UNIVERSITI MALAYA

Syarahan Perdana

Modelling Data: Knowledge Discovery or Empirical Dexterity?

Profesor Dr. Shyamala Nagaraj

UNIVERSITI MALAYA

Syarahan Perdana

Modelling Data: Knowledge Discovery or Empirical Dexterity?

Professor Dr. Shyamala Nagaraj
Department of Applied Statistics
University of Malaya

Inaugural Lecture
Dewan FEA
29 March 2005
(Revised 30 March 2005*)

* *I am grateful to Lee Kiong Hock and Vijay N. Nair for useful comments on an earlier draft. I am, of course, solely responsible for any errors therein.*

Shyamala Nagaraj
Department of Applied Statistics
University of Malaya

Modelling Data: Knowledge Discovery or Empirical Dexterity?

Modelling with data is an important academic activity in almost every area of study. Modelling uses statistically based methodologies to seek out patterns amidst variability in data. Such patterns may be described through theoretical models or established from data analyses. The data can come from tightly controlled and replicable experiments that make for clear statement of the nature of variability. Alternatively, data may be obtained through observation; in which case the nature of variability may be less clear and often the exploratory nature of the investigation make the statistical questions less well defined. Although in both cases sound statistical methods may be used, going from data to the answers requires careful selection, development and testing of appropriate models that seek to quantify and synthesize often complex relationships. The uncertainty in variability makes this process less well-defined and contentious. The merit in statistical modelling in creating knowledge has been

acknowledged in many areas; for example, it remains an essential step in the development and adoption (usually based on experimental data) and even discontinuance (often based on observational data) of medical drugs. On the other hand, it led to the accusation that econometrics (the application of statistics in the field of economics) is little more than "junk science" in reference to diametrically opposed conclusions about the efficacy of gun control laws by two groups of researchers working on the same set of data! Does modelling lead to true knowledge creation or does it merely reflect empirical dexterity on the part of the researcher? This lecture considers the debate and highlights the need for a proper modelling procedure to get from data to knowledge.

Introduction

It is only appropriate that I begin this talk by recognizing those who have led me up this fascinating path of statistics and data analysis, as well as those who have supported my travails. I start by acknowledging:

- My teachers:
  - Brother Celestine, my Form Six Economics teacher at the St John's Institution, who introduced me to the skills and rewards of analytical thinking, and
  - My lecturers at the then Division of Statistics here at the Faculty of Economics and Administration, who showed me how exciting statistics and mathematics could be.
- Statisticians who have encouraged me:
  - Dr Cheong Kee Cheok (Econometrics), my M. Ec. Dissertation supervisor,

- Professor Paul Shaman (Time Series), my Ph. D thesis supervisor,
- Professor Donald Morrison (Multivariate Analysis), my graduate advisor,
- Professor George Box (Multivariate Time Series Analysis & Experimental Design) – a remarkable person, a prolific statistician writing new material even today at the age of 85, and
- Professor Vijay N. Nair (Much needed life line!).

➢ My co-researchers, many of whom I am proud to say, have been co-authors:
- Dr Lee Kiong Hock
- Dr Lim Lin Lean and Dr Mavis Puthucheary
- My colleagues
- My students

➢ And last but most certainly not least, my parents, husband and children who have provided the kind of support necessary for a very engrossing activity such as research.

My work began with modelling in macroeconomics, but has subsequently widened to include areas in social science and medicine, particularly of observational data. I will talk of modelling in a broad way; that is, not of the types of models but how a model may be useful for analyzing data. I take knowledge to be, as per Wikipedia's definition, "the awareness and understanding of facts, truths or information gained" in the "possession of interconnected details which, in isolation, are of lesser value." That is, the modelling of data can be

considered to have made a contribution to knowledge *only when it has added value to the knowledge already available.*

I first provide some examples to highlight the value of modelling. I then provide two examples where data modelling has led to knowledge discovery. In this respect, I have selected examples that I believe have been critical to the development of *more* knowledge. Modelling data has led to disillusionment in certain quarters, especially when decisions that touch on our personal lives are made on the basis of the findings of empirical analyses. I consider two examples that highlight the nature of concerns with modelling. I go on to discuss what modelling data encompasses, and highlight the essentials of the modelling exercise, one that I have found useful in the search for structure irrespective of area of research. I conclude with some remarks regarding the future for statistical modelling.

The Value of Modelling

It is natural that I speak of modelling data, an activity that I have been engaged in the past thirty odd years since I began graduate work here at the Faculty of Economics and Administration. The more and diverse the data sets that I have worked on, the stronger is my view that good data analysis is less of method and more of clear reasoning. An example shows this best.

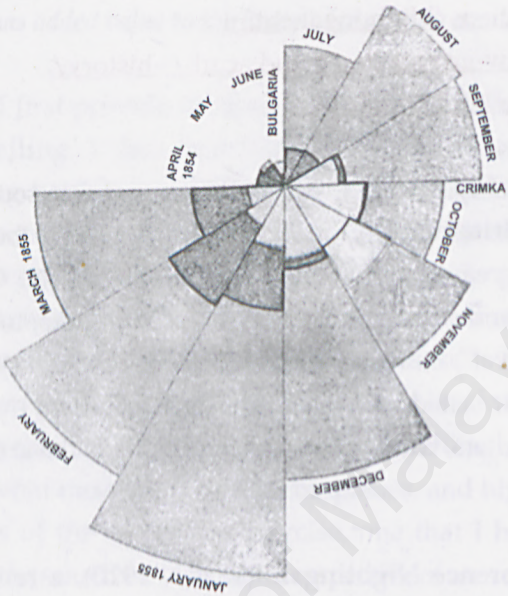
The Work of Florence Nightingale
http://www.uh.edu/engines/epi1712.htm;
http://www.pims.math.ca/education/2001/women/july/.

http://www-groups.dcs.st-and.ac.uk/~history/Mathematicians/Nightingale.html
http://www-groups.dcs.st-and.ac.uk/~history/Quotations/Nightingale.html

In 1853, Turkey declared war on Russia and was soon joined by Great Britain and France. The Crimean War began the following year when the British landed on the Crimean Peninsula and set out, with the French and Turks, to take the Russian naval base at Sevastopol. They succeeded a year later and the war lasted two years, during which time over half a million soldiers, battling terrible conditions in the armies, lost their lives.

Florence Nightingale (1820 – 1910), a remarkable woman by all measures, arrived in Turkey as a nursing administrator in 1854. In order to convince her superiors of the unsanitary conditions of army hospitals, she collected data and systematized record-keeping practices, and created a graphical display, the Polar-Area Diagram, to show the information. Figure 1 reproduces a copy of the diagram. Nightingale's graph is like a pie chart, cut into twelve equal angles. But the comparison ends there. The area of each coloured wedge, measured from the centre, is proportional to the statistic being represented. These slices advance in a clockwise direction. Each shows what happened in one month of one year. The outward reach of each slice shows how many deaths occurred in that month. We see little short slices in April, May and June of 1854. After the troops land in the Crimea, the slices begin reaching far outward in the radial direction. There is detail in each wedge. Blue wedges (the outer area) represent

Figure 1. *Florence Nightingale's Polar Area Diagram*



Source: http://www.pims.math.ca/education/2001/women/july/;

deaths from preventable or contagious diseases, pink wedges (in the centre) deaths from wounds and grey wedges (in-between) deaths from all other causes. Except for the bloodiest month in the siege of Sevastopol, battle deaths take up a very small portion of each slice. Even the Charge of the Light Brigade yielded only a modest fraction of the total deaths in that month. The diagram shows clearly that soldiers died from diseases more than bullets.

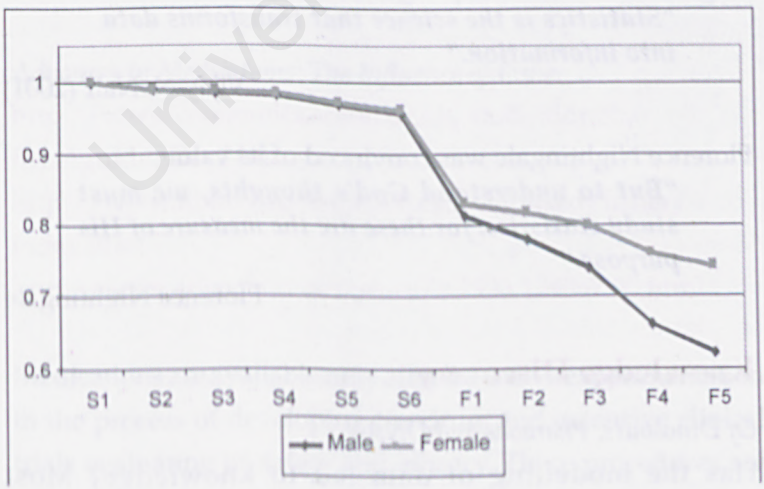
Nightingale's work led to improved city and military hospitals. She developed a Model Hospital Statistical Form for hospitals to collect and generate consistent data. She became a Fellow of the Royal Statistical Society in 1858 and an honorary member of the American Statistical Association in 1874. Karl

Pearson acknowledged Nightingale as a "prophetess" in the development of applied statistics..

Understanding What's Happening

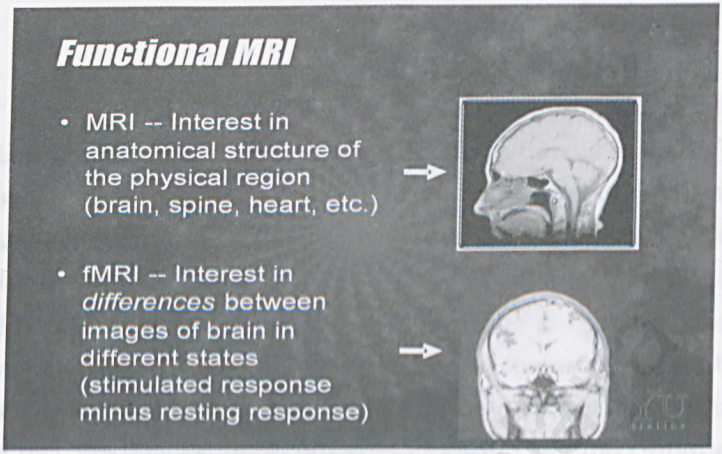
Two more examples of the value of modelling are provided. In trying to understand the observed higher educational level of females among young workers in the labour force (a reversal of earlier patterns), data on transition was collected on a single cohort (Nagaraj and Lee, 2003). Figure 2 shows the steadily declining presence of males as the 1988 cohort made its way through Malaysian public school system. Figure 3 shows that a more complex question, how the brain remembers, can be studied by comparing the changes observed in the brain when it remembers a phone number. The preceding examples demonstrate that clear reasoning and an organized way of learning are essential to effective and useful knowledge. That is, thinking drives the method, not the other way around.

Figure 2. *Persistence to Form Five, 1988 Cohort, Malaysia*



Source: Nagaraj and Lee 2003

Figure 3. *How Does the Brain Remember?*



Source: Christensen 2003

The modelling of data brings us to the realm of statistics, a discipline recognized in its own right only in the early twentieth century through Ronald Fisher's contributions. Sound modelling in fact needs good statistics as

> ***"Statistics is the science that transforms data into information."***
>
> Vijay N. Nair (2001)

Florence Nightingale was convinced of its value:

> ***"But to understand God's thoughts, we must study statistics, for these are the measure of His purpose."***
>
> Florence Nightingale

Knowledge Discovery

Of Dinosaurs, Pharaohs and DNA

Has the modelling of data led to knowledge? Most emphatically, yes! Some dramatic examples of knowledge

discovery come to mind. Archaeology has used principal components analysis to search for structure in bones, and discriminant analysis to classify unknown specimens, in order to tell us about ancient times. The information available today on pharaohs, mummies and their tombs is considerable. Books abound on dinosaurs, their many forms, even how their eggs looked! Urine testing for drug abuse is a regular feature of many sport competitions, while DNA testing is now routine in establishing identity. Identification is fundamentally a statistical pattern recognition problem. Fingerprinting, the weather map, the Human Genome Project, the list is quite extensive. These developments in statistics have been driven by the desire for the knowledge, and the stories of the personalities behind them make fascinating reading (see, for example, Salsburg, 2001).

I describe below two examples of knowledge discovery that I believe have promoted even greater discovery and which have become an integral part of research today.

Advances in Medications: The Influence of Fisher

http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/Doc1.htm

http://www.library.adelaide.edu.au/digitised/fisher/index.html

http://www.foxnews.com/story/0,2933,134057,00.html

Medications are available only after extensive experimentation in the process of developing the drug, and extensive clinical trials evaluating its safety and efficacy. These procedures are required by the U.S. Food and Drug Administration (USFDA)

prior to approval. Indeed, virtually every scientist must subject their data to some form of statistical analysis. The notion of an "objective proof" in a statistical analysis (in which there is really no proof) was largely due to R.A. Fisher (1890-1962). Fisher made many contributions to both statistics and genetics. Much of this work was done at Rothamsted Agricultural Experiment Station, the oldest agricultural research institute in the United Kingdom, established in 1837. There he studied the design of experiments by introducing the concept of randomization and the analysis of variance, procedures now used in many fields including drug development. Fisher's idea was to arrange an experiment as a set of partitioned sub-experiments that differ from each other in having one or several factors or treatments applied to them. The sub-experiments were designed in such a way as to permit differences in their outcome to be attributed to the different factors or combinations of factors by means of statistical analysis. He also provided a number of tests and wrote of the importance of significance in reaching conclusions about the results of an experiment. In 1925, Fisher published his techniques in a book, *Statistical Methods for Research Workers*, which lays claim to being one of the most influential texts in the history of science.

Survey Sampling: The Works of J Neyman (Berkeley), P C Mahalanobis (Indian Statistical Institute), MH Hansen (US Bureau of Census)

http://www.isical.ac.in/prof.html

http://www.mrs.umn.edu/~sungurea/introstat/history/w98/Jerzy_Neyman.html

http://stills.nap.edu/html/biomems/mhansen.html

In the early 20th century, as governments and populations grew, it became clear that there were problems of coverage of censuses. Jerzy Neyman (1894-1981) who made great contributions in probability theory, hypothesis testing, confidence intervals and other areas of mathematical statistics, also developed a theory of survey sampling in 1934; most importantly, the use of probability sampling for cluster samples. This was then used for a labour survey in Poland. The practical aspects of survey methodology were developed separately by M.H. Hansen (1910-1990) at the U. S. Bureau of Census and P.C. Mahalanobis (1893-1972) at the Indian Statistical Institute. Hansen developed the sampling theory necessary for the efficient conduct of large-scale national surveys, the establishment of formal quality control methods for surveys, and the derivation of theory and models for analyses of non-sampling errors. Mahalanobis' contributions include optimal choice of sampling design using variance and cost functions, and the technique of interpenetrating network of sub-samples (Deming's replicated sampling) for assessment and control of errors, especially non-sampling errors, in surveys.

Hansen, an accountant by training, was first involved when he designed a survey for unemployment in 1937. His report, with colleague C.L. Dedrick, was considered a major innovation in the philosophy of sampling. He also devised techniques to address non-response. Mahalanobis persevered for almost a decade to get sample surveys to be recognised as being more accurate and cost effective than complete enumeration. Practices common today for the management and quality control of data, such as pilot surveys, field

supervisors and the like were devised by these pioneers. As Rao (1999) notes, Hansen and Mahalanobis were the advocates of 'Total Sample Survey Design'. Under the chairmanship of Mahalanobis (1947-1951), the United Nations Sub-Commission on Statistical Sampling recommended that sampling methods be extended to all parts of the world.

These examples represent but only a tiny fraction of the areas in which modelling data has contributed significantly to knowledge. Its noticeable contributions range across diverse fields, from archaeology to zoology, etc., and it has indeed increased our understanding of our world and that has, in turn, increased the welfare of mankind in general.

Empirical Dexterity

Does modelling lead to true knowledge creation or does it merely reflect empirical dexterity on the part of researchers? This is an important question as modelling data, doing statistical analyses, and discussing the results are core to research activity anywhere in the world. It appears that where findings have impacted on personal rights, the very core of the analyses – the modelling, the testing, the assumptions - have been questioned. I reproduce here abstracted texts and a picture from two particularly provocative articles.

With regards to medicine:

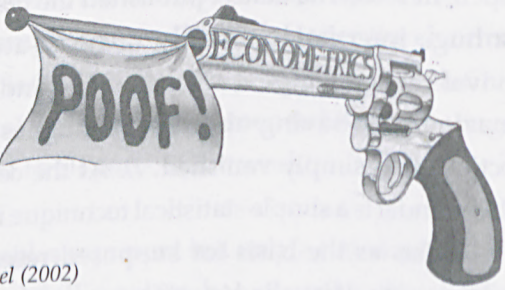
"If you were going to have a heart attack, it seemed there was never a better time than the early 1990s. Leading medical journals were regularly reporting results from trials of new treatments for heart attacks that weren't just good - they were

incredible. ... But then something odd began to happen. In 1995, The Lancet published the results of a huge international study of heart-attack survival rates among 58,000 patients - and the "amazing" life-saving abilities of magnesium injections had simply vanished. ... At the centre of this scandal is a simple statistical technique used by scientists as the basis for supposed research breakthroughs. It is called 'significance testing'.... And it is fatally flawed. ... Just why has the scientific community failed to act? The answer lies in its squeamishness about subjectivity. It is hard to convey the strength of emotion aroused within the scientific community by the "S-word". Subjectivity is seen as the barbarian at the gates of science, the enemy of objective truth. ... The implications are stark. It means that vital scientific questions - whether a new heart drug is seen as effective or whether breast implants trigger disease, for example - are being decided by an entirely arbitrary standard [p-value]. ... Curiously for a profession supposedly dedicated to discovering truths, the reliability of research findings is never mentioned." (Matthews, 1998).

And with regard to econometric modelling (see also Figure 4):

"....Do you believe that a 1% increase in the number of citizens licensed to carry concealed weapons causes a 3.3% *decrease* in the state's murder rate? Do you believe that 10 to 20% of the decline in crime in the 1990s was caused by an

Figure 4. *Goertzel Dismisses Econometerics*



Source: Goertzel (2002)

increase in abortions in the 1970s? .....If you were misled by any of these studies, you may have fallen for a pernicious form of junk science: the use of mathematical models with no demonstrated predictive capability to draw policy conclusions. These studies are superficially impressive. Written by reputable social scientists from prestigious institutions, they often appear in peer-reviewed scientific journals. ... John Lott, an economist at Yale University, used an econometric model to argue that "allowing citizens to carry concealed weapons deters violent crimes, without increasing accidental deaths."... [Subsequently] Dan Black and Daniel Nagin (1998) published a study showing that if they changed the statistical model a little bit, or applied it to different segments of the data, Lott and Mustard's findings disappeared.....The *Philadelphia Inquirer's* David Boldt [a journalist], after hearing John Lott speak on concealed weapons and homicide rates, and checking with other experts, lamented that "trying to sort out the academic arguments is almost a

fool's errand. You can drown in disputes over t-statistics, dummy variables and 'Poisson' vs. 'least squares' data analysis methods." (Goertzel, 2002).

Statistics is a subject that deals with data, *variation* and *chance*. In a single analysis, a statistical sample provides only an estimate of a value that will never be known. The solution is to get an approximation that is close enough to the (unknown) true value to be usable. In some cases the line of reasoning is clear. For example, how cigarette smoking causes lung cancer is not exactly known, but the fact that cigarette smoking does indeed cause lung cancer is nevertheless shown by a clear correlation between smoking and the instance of lung cancer. On the other hand, the difficulty with measurement is shown by the results of the 2000 presidential election in the United States, where the vote was essentially a tie, and the differences were well within all possible measures of sampling error. Even more disturbing is the situation with the drug Vioxx, the withdrawal of which was financially debilitating for Merck (US$25 billion losses on the share market overnight) (Mandell, 2001; Shatz, 2004; Edwards, 2005). This drug, hailed as a wonder drug when first introduced, would have cleared the rigorous requirements of the USFDA before being approved. Yet, the drug's safety record was questioned based on findings from an observational study.

Modelling Data

In this section, I highlight several aspects that need to be considered in any modelling exercise, irrespective of area of research.

Initial Knowledge

The modelling process requires that there is some knowledge to begin with, sufficient in fact to describe a tentative mathematical or conceptual model. A model is usually a simplified representation of a complex reality, and is meant to assist clear thinking and reasoning. It seeks to tie the object or outcome of interest to related factors or objects. The conception of this model ensures that sufficient thought is given to the way the outcome relates to other objects. Without a well-thought out conceptual model, it can be difficult to appreciate the data that are needed, how they can be measured and what the data show, and to establish whether knowledge has been enhanced.

Getting the Data

Data collection, whether by experimentation or observation is a major step. Despite the advances made in respect of accuracy and reliability of data collection, there remain however many difficult issues. The Vioxx controversy is a good example. For example, how close should the sample be to the population? In the clinical trials, many factors would have been controlled for, factors that cannot be controlled in the general populace that would consume the drug. Can indeed the results of controlled experiments then be considered valid for a population towards which the drug is targeted? The observational sample was made up of persons with a specific medical history. Are the results then valid for the general population for which the drug is useful? Should there have been a narrower definition of the population for whom the drug would have been useful?

Measurement

The design and collection of data impact upon the quality of data, their coverage, reliability as a measuring tool and validity (representativeness or relevance). The measurement model (Griliches, 1986) is the model used to empirically define the conceptual variables of interest. While this is an obvious issue in the social sciences where abstract matters such as opinions are quantified through various methodologies, it also occurs in the sciences. Take, for example, the Vioxx controversy. Did the sample design control for factors that may confound the measured outcome? The dosage and frequency of use of the drug, as well as the length of time after which the event was measured differed between the clinical trials and the observational study. Another dramatic example is that of the Challenger Space Shuttle which crashed in 1986 killing seven astronauts. The O-rings used had been tested in trials where the temperature settings were much warmer than the temperature on that day.

The issue becomes more complicated when the data are not collected directly by the analyst. Often information on the process of measurement may not be available, or as is more often the case, not in the form suitable for analysis (Nagaraj and Mahani, 1986; Nagaraj and Kok, 1990). For example, it is not always apparent that we may be dealing with essentially an identity. Some cases are obvious, such as an analysis of profit, cost and revenue. Others are much less so.

Take the case of growth accounting which seeks to explain the remainder after subtracting changes in labour and capital from changes in growth as measuring as total factor

productivity (TFP), an approach to growth made popular through the works of Dennison (1967). However, as Felipe (2004) shows, the residual is no more than a weighted average of the contributions of labour and capital, arising from the functional distribution of income expressed as an identity of the national accounts of a country. Conceptually, productivity being obtained as a function of how labour interacts with capital may be appealing. However, the division of changes in growth into components depends on the way the labour and capital series are related. Indeed, as Felipe points out, a TFP value of zero does not mean a lack of productivity but that the two components in the residual cancel out each other.

The Model

The statistical model that is used with the data has assumptions in order to explain the nature of variation in the data. The following often-cited quote emphasises tentative nature of models:

> "All models are wrong, but some are useful." George E. P. Box (1979a: 202)

At each step of the way, there are decisions to be made. Although there are procedures, methods and common practices to guide the data analyst, there is nevertheless sufficient room for what Leamer (1983) calls the "whimsical character of inference" especially when facts are not available:

> " Sometimes I take the error terms to be correlated, sometimes uncorrelated; sometimes normal and sometimes nonnormal; sometimes

> I include observations from the decade of the fifties, sometimes I exclude them; sometimes the equation is linear and sometimes nonlinear; sometimes I control for variable z, sometimes I don't. Does it depend on what I had for breakfast?"

The problem is, as those of us who have engaged in this activity would have come to realize, modelling is an absorbing activity both in terms of time and effort. And it is not surprising then we can recognize ourselves in the following quote ascribed to George Box (as quoted in De Veaux, 2004):

> "Statisticians, like artists, have the bad habit of falling in love with their models."

Tests of Significance

Evaluation of the importance of findings uses some form of statistical testing. The decision to reject or accept a hypothesis is based on the level of significance (entirely arbitrary but with conventional values regarding the probability of rejecting the null hypothesis when it is true) and power (a function of sample size, and measuring the ability to reject the null hypothesis when it is false). There is also the distinction between statistical significance and substantive significance, one that is confounded by sample size. In a small sample, an important effect may go undetected, but in a large sample, a substantive effect may be significant.

That the p-value has been used as a tool for the mechanical selection of "important" factors by many

researchers is undeniable. Even some academic journals have been guilty of approving articles that report only significant results. The debate over tests of significance is not new and dates back to the time of Fisher and the disagreements with Pearson and Neyman (see, for example, Morrison and Henkel, 1970). It is not surprising then that many data analysts (including myself) in very different fields eventually write or discuss these matters either generally or in reference to particular statistical issues (for example, Tukey, 1954; Box, 1979b; Nagaraj, 1993, 1994a, 1994b). For example, Selvin (1957) declares:

> *"In design and in interpretation, in principle and in practice, tests of statistical significance are inapplicable in non-experimental research."*

Winch and Campbell (1969) go a step further:

> *"To do or not to do a test of significance – that is a question that divides men of goodwill and sound competence."*

Robert Matthews (1998), a highly qualified science journalist, in his article cited earlier demands:

> *"…. 70 years ago Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding. It is time to pull the plug."*

Yet in equal measure, and more, significance testing has surely facilitated the iterative process of knowledge gain. In a special issue on the value of p, the editors of *Epidemiology* (2001) invited three commentaries on the subject (Weinberg, 2001; Poole, 2001; Goodman, 2001) in deciding whether to continue the previous (and founder) editor's stringent policy on discouraging the use of p values. They conclude that there are settings in which p-values can be informative but they would have to always be on the alert for possible abuse.

Studying the Data

The most important aspect in modelling is to be fully acquainted with the data. With the availability of speed and ease of statistical software packages, it is often all too easy to just throw in all the variables, apply a method and try to decipher the output. That should in fact be the last step, a step taken only if the data support it. The first step would be to inspect the data visually, and to then understand the basic patterns through a tabular analysis. To come to decisions about the nature of the object of interest, the data analyst must look at an adequate range of values, especially at the extremes. It is necessary that the data be visually explored to know whether a transformation is needed, or whether a postulated relationship is apparent.

The analyst must ascertain that the selected method of statistical analysis is appropriate to achieve the objective of the study. Florence Nightingale showed that an analyst does not need to use advanced techniques to show knowledge. But complex problems may need advanced techniques if there is

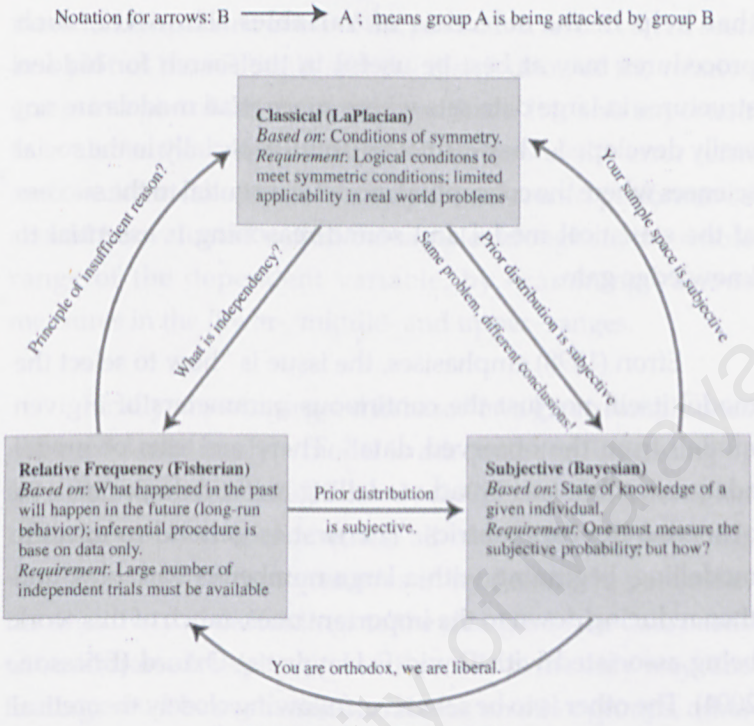
to be gain in knowledge. Efron (1998) notes:

> *"...statisticians are being asked to solve bigger, harder, more complicated problems, under such names as pattern recognition, DNA screening, neural networks, imaging and machine learning. New problems have always evoked new solutions in statistics, but this time the solutions might have to be quite radical ones."*

Modelling is complicated by a great many decisions that need to be made. For example, the stochastic assumption can lead to many different models, the functional form may not be easily established, estimation is not always easy if the method is not least squares, and tests can lead to different conclusions. Which variables are exogenous and which are endogenous? Which should be controlled for? Which variables should be in the model, and if so, in what form? A related issue is the position regarding inferential statistics. Figure 5 shows the differences between the classical, Fisherian and Bayesian schools of thought regarding inference. In practice, as Arsham (2002) emphasizes, the statistician "uses whatever method that comes in handy."

Freedman (1994) observes that assumptions are often not tested. He highlighted the frequent use in the social sciences of the dummy variable to represent group behaviour in regression analysis without testing the validity of the implied assumptions. If, for example, group effects are being evaluated, the proposed functional form and relationship should be first established for each group before the data are pooled (Nagaraj,

Figure 5. *The Three Major Schools of Thought in Inferential Statistics*



***Conclusion*: Working statisticians use whatever methods come in handy from a variety of approaches.**

Source: Hossein Arsham;2002. Business Statistics:Revealing Facts from Fifures http://163.121.24.109/decision_making_tools/opre504.htm

2001). Furthermore, only some assumptions are testable. Indeed as Joiner (1982) clarifies:

> ***"in fitting the data, we pretend that the data can be described in the fashion assumed."***

The selection of relevant variables and of the model after estimation is another tricky question. There are automated

procedures such as forward and backward stepwise regression that help in the selection of variables. However, such procedures may at best be useful in the search for hidden structures in large data sets where conceptual models are not easily developed. They cannot be useful especially in the social sciences where the conceptual model is essential to the success of the statistical model and sound reasoning is essential to knowledge gain.

Efron (1998) emphasises, the issue is "how to select the model itself, not just the continuous parameters of a given model, from the observed data." There are tests of model adequacies, but two broad modelling methodologies may be observed in econometrics. The first is general-to-specific modelling, beginning with a large number of variables, and then reducing down to the important ones, much of this work being associated with David F. Hendry at Oxford (Ericsson, 2004). The other is to be selective, following closely theoretical constructs or prior information about the variable of interest, identifying structure in the data until the residuals are white noise, the approach of the Box and Jenkins (1970) methodology. It is interesting that the adoption of these methodologies have been tied to developments in software, computational techniques and computer speeds. The search for structure gets harder the more the data. Sala-i-Martin (1997), for example, ran two million regressions in the context of growth regressions. Hendry (2004) shows that his analysis could have been achieved with one regression using the general unrestricted model with *PcGets*, a program that does automated econometric model selection (see a very positive review by Bardsen, 2001 arguing that the program actually facilitates reasoning).

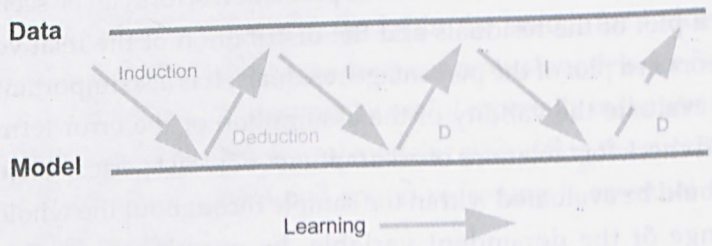
Diagnostics are important to test for the adequacy of model fit. The distribution of the prediction errors can be seen in a plot of the residuals and the distribution of the relative errors in a plot of the percentage residuals. It is also important to evaluate the validity of the assumption of the error term and check for violations of assumptions. Forecast performance should be evaluated within the sample throughout the whole range of the dependent variable, by examining various measures in the lower-, middle- and upper- ranges.

Finally, ensuring a good fit based on a given sample does not necessarily imply a model is adequate for forecasting or for a different set of data. Kennedy (2002) provides a list of ten rules for applied econometrics, all of which emphasise the nature, pitfalls and uncertainties in modelling. In particular, the sensitivity of the findings to the changes in the data needs to be addressed. Replication of the study is the only way. The findings must be evaluated on subsets, on other groups, other time periods.

Concluding Remarks

So does modelling of data lead to knowledge discovery or is it merely empirical dexterity? Karl Popper, the famous philosopher of science, argues that for any claim to be declared scientific (and therefore knowledge adding) it must be falsifiable (Popper, 1963). Then this is where modelling has a central role. So the modelling of data should indeed add to knowledge. The process of modelling and data analysis that leads to knowledge gain is essentially iterative (Figure 6). Box

Figure 6. *The Process of Knowledge Gain*



(George Box, 2000 Deming lecture,
Statistics for Discovery, Box and Youle 1953)

(1979b; 2000) describes it thus:

"Data or facts lead us to induce a possible model, theory, hypotheses, conjecture or idea. (I will regard these as essentially the same thing and use the word "model" to encompass them all.) The induced model then leads us by a process of deduction to consider the kinds of things that should happen if that model were true and what data we ought to get to compare what we thought would happen with what actually occurred. The nature of the discrepancies would then lead us to induce an appropriately modified model and so on".

That there will be detractors is inevitable, for that is the nature of the modelling. Knowledge will continuously be added to, and there can be no final word on any matter; just greater confidence in the knowledge that has been found, and that added knowledge adds also to our reasoning powers. However, knowledge gain that is useful or that adds value must be based on sound reasoning, sound observation and sound modelling and testing. Irrespective of differences in

thought regarding inferential statistics (Fisherian versus Bayesian), or modelling strategies (Hendry versus Box-Jenkins), all serious data analysts will have some form of prior [reasoning and] knowledge.

Modelling data is essential to knowledge discovery. And modelling requires good statistics to get the knowledge from the data:

> *"I like to think of statistics as the science of learning from data."*
>
> John Kettenring, ASA President, 1997
> (http://www.amstat.org/careers/index.cfm?fuseaction=whatisstatistics)

The word science comes from the Latin 'scientia' which means knowledge, and knowledge can be gained from reasoning, as well as from observation. Statistics and modelling will continue to be important in the future more so in a world that is seeing an explosion in the amount of data. The future lies on the analysis of really large data sets such as in astronomy (Maindonald, 1998; Szalay, 1999). I provide below an example that hints at the dramatic prospect for the future of modelling and knowledge discovery.

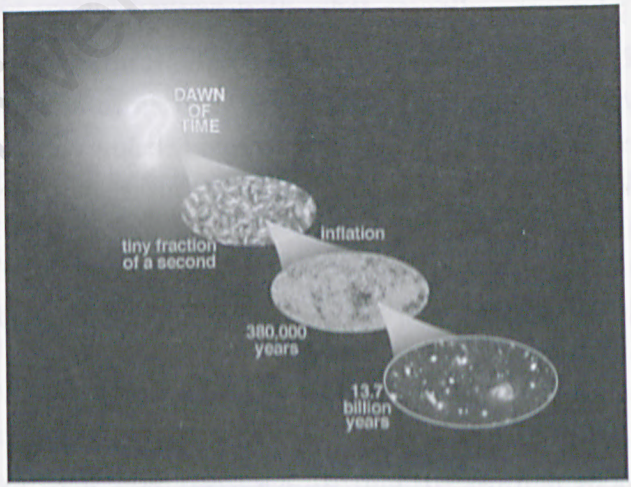
The Work of the Wilkinson Microwave Anisotropy Probe (WMAP) Team

http://map.gsfc.nasa.gov/

The Wilkinson Microwave Anisotropy Probe (WMAP) was launched in June of 2001 and has made a map of the

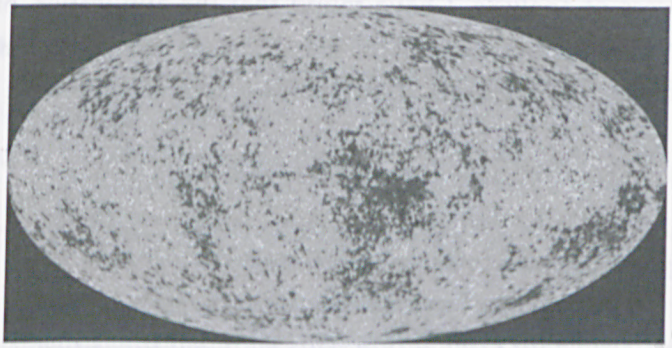
temperature fluctuations of the cosmic microwave background (CMB) radiation with much higher resolution, sensitivity, and accuracy than before. The CMB radiation is the radiant heat left over from the Big Bang, and was first observed in 1965. The properties of the radiation contain a wealth of information about physical conditions in the early universe and a great deal of effort has gone into measuring those properties since its discovery. With the data collected so far, the universe has been estimated as being 13.7 billion years old. Figure 7 shows the postulated growth of the universe, while Figure 8 is the first detailed full-sky map of the oldest light in the universe. This radiation (and by extension, the early universe) is remarkably featureless; it has virtually the same temperature in all directions in the sky. Colours indicate "warmer" (red) and "cooler" (blue) spots. The oval shape is a projection to display the whole sky; similar to the way the globe of the earth can be projected as an oval. The data brings into high resolution

Figure 7. *Cosmic History*



Source: NASA/WMAP Science Team http://map.gsfc.nasa.gov/m_or.html

Figure 8. *The Microwave Sky*



Source: NASA/WMAP Science Team http://map.gsfc.nasa.gov/m_or.html

the seeds that generated the cosmic structure we see today. The new data support and strengthen the Big Bang and Inflation Theories.

This example demonstrates aptly that modelling is more than just a science. Can the modelling of data be as objective as the test of significance used in its analysis purports to be? This is where its success in adding to knowledge makes it an art. In fact, Moore (1998) contends that it is time that statistics is placed among the liberal arts because:

> ***"The liberal arts image emphasizes that statistics involves thinking. It is because statistics involves distinctive and powerful ways of thinking that we will not be swallowed up by information technology."***

I hope that I have been able to argue that statistical reasoning is more important than the method itself in order to ensure

that value-added knowledge is continually created by the process of analysis. I have found that the application of a scientific method to explain the variation around us most fulfilling in this respect. Kish (1959) explains this life well:

> *"The statistical consultant spends much of his [to which I might want to add 'her'] time in the borderland between statistics and other aspects, philosophical and substantive, of the scientific search for explanation. This marginal life is rich both in direct experiences and in the discussion of fundamentals."*

References

Hossein, A., 2002. Business Statistics: Revealing Facts from Figures. http://163.121.24.109/decision_making_tools/opre504.htm. [Accessed 25 March 2005].

Bardsen, G., 2001. Review of PcGets 1 for Windows, *Econometrics Journal,* 4: 311–318.

Below, P., 2003. Extreme Metrics Analysis for Fun and Profit. www.sasqag.org/pastmeetings/below_slides.ppt. [Accessed 24 March 2005].

Edward F. Dennison 1967. *Why Growth Rates Differ*, Washington D.C.: Brookings Institution

Black, D. and D. Nagin, 1998. Do right-to-carry laws deter violent crime? *Journal of Legal Studies,* 27: 209-219.

Box, G.E.P., and P. V. Youle, 1955. The Exploration of Response Surfaces: An Example of the Link between the Fitted Surface and the Basic Mechanism of the System, *Biometrics*, 11, 287-323.

Box, G.E.P. and G.M. Jenkins, 1970. *Time Series Analysis: Forecasting and Control.* New York: Holden-Day.

Box, G. E. P., 1979a. Robustness in scientific model building. In R. L. Launer, & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201-236). New York: Academic Press.

Box, G. E. P., 1979b. Some Problems of Statistics and Everyday Life, *Journal of the American Statistical Association* 74(365):1-4.

Box, G. E. P., 2000. Statistics for Discovery, *Journal of Applied Statistics*, 2001, 28:285-299..

Christensen, W., 2004. Discovering Hidden Structures. Brigham Young University House of Learning Lecture. statistics.byu.edu/faculty/wfc/Lectures/WFChouseoflearning.ppt. [Accessed 25 March 2005].

De Veaux, R D., 2004. Predictive Analytics: Making it Pay Off www.spss.gr/events/content_03/ Keynotes/DeVeaux_proceedings.ppt. [Accessed 25 March 2005]

Deming, W. E., 1986 *Out of the Crisis*, Cambridge, Mass.: Massachusetts

Institute for Technology, Center for Advanced Engineering Study.

Edwards, B. S., 2005 Vioxx, Celebrex and Aleve: What's a consumer to do? Editor's note www.mayoclinic.com/invoke.cfm?id=PN00065. [Accessed 25 March 2005]

Efron, B., 1998. R. A. Fisher in the 21st Century, *Statistical Science,* 13: 95-122.

The Editors, 2001. The Value of *P*, *Epidemiology*, 12: 286.

Ericcson, N. R., 2004. The ET Interview: Professor David F. Hendry, *Econometric Theory,* 20: 743-1404.

Felipe, J. and J. S. L. McCombie, 2004. Is a Theory of Factor Productivity Really Needed. Working Paper, Centre for Applied Macroeconomic Analysis, Australian National University, http://cama.anu.edu.au/Working%20Papers/Papers/FelipePaper122004.pdf. [Accessed 25 March 2005]

Freedman, D., 1995. Some Issues in the Foundation of Statistics. www.stat.berkeley.edu/users/freedman. [Accessed 25 March 2005]

Goertzel, T., 2002. Myths of Murder and Multiple Regression. *The Skeptical Inquirer*, 26: 19-23. Extended version: Econometric Modeling as Junk Science. www.crab.rutgers.edu/~goertzel/econojunk.doc. [Accessed 23 March 2005]

Goodman, S.N., 2001. Of P-values and Bayes: A modest proposal. *Epidemiology*, 12: 295-297.

Griliches, Z., 1986. Economic Data Issues, in Z. Griliches and M.D. Intriligator (eds),
Handbook of Econometrics, Vol. III, Chap. 25.

Hendry, D. F. and H. Krolzig, 2004. We Ran One Regression, *Oxford Bulletin of Economics & Statistics*, 66: 799-810.

Joiner, B., 1982. Practising Statisticians or What They Forgot to Say in the Classroom. In J. S. Rustagi and D. A. Wolfe (eds), *Teaching of Statistics and Statistical Consulting,* New York: Academic Press: 327-42.

Kennedy, Peter, 2002. Sinning in the Basement: What are the Rules? The Ten Commandments of Applied Econometrics, *Journal of*

Economic Surveys, 16: 569-89.

Kihn, E., 2001. Stopping the Tidal Wave: Techniques for Handling Incredible Data Volume. globalchange.gov/workshop2001/proceedings/Kihn/Eric_Kihn.ppt. [Accessed 25 March 2005]

Kish, L., 1959. Some Statistical Problems in Research Design, *American Sociological Review*, 24: 328-338.

Lang J. M., K. J. Rothman, and C. I. Cann, 1998. That confounded. P-value, *Epidemiology*, 9: 7-8.

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

Lott, John, 2000. *More Guns, Less Crime: Understanding Crime and Gun Control Laws*. Chicago: University of Chicago Press, second edition.

Maindonald, J. H., 1998. New Approaches To Using Scientific Data – Statistics, Data Mining And Related Technologies In Research And Research Training. Occasional Paper, GS 98/2, ANU Graduate School.
eprints.anu.edu.au/archive/00001150/.[Accessed 25 March 2005]

Mandell, B. F., 2001. Cox-2 Inhibitors and Cardiovascular Risk: Point and counterpoint, *Cleveland Clinic Journal of Medicine*, 68: 957-960.

Matthews, R., 1998. The Great Health Hoax, *The Sunday Telegraph* September 13. Full –length text: Facts versus Factions: the use and abuse of subjectivity in scientific research, http://chetday.com/healthhoax.html. [Accessed 25 March 2005]

Moore, D. S., 1998. Statistics Among the Liberal Arts, *Journal of the American Statistical Association*, 93, No. 444, Theory and Methods.

Morrison, D. E. and Henkel, R. E. (eds.), 1970. *The Significance Test Controversy*. Chicago: Aldine Publishing Company.

Nagaraj, S. and Mahani Z. A., 1986. Data Considerations in Econometric Modelling: Some Views on the Malaysian Situation. Discussion Paper No 2, Kuala Lumpur: Malaysian Institute for Economic Research.

Nagaraj, S. and K.L. Kok, 1990. Data methodology and socio-economic

modelling in Malaysia', in *Issues and Challenges for National Development,* Faculty of Economics and Administration, University of Malaya.

Nagaraj, S. and Lee Kiong Hock, 2003. Human Resource Development and Social Reengineering: Which Part of the Field are We Levelling? in J. Yahaya, N. P. Tey and K. K. Yeoh (eds), *Sustaining Growth, Enhancing Distribution: The NEP and NDP Revisited,* Kuala Lumpur: Centre for Economic Development and Ethnic Relations, University of Malaya.

Nagaraj, S., 1994. Significance Testing: Future Directions. Research Seminar, Center for Quality and Productivity Improvement, University of Wisconsin, January.

Nagaraj, S., 1994. Significance Testing: The Controversy. Research Seminar, Center for Quality and Productivity Improvement, University of Wisconsin, January.

Nagaraj, S., 1993. How Significant is Significant? Lecture organised by the Mathematics Department, Institut Teknologi Malaysia, June.

Nagaraj, S., 2001. Some Considerations in Evaluating Group Effects in Non-Linear Regression Models, *Proceedings of the International Conference and Workshop on Recent Developments in Statistics and its Applications, Kuala Lumpur, June 26-30.* 207-214.

Nair, V. N., 2001. Statistics in Industry: Research Opportunities & Challenges, http://www.itl.nist.gov/div898/conf/jrc/abstracts/w1r.html. [Accessed 27 March 2005].

Poole, C., 2001. Low P- Values or Narrow Confidence Intervals: Which are more Durable? *Epidemiology,* 12: 291-294.

Popper, K., 1963. *Conjectures and Refutations: The Growth of Scientific Knowledge.* London: Routledge and Kegan Paul

Rao, T. J., 1999. Early Developments in Sample Surveys: A Tale of Two Personalities. www.stat.fi/isi99/proceedings/ arkisto/varasto/rao_0819.pdf. [Accessed 25 March 2005]

Sala-i-Martin, X. X., 1997. I have just run Two Million Regressions. *American Economic Review,* 87, 178–183.

Salsburg, D., 2001. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, London: Palgrave Macmillan.

Selvin, H., 1957. A Critique of Tests of Significance in Survey Research, *American Sociological Review* 22: 519-527.

Shatz, A., 2004. VIOXX pain – a story of business and statistics errors. www.interconus.com/Vioxx_Pain.doc. [Accessed 25 March 2005]

Szalay, A., 1999. New Astronomy from the Sloan Digital Sky Survey to the National Virtual Observatory. tarkus.pha.jhu.edu/~szalay/powerpoint/euus.ppt .

Tukey, J. W., 1954. Unsolved problems of experimental statistics. *J. Amer. Statist. Assoc.* 49: 706- 731.

Weinberg, C. R., 2001. It's time to rehabilitate the P-value. *Epidemiology*, 12: 288-290.

Winch, R. F., and D. T. Campbell, 1969. Proof? No. Evidence? Yes. The Significance of Tests of Significance. *American Sociologist*, 4: 140-143.