Making of Maximum Likelihood, 1912-1922," *Statistical Science, 12:* 162-176, 1997.

Anselin, L., *Spatial Econometrics: Methods and Models,* Dordrecht: Kluwer Academic, 1988.

Azzalini, A., *Statistical Inference Based on the Likelihood,* London: Chapman & Hall, 1996.

Barndorff-Nielsen, O., *Information and Exponential Families,* New York: Wiley, 1978.

Barnard, G. A., "Statistical Inference," *Jour. Royal Statistical Society, Ser. B 11:* 115-149 (1949).

Barnard, G. A., "The Theory of Information," *Jour. Royal Statistical Society, Ser. B 13:* 46-64 (1951).

Barnard, G. A., "The Use of the Likelihood Function in Statistical Practice," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1:* 27-40 (1966).

Barnard, G. A., "The Bayesian Controversy in Statistical Inference," *Journal of the Institute of Actuaries, 93:* 229-69 (1967).

Barnard, G. A., G. M. Jenkins and C. B. Winsten "Likelihood Inference and Time Series," *Jour. Royal Statistical Society, Ser. A 125:*321-72 (1962).

0.134, b_0: 5 .366, b_1: 0.331, b_2: 0.090, σ^2: unreported. The main difference, as one might expect, is in the value of the spatial dependency parameter λ. Row normalization changes the interpretation of this parameter, not significant here, but potentially problematic if the dependencies are variable, for example depending on distance. I return to this point in Chapter 11.

Barndorff-Nielsen, O. E., *Parametric Statistical Models and Likelihood,* Lecture Notes in Statistics, No. 50. Berlin: Springer-Verlag, 1988.

Bayes, T., "An Essay towards Solving a Problem in the Doctrine of Chances," *Philosophical Transactions of the Royal society of London for 1763, 53:* 370-414, 1764. Reprinted with an introduction by G. A. Barnard, pp. 131-153, in E. S. Pearson and M. G. Kendall, eds., *Studies in the History of Statistics and Probability,* London: Chas. Griffin, 1970.

Bennett, J. H., ed., *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher,* New York: Oxford University Press, 1990.

Berger, J. O., and R. L. Wolpert, *The Likelihood Principle,* Hayward, CA: Institute of Mathematical Statistics, 1984.

Bernnoulli, Jacob, *Ars Conjectandi*, Basel: Thurnisiorum, 1713.

Bernoulli, Daniel, "The Most Probable Choice between Several Discrepant Observations and the Formation therefrom of the Most Likely Induction," in Latin in *Acta Academiae Scientiarum Imperialis Petropolitanae, 1777.* Reprinted with an introduction by M. G. Kendall and an extended commentary by Leonard Euler, Observations on the foregoing Dissertation of Bernoulli, pp. 155-172 in E. S. Pearson and M. G. Kendall, eds., *Studies in the History of Statistics and Probability,* London: Chas. Griffin, 1970.

Billingsley, P., *Probability and Measure, 3rd ed.,* New York: Wiley, 1995.

Birnbaum, A., "On the Foundations of Statistical Inference," *Jour. American Statistical Association, 57:*269-306 (1962). Reprinted in S. Kotz and N. L. Johnson, *Breakthroughs in Statistics, Vol.1,* with an introduction by J. F. Bjørnstad, pp. 461- 518, New York: Springer-Verlag, 1992.

Box, G.E.P., and D. R. Cox, "An Analysis of Transformations," *Jour. Royal Statistical Society, Ser. B, 26*: 211-243 (1964).

Cliff, A. D., and J. K. Ord, "The Problem of Spatial Auto Correlation, pp. 25-55 in A. J. Scott, ed., *Studies in Regional Science,* London: Pion, 1969.

Cox, D. R., "Some Problems Connected with Statistical Inference," *Annals of Mathematical Statistics, 29:* 357-372, 1958.

Cramer, J. S., *Econometric Applications of Maximum Likelihood Methods.* Cambridge: University Press,1986.

Dale, A. L., *A History of Inverse Probability from Thomas Bayes to Karl Pearson,* New York: Springer Verlag, 1991.

Davidson, R. and J. G. MacKinnon, *Estimation and Inference in Econometrics,* New York: Oxford University Press, 1993.

Daw, R. H., and E. S. Pearson, "Abraham de Moivre's 1733 Derivation of the Normal Curve: A Bibliographic Note," *Biometrika, 59:* 677-680, 1972.

de Morgan, A., *An Essay on Probabilities and Their Application to Life Contingencies and Insurance Offices,* London: Longmans, 1838.

de Moivre, Abraham, *The Doctrine of Chances, 2nd. ed.,* London: Woodfall, 1738.

de Moivre, Abraham, *Approximatio ad Summam Terminorum Binominii $(a+b)^n$ in Seriem Expansi,* printed for private circulation, cited in Hald (1998, p. 795) and in Daw and Pearson (1972).

Edgeworth, F. Y., "On the Probable Errors of Frequency-Constants," *Journal of the Royal statistical Society, 71:* 381-397, 499-512, 651-678; "Addendum," *72:* 81-90, 1908-1909.

Edwards, A. W. F., *Likelihood*. Cambridge: University Press, 1972.

Edwards, A. W. F., "The History of Likelihood," *International Statistical Review, 42:* 9-15, 1974.

Edwards, A. W. F., "What Did Fisher Mean by 'Inverse Probability' in 1912-1922?", *Statistical Science, 12:* 177-184, 1997.

Fisher, R. A., "On an Absolute Criterion for Fitting Frequency Curves," *Messenger of Mathematics, 41:* 155-160, 1912. Reprinted in A. W. F. Edwards, "Three Early Papers on Efficient Parametric Estimation," *Statistical Science, 12:* 39-41, 1997.

Fisher, R. A., "On the Probable Error of a Coefficient of Correlation Deduced from a Small Sample," *Metron, 1:* 3-32, 1921.

Fisher, R. A., "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London, Series A, 222:*309-368,1922. Reprinted in S. Kotz and N. L. Johnson, *Breakthroughs in Statistics, Vol.1,* with an introduction by S. Geisser, pp. 1-44, New York: Springer-Verlag, 1992.

Fisher, R. A., "Theory of Statistical Estimation," *Proceedings of the Cambridge Philosophical Society, 22*:700-725 (1925).

Fisher, R. A., "Inverse Probability and the Use of Likelihood," *Proceedings of the Cambridge Philosophical Society, 28:* 257-261, 1932.

Fisher, R. A., *Statistical Methods, Experimental Design, and Scientific Inference,* being a reprint of *Statistical Methods for Research Workers* (1925), *The Design of Experiments* (1935), and *Statistical Methods and Scientific Inference* (1956). Oxford: University Press, 1990.

Fisher, R. A. , "Inverse Probability and the Uses of Likelihood, *Proceedings of the Cambridge Philosophical Society, 28:* 256-261 (1932).

Fisher, W. D., "Econometric Estimation with Spatial Dependence," *Regional and Urban Economics, 1:* 19-40, 1971.

Greene, W. H., *Econometric Analysis,* 2nd ed., New York: Macmillan,1993, 3rd ed., Upper Saddle River, NJ: Prentice-Hall,1997.

Gauss, Carl Friedrich, *Theoria motus corporum celestium,* Hamburg: Perthes und Besser, 1809. Translation by C. H. Davis in *Theory of Motion of Heavenly Bodies,* New York: Dover, 1963.

Gauss, Carl Friedrich, *Theoria Combinationis Observationium Erroribus Minimis Obnoxiae,* Göttingen: Dieterich, 1823. French translation pp. 1-69 in J. Bertrand, *Méthode des moindres carrés. Mémoires sur la combination des observations,* Paris: Mallet-Bachelier, 1855.

Hald, Anders, *A History of Mathematical Statistics from 1750 to 1930,* New York: Wiley, 1998.

Hogg, R. V., and A. T. Craig, *Introduction to Mathematical Statistics, Fourth Edition.* New York: Macmillan,1978.

Jeffreys, H., "Probability and Scientific Method," *Proceedings of the Royal Society, Ser. A, 146:* 9-16 (1934).

Jeffreys, H., *Theory of Probability, 3rd ed.,* Oxford: University Press, 1961; 1st ed. 1939.

Judge, G. G., R. C. Hill, W. E. Griffiths, H. Lütkepohl and T-C. Lee, *Introduction to the Theory and Practice of Econometrics, 2nd Edition.* New York: Wiley, 1988.

Kelejian, H. H., and D. P. Robinson, "Spatial Autocorrelation: A Suggested Alternative to the Autoregressive Model," pp. 75-93 in L. Anselin and R. Florax, *New Directions in Spatial Econometrics,* New York: Springer-Verlag, 1995.

Kingman, J. F. C., and S. J. Taylor, *Introduction to Measure and Probability,* Cambridge: University Press, 1966.

Koopmans, T. C., and W. C. Hood, "The Estimation of Simultaneous Linear Economic Relationships," pp. 112-199 in W. C. Hood and T. C. Koopmans, eds., *Studies in Econometric Method,* New York: Wiley, 1953.

Laplace, Pierre Simon, "Mémoire sur la probabilité des causes par les évèments," *Mémoires de l'Académie Royale des Sciences Presentés par Divers Savans, 6:* 621-656, 1774. Translated in S. Stigler "Laplace's 1774 Memoir on Inverse Probability, " *Statistical Science,1:* 359-378 (1986).

Laplace, Pierre Simon, "Mémoire sur les approximations des formules qui sont fonctions de très grands nombres et leur application aux probabilités," *Mémoires de l'Académie des Sciences de Paris, 1re Serie, 10:* 353-415; "Supplément," 559 - 565, 1810.

Laplace, Pierre Simon, *Théorie analytique des probabilités,* Paris: Courcier, 1st ed., 1812, 3rd ed. 1820, with supplements.

Larkin, J. H., and J. B. Kadane, "A Method for Maximizing Likelihood Functions," pp. 453-472 in S. Geisser, J. S. Hodges, S. J. Press and A. Zellner, *Bauesian and Likelihood Methods in Statistics and Econometrics,*Amsterdam: North-holland, 1990.

Leamer, E. E., "Let's Take the Con out of Econometrics," *American Economic Review, 73:* 31-43, 1983.

Legendre, Adrien Marie, *Nouvelles méthodes pour la détermination des orbites des comètes,* Paris: Courcier, 1805.

Lindsey, J. K, *Parametric Statistical Inference,* Oxford: Clarendon Press, 1996.

Neyman, J., "Frequentist Probability and Frequentist Statistics," *Synthèse, 36*: 97-131, 1977.

Neyman, J., and E. S. Pearson, "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," *Biometrika, 20A:* 175-240, 263-294, 1928.

Neyman, J., and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London, Ser. A, 231:* 289-337. 1933a.

Neyman, J., and E. S. Pearson, "The Testing of Statistical Hypotheses in Relation to Probabilities A Priori," *Proceedings of the Cambridge Philosophical Society, 29:* 492-510, 1933b.

Norden, R., "A Survey of Maximum Likelihood Estimation," *International Statistical Review,* Part 1, *40*: 329-354, (1972), Part 2, *41:* 39-58 (1973).

Pearson, Karl, *The Grammar of Science,* London: Walter Scott, 1892.

Pearson, Karl, "On the Criterion that a Given System of Deviations from the Probable Case of a Correlated System of Variables is Such that It Can Reasonably Be Supposed to Have Arisen from Random Sampling," *Philosophical Magazine, 5th Ser., 50:* 157-175, 1900. Reprinted in S. Kotz and N. L. Johnson, *Breakthroughs in Statistics, Vol.2,* with an introduction by G. A. Barnard, pp. 1-28, NewYork: Springer-Verlag, 1992.

Pearson, Karl, "On the Systematic Fitting of Curves to Observations and Measurements, Parts I and II," *Biometrika, 1:* 265-303 and *2:* 1-23, 1902-1903.

Plackett, R. L., "The Discovery of the Method of Least Squares, *Biometrika, 59:* 239-251, 1972.

Pratt, J. W., "F. Y. Edgeworth and R. A. Fisher on the Efficiency of Maximum likelihood Estimation," *Annals of Statistics, 4:* 501-514, 1976.

Press, S. J., *Bayesian Statistics: Principles, Models and Applications.* New York: Wiley, 1989.

Rao, C. R., "Apparent Anomalies and Irregularities in Maximum Likelihood Estimation," Sa*nkhyã,24:* 73-102 (1962).

Royall, R., *Statistical Evidence: A Likelihood Paradigm,* London: Chapman & Hall, 1997.

Santayana, George, *The Life of Reason, or, The Phases of Human Progress, Vol. 1, Reason in Common Sense,* New York: Chas. Scribner & Sons, 1905.

Savage, L. J., *The Foundations of Statistics,* New York: Wiley, 1954.

Savage, L. J., "On Rereading R. A. Fisher," *Annals of Statistics, 4:* 441-483 (1976).

Saxe, John Godfrey, *The Poems: Complete Edition,* Boston: Houghton, Mifflin & Co., 1880.

Simpson, T., "A letter to the Right Honorable George Earl of MacClesfield, President of the Royal Society, on the advantage of taking the mean of a number of observations, in practical astronomy," *Philosophical Transactions of the Royal Society of London, 49:* 82-93, 1755.

Stigler, S. M., "Laplace, Fisher, and the Discovery of the Concept of Sufficiency," *Biometrika, 60:* 439-445 (1973).

Stigler, S. M., "Laplace's 1774 Memoir on Inverse Probability," *Statistical Science, 1*: 359-378, 1986. Introduction to and translation of P. S. Laplace, "Mémoire sur la probabilité des causes par les évènemens," 1774, p. 27-65 in *Oeuvres Complètes de Laplace Vol. 8,* Paris: Gauthiers-Villars, 1886-1912.

Stigler, S. M., *The History of Statistics: The Measurement of Uncertainty before 1900,* Cambridge, MA: Harvard University Press, 1986.

Stigler, S. M., "Daniel Bernoulli, Leonard Euler, and Maximum Likelihood," unpublished, dated May 15, 1998.

Tanner, M. A., *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, Third Edition.* New York: Springer-Verlag, 1996.

Whittle, P., "On Stationary Processes in the Plane," *Biometrika, 41:* 434-449, 1954.

Zabell, S., "R. A. Fisher on the History of Inverse Probability," *Statistical Science, 4:* 247-263, 1989.

Zellner, A., *An Introduction to Bayesian Inference in Econometrics,* New York: Wiley, 1971.