STATISTICS AND THE SCIENTIFIC METHOD

BY

DAVID A. FREEDMAN

COMMENTS ON AND REACTIONS TO FREEDMAN

BY

STEPHEN E. FIENBERG

A REJOINDER TO FIENBERG'S COMMENTS

BY

DAVID A. FREEDMAN

TECHNICAL REPORT NO. 19

REVISED APRIL 1983

DEPARTMENT OF STATISTICS

UNIVERSITY OF CALIFORNIA, BERKELEY

RESEARCH PARTIALLY SUPPORTED BY

NATIONAL SCIENCE FOUNDATION GRANT MCS-80-02535

AND

NATIONAL SCIENCE FOUNDATION GRANT SES-80-08573

Table of Contents

C. A rejoinder to Fienberg's comments
by David A. Freedman

Statistics and the Scientific Method

by

David A. Freedman

Abstract. Regression models have not been so useful in the social sciences. In an attempt to see why, such models are contrasted with successful mathematical models in the natural sciences, including Kepler's three laws of motion for the planets.

Author's footnote. I would like to thank the following persons for their help with the essay--in one or two cases, rendered despite serious disagreement with some of the contents: Persi Diaconis, Morris Eaton, John Heilbron, David Hopelain, David Lane, Thomas Rothenberg, Richard Sutch, Amos Tversky.

1. Introduction

In the social sciences today, much effort is spent running regressions. As far as I can see, the return on this intellectual investment has been meager.[1] If so, this raises two very difficult questions: Why is it that regression models have had so little success? And why are they so popular? As a partial answer, it may help to look at some mathematical models that have succeeded. I have chosen a few, largely on the bais of familiarity. They are not regression models. They happen to belong to the natural sciences rather than the social sciences.

The comparison between typical regression models in the social sciences and a select handful of the great natural science models may seem unfair or even irrelevant. Hence it requires some justification. The point is not to demonstrate the superiority of natural science: there are plenty of bad models in biology or physics, and much good work in economics or psychology. The idea is rather to examine some highly successful mathematical models for natural phenomena, in order to understand the sources of their strength. The history will be interesting in its own right, and may--or may not--shed some light on the present issue: why have mathematical methods succeeded so well in the physical sciences, and not so well in the social sciences?

My view is that regression models are in vogue in the social sciences largely because of the success of other kinds of mathematical models in the natural sciences. However, these notions of "model" seem so different to me that covering both by the same term may be a source of real confusion. Saying clearly what the differences are is not so simple , but the following headings seem relevant: natural law, originality, depth, prediction, stochastics, measurement, and replication. In brief, my points will be as follows:

On natural law. The great models in the natural sciences result from a search for truth, namely the laws governing the phenomena under investigation. Such a model expresses in definite mathematical form the

---

[1]For some evidence to support this contention see Freedman (1981) or Freedman-Rothenberg-Sutch (1983), and the references cited there, especially Christ (1975) and Zarnowitz (1979). For similar critiques in other contexts, see Hendry (1970), Karlin (1979), or Kiefer (1979). For related discussions on handling experimental data, see Hausman and Wise (1982), or Zeisel (1982). For complementary discussions, see Lucas and Sargent (1978) or Sims (1980, 1982).

investigator's idea as to how the phenomenon really behaves. By comparison, social scientists who do regressions are usually fitting curves: they are modeling the data: see Neyman (1977). It should come as no surprise when such curves lose their fit after a short time. An investigator who is not looking for the truth will not find it.

On originality. The great models are brilliantly original; no two are alike. Each one was discovered through an act of intellectual creativity of high order. Regression models, by comparison, are right off the shelf.

On depth. In the natural sciences, the great models reflect profound insight, and show real intellectual elegance. Typically the model succeeds in explaining some very diverse set of facts on the basis of a few simple axioms--including facts not available when the model was developed. However, much hard thinking is needed to get from the axioms defining the model to the conclusions about the world. With regression models, there is seldom much real difference between the inputs and the outputs.

On Prediction. Models in the natural sciences are expected to make sharp and nontrivial predictions about the future, predictions which can be verified by direct observation. Some models even give a large measure of control over the phenomena. Models which fail such tests eventually get scrapped. For a brilliant, if sometimes perverse account, see Kuhn (1970). In the social sciences, regression models are seldom exposed to this kind of risk. But models which are not subjected to rigorous empirical testing cannot be expected to have much empirical content. And the standard statistical tests are usually irrelevant, as will be argued below: the main reason is that the tests themselves make assumptions which are not tenable.

On stochastics. There are great stochastic models in the natural sciences, and a lot of attention goes into testing their basic assumptions. After all, it is the assumptions which define the model. Regression models too make quite strict assumptions, explicitly or implicitly, about the stochastic nature of the world. In most social-science applications, these assumptions do not hold water. Neither do the resulting models.

On measurement. The great models in the natural sciences involve variables which have been defined clearly and measured carefully. Such claims can be made for few regression models in the social sciences. Good models are hard to build on the basis of bad data.

On replication. In the natural sciences, the crucial experiments to validate important models get replicated as a matter of course. In the social sciences, few regressions get replicated. This comes back to the point that few social science regression models are exposed to rigorous empirical testing.

I hope this paper will not be construed as making invidious comparisons between the social sciences and the natural sciences: there are plenty of bad models in the natural sciences. It is not an attack on the social sciences, or even on the use of quantitative methods in the social sciences. Indeed, statistics are clearly very useful in descriptive work; so are survey methods. My critique is much narrower: the focus is on regression models and variants like structural equation models. The criticisms seem to apply to many of the papers in the present volume; for specifics, see Appendices I and II below.

2. Some models in the natural sciences

In 1609, Kepler published his laws of planetary motion: the first law, for instance, is that planets move in elliptical orbits with the sun at one focus. These laws constitute a "model" for the solar system.[1] I use quotation marks because Kepler viewed his laws not as a model, but as a description of how the planets actually moved in space. The story of Kepler will be discussed in more detail in section 7, to show his attitude toward natural law, as well as the originality, depth and power of his discovery.

Kepler gave a brilliantly simple description of a very complicated set of planetary motions. Twenty years later, the "Rudolphine Tables" were published; these used Kepler's laws to predict the positions of the planets in the sky, and were an immediate practical success. The earlier Ptolemaic and Copernican tables were often in error by up to 5 degrees in predicting planetary positions; the Keplerian tables reduced the error by a factor of 30, to below 10 minutes of arc.[2] Still, the mechanism behind Kepler's laws was unknown. Half a century later, Newton provided the mechanism, whose centerpiece was the law of gravity. Newton's theory looks very different from the facts it explains. The law of gravity, for example, is that any two bodies attract each other with a force proportional to the product of their masses, divided by the square of the distance between them. This does not seem to have much to do with an ellipse. However, Newtonian mechanics implies Kepler's laws (including the elliptical orbits) by a strict mathematical

---

[1] The language of "laws" and "models" does not cohere so well. For present purposes, I will construe a "law" as one of the axioms defining a model. However, such axioms are also truths about the world--although not necessarily self-evident ones. This is an old-fashioned view. Friedman (1953) argues that useful theory about the world can be developed from axioms which are false as statements about the world. But this is only another proof of how clever Milton Friedman really is. Other investigators would be well advised not to accept the handicap imposed by false assumptions.

[2] See Gingerich (1971).

argument. To bring off this argument, Newton was obliged to invent large parts of what is now called "the calculus." In this example, the inputs to the model are very different from the outputs, and a lot of hard thinking is needed to get from the axioms to the conclusions.

The word "gravity" is so much a part of our vocabularies now that the idea may be difficult to appreciate. But in its time it was brilliantly original. In fact, it was almost unthinkable for many of Newton's contemporaries, and rather hard even for Newton himself, because it involved the idea of action at a distance.

Newtonian mechanics has come to dominate our view of the physical world, and has given us substantial mastery over that world. Guided by Newton's theory, investigators can discover new planets by the anomalies created in the orbits of the old ones;[1] and in this century, astronauts land on the moon. The great models have empirical consequences.

Of course, Newtonian mechanics is not the last word. Einstein discovered that Newton's laws were only a first approximation, applicable to relatively small masses moving at negligible fractions of the speed of light. But Newton's mysterious force of gravity turns out to be a consequence of the very geometry of space--and the tensor calculus.

This brief history may indicate some of the originality, diversity, and depth of the great mathematical models in physics, as well as their

---

[1] See Grosser (1962). The existence and position of the planet Neptune was deduced by Leverrier, from a study of anomalies in the motion of Uranus. Neptune was first observed (through a telescope) by Galle. Two quotes:

> It is impossible to satisfy the observations of Uranus without introducing the action of a new Planet, thus far unknown... Here are the elements of the orbit which I assign to the body...
>
> Leverrier to Galle, September 18, 1846

> The Planet whose position you have pointed out actually exists...
>
> Galle to Leverrier, September 25, 1846

explanatory power. There are similar stories in biology. In 1865, Gregor Mendel proposed a statistical model to explain the mechanism of heredity.[1] Seed color in peas, to take a famous example, was postulated to depend on a pair of "entities," one from each parent. The transmission of these entities from one generation to the next obeyed carefully-formulated probabilistic laws, which successfully explained a maze of empirical data. In this century, the physiological basis for the model has been thoroughly explored, and most of it can now be photographed under an electron microscope. A cell divides: and a chromosome is either on one side of the line or the other, with a 50-50 chance. There is even a successful model for the structure of the chromosome itself: the Watson-Crick double helix. Again, it is hard to over-estimate the degree of understanding and control that Mendelian genetics gives us. In the third world, for example, millions of people live on the "miracle rice" developed at the International Rice Research Institute, using Mendel's principles. The genetic model too captured the truth. It was strikingly original and very deep. Its conclusions are quite different from its assumptions. And it has great explanatory power.

[1] Some references are Judson (1979), Rosenberg (1979), Freedman-Purves-Pisani (1978, Chapter 25). An interesting sidelight is that Mendel's theory was overlooked for nearly half a century, and then rediscovered independently by some ordinary, working scientists. The triumph is of method, as well as genius.

3. Regression models in the social sciences

In social-science regression analysis, the approach is very different. Usually the idea is to fit a curve to the data, rather than figuring out the process which generated the data. As a matter of fact, investigators often talk about "modeling the data." This is almost perverse: surely the object is to model the phenomenon, and the data are interesting only because they contain information about that phenomenon. Whatever it is that most social-science investigators are doing when they construct regression models, discovering natural laws does not seem to be uppermost in their minds.

The next point to make is that most statistical models in the social sciences bear a strong resemblance to one another. Investigators have the normal curve and regression, the multinomial distribution and logits, time series and autoregression, used over and over again. Indeed, the choice of statistical model is usually governed not by logic of the situation but by the layout of the data files in the computer.

What about the use of social-science regression models to make predictions? In such models, there is very little difference between the inputs and the outputs. Investigators postulate a linear relationship between the dependent variable and some explanatory variables, which may even include thinly disguised versions of the dependent variable itself. They fit the model by least squares, and theorize retrospectively about the coefficients. (If the coefficients come out wrong, they respecify the equations.) Sometimes they use the model to do simulations--for a world which will be never observed. And that is where such modeling exercises usually seem to stop. This kind of work is unlikely to lead to any real advances in the understanding of social phenomena.

4. Stochastics

Even off-the-shelf statistical models make quite strong assumptions about the processes generating the data, and are likely to produce nonsense if these assumptions fail. However, it is rare indeed to find an investigator who takes these assumptions seriously--or who backs off when confronted with the fact that the assumptions are clearly violated by the phenomenon under analysis. Specific examples will be discussed in the appendices to this paper.

In Mendelian genetics, by contrast, the stochastic assumptions are taken very seriously indeed. Maybe that is one reason why statistical models work so much better in genetics than in the social sciences.

5. Measurement

In the natural sciences much importance is attached to careful measurement work. The key variables get defined very cleanly. A lot of ingenuity, and years of patient work, go into determining the fundamental parameters of physical models: these parameters get connected to observable quantities. And the investigators often manage to design the experiments so that measurement error is held to a very low level indeed.

I will cite two important physical constants that seem almost impossible to measure. The first is the charge on the electron. How can you measure something that small? Millikan did it, using a drop of oil. Or what about the speed of light, which is practically infinite? Michelson is famous because he measured it. See Franklin (1980), Holton (1978), Livingston (1973), and Swenson (1972).

With social-science regression models, it is altogether different. Few investigators do careful measurement work. Instead, they factor-analyze questionnaires. If the scale is not reliable enough, they just add a few more items. Such techniques are not serious, by comparison with the sort of measurement work done in the natural sciences.[1]

[1] For a well-known critical review of the accuracy of economic data, see Morganstern (1963).

6. Replication

In the natural sciences, most of the crucial experiments and observations are replicated not just once but dozens and hundreds of times. The really classic ones even get incorporated into high school and college lab courses. Replication is another characteristic feature of the natural sciences.

In the social sciences, by comparison, few studies are done more than once. But replication is a relevant idea, even for ingestigators fitting regression models to observational data. After running the regressions, the investigators can collect some more data and see if the equations survive. Econometricians are almost forced into this, because year by year new data comes pouring in. And the half-life of a coefficient in an econometric model is measured in months, not years. I do not wish to be unkind, but the contrast is stark. Astronomers still use Kepler's model. He got it right.

7. Some responses

There are three standard objections to my line of argument:

- Natural scientists can do controlled experiments; social scientists cannot.
- Social scientists deal with more complicated problems than natural scientists.
- Social science models should not be judged so harshly, because the investigators are only doing data analysis.

My reactions are as follows.

On controlled experiments. Astronomy, for example, is mainly observational. And learning theorists in psychology do a lot of experimentation on human subjects. Controlled experiments are very useful, but not crucial.

On complexity. Some problems in the natural sciences now look very clean and simple, but only because of the analytical work that has been done. To appreciate this point, imagine trying to figure out the orbit of Mars for yourself. You go out on a clear night, look up into the sky, and see thousands of points of light. Which one is Mars? To begin closer to the beginning, which ones are the planets and which the stars? Continuing to watch for several hours might only confuse matters further: for the pattern of the stars will gradually change as the night wears on. Even recognizing this change depends on prior knowledge; for it is hard to see the shifting pattern of the stars without using the constellations.

It took many thousands of years of patient study before astronomers were able to recognize the existence of the planets ("planet" derives from a Greek word meaning "wanderer"), or stars which moved against the background created by the constellations of fixed stars. And even after astronomers recognized the fixed stars and the planets, they needed a theory to enable them to measure positions in the sky. Such a theory was developed by the Greeks, who imagined the stars as fixed to a heavenly sphere, with the earth at its center. This sphere revolved once a day. The sun moved along the sphere in a path called the ecliptic, completing one circuit every year.[1] And each planet moved along the heavenly sphere in its appointed orbit. This theory now seems quaint or even absurd. But in its time it was a brilliant advance, for it permitted astronomers to record the positions of the planets by separating their movement against the stars from the apparent daily rotation of the heavens. So when Brahe and Kepler went to work, they could draw on centuries of skillful observation and theorizing. The main elements of the problem had been identified, and some techniques for making relevant measurements had been well developed.

One conclusion from this history: before scientists can make good measurements, they need to have a clear idea of what it is they are going to measure. In other words, good measurements often depend on good theory. A second conclusion: insofar as the problem of the planetary motions now looks clean and simple, that is the result of many centuries of hard work. A final conclusion: the social sciences may well be at the pre-Keplerian stage of investigation--the equivalent of figuring out which are the planets and which the fixed stars. If so, using sophisticated analytical techniques like regression is bound to add to the confusion. The problem is to define the basic variables, to figure out ways of measuring them, to perceive the main empirical regularities. Estimating coefficients by least squares before

---

[1] The Signs of the Zodiac are the constellations through which the sun moves, along the ecliptic. Today, the ecliptic is defined as the apparent path of the sun against the fixed stars.

the basic variables have been understood is like using a scalpel to clear a path through the jungle.

On data analysis. "Data analysis" is one of the current slogans in statistics, but it is not a fair description of the kind of statistical work usually done in the social sciences. Data analysts emphasize close inspection of the numbers, displayed graphically. Conventional methods involve histograms, scatter diagrams, and residual plots. The more radical ones involve the stem-and-leaf, hinges, and hanging root-a-grams. But all data analysts draw pictures. In the social sciences, graphical analyses of the data are quite rare. Data analysts work very hard to develop models for their data, and run diagnostics to see if the models are sensible. They spend endless hours dealing with outliers, or changes in the relationships from one region to another, or non-linearities. In the social sciences, this kind of analysis too is quite rare. Data analysts seldom make conventional statistical tests, like t, χ^2, or F, because such tests are valid only under severely restrictive mathematical assumptions. There are few statistical papers in the social sciences without a battery of such tests.[1]

The conclusion seems inescapable. In general, social scientists who run regressions are not doing data analysis. Instead, they are mechanically applying regression models in situations where the assumptions do not hold. The computer outputs--the parameter estimates, the standard errors, the t-tests--are usually devoid of scientific meaning. Rather than facing up to this issue, however, the investigators just label the outputs as "descriptive

[1]The paper by Duncan in this volume is an example of what can be done by using simple, and appropriate, statistical techniques for looking at data.

statistics." This is a swindle. If the assumptions of the regression model do not hold, the computer outputs do not describe anything: they are mere numerical artifacts.

8. The story of Kepler[1]

One of the great mysteries of the ancient world was how to explain the motions of the planets, for their apparent paths in the sky were extremely complicated. The problem was made even more difficult by the doctrines of Aristotle and Plato, who held that the earth was stationary at the center of the universe, and furthermore that the planets went around the earth in perfect circles at constant speed--the rule of absolute motion. By ignoring some of these obstacles, the Greek astronomer Ptolemy (second century AD) perfected an astronomical system which explained the apparent paths of the planets reasonably well.

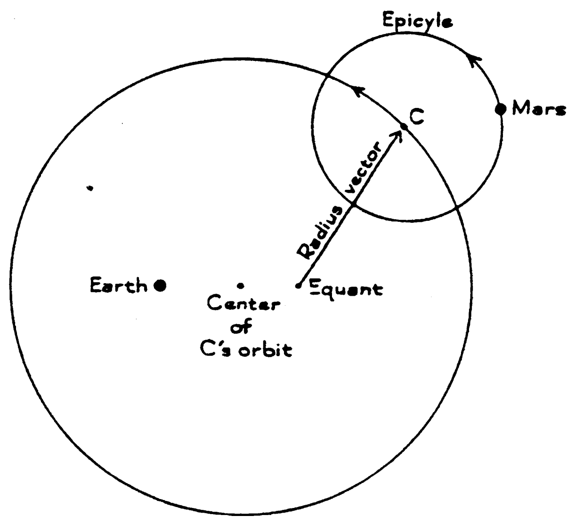
Ptolemy had to depart from Aristotelian physics in several ways. Perhaps most important was his use of the epicycle and equant. Take Mars as an example. Ptolemy postulated that this planet moved at constant speed in a circle called the epicycle (figure 1). The center C of this epicycle moved in a larger circle (the eccentric) around the earth. However, Ptolemy placed the center of this larger circle in space off the earth, and he did not make C move at constant speed around this center. Instead, he introduced another point called the equant. In his scheme, the radius vector joining C to the equant swept out equal angles in equal amounts of time. As we know today, Ptolemy was forced into these complications because Mars goes around the sun, not the earth, and does not follow a circular orbit.

Using a half-dozen equants and two dozen epicycles, Ptolemic astronomers were able to account fairly well for most of the observed planetary motions.

---

[1] I am grateful to John Heilbron, Professor of History, UC Berkeley, for his help with this section. Needless to say, I am responsible for all the faults which remain. Some references are: Butterfield (1949), Dreyer (1953), Koestler (1963), Koyré (1973), Kuhn (1969), Russell (1964), Wilson (1968, 1972).

Figure 1. Ptolemy's theory. Mars moves at constant speed in an auxiliary circle called an <u>epicycle</u>. The center C of the epicycle goes around the earth in a larger circle called the <u>eccentric</u>. The center of the eccentric is located somewhere in space, off the earth. There is a further imaginary point called the <u>equant</u>, placed as far in one direction from the center of the deferent as the earth is in the other. The radius vector joining C to the equant covers equal angles in equal amounts of time.



However, there were always some discrepancies between the predictions and the observations. These discrepancies were especially noticeable for the motions in latitude, that is, north or south of the ecliptic. The observed changes in the distances between the planets and the earth, indicated by changes in their apparent brightness, also presented a serious problem for the theory. For these reasons, and perhaps because of the artificiality of their constructions, Ptolemaic astronomers tended to regard their system as a device for predicting the apparent positions of the planets, and not as a true description of paths in space.

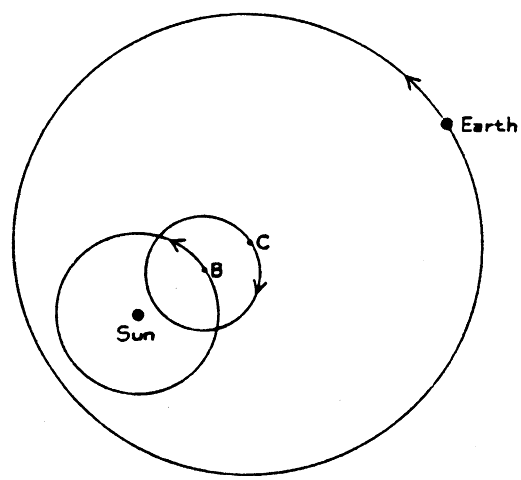
In the Middle Ages, however, many astronomers took Ptolemy very literally indeed. They equipped the heavens with crystalline spheres revolving on spheres to carry the planets along the epicycles. There were endless arguments about how many spheres were needed, how they fitted together, whether motion could be transferred from one to another, and how many angels were needed to keep the spheres going. This state of affairs went on until the time of Copernicus (1473-1543).

As most high school texts explain, Copernicus rebelled against the Ptolemaic system, and taught that the earth went around the sun. However, it was not that simple. What Copernicus seems mainly to have rebelled against was Ptolemy's use of equants, regarding them as a violation of Aristotle's rule of absolute motion. As Copernicus himself put it, according to Koestler (1963, p. 145)

> Having become aware of these defects, I often considered whether there could perhaps be found a more reasonable arrangement of circles...in which everything would move uniformly about its proper center, as the rule of absolute motion requires.

After many years of labor, Copernicus found such a system. Part of it --the orbit of the earth--is illustrated in figure 2. As this shows, Copernicus imagined the earth to move at constant speed in a perfect circle around the point C. This imaginary point in turn moves at constant speed in a perfect circle around the point B. Copernicus is not done yet, for B also moves, at constant speed in a perfect circle around the sun.

Figure 2. The Copernican theory of the orbit of the earth. The earth moves at constant speed in a perfect circle around the center C, completing one orbit each year. The center C moves at constant speed in a perfect circle around the point B, completing one orbit every 3434 years. The point B moves at constant speed in a perfect circle around the sun, completing its orbit in about 53,000 years. This figure is adapted from Dreyer (1953, p. 332).



The other planets all move on similar nests of circles, ultimately centered on the imaginary point C, the moving center of the earth's orbit. Copernicus says with evident pleasure (Dreyer, 1953, p. 343).

> Thus Mercury runs in all on seven circles, Venus on five, the earth on three, and round it the moon on four, lastly Mars, Jupiter and Saturn on five each. Thus altogether thirty-four circles suffice to explain the whole construction of the world and the whole dance of the planets.

Koestler (1963, p. 572) has complained that when Copernicus got down to brass tacks, he actually needed forty-eight circles. Whether this is true or not, the verdict of Butterfield (1949, p. 30) seems right:

> When you go down, so to speak, for the third time, long after you have forgotten everything else in this lecture,

there will still float before your eyes that hazy vision, that fantasia of circles and spheres which is the trademark of Copernicus.

Tycho Brahe (1546-1601)
From the Wolff-Leavenworth Collection
George Arents Research Library, Syracuse University

The next figure in the story is Tycho Brahe (1546-1601). Brahe recognized that in order to solve the mystery of the planetary motions, thousands of very accurate observations of their positions would be needed, over a period of many years. He devoted his life to this task, and his achievement marks the beginning of modern observational astronomy. This approach was strikingly different from that of Copernicus, who so far as is known made only twenty-seven observations on the planets.[1] Eventually,

[1] And in fact, many of the Copernican epicycles were needed to get the theory to conform to certain observations made by the Arab astronomers--observations which have since been shown to be wrong. It is hard to build a good model on the basis of bad data.

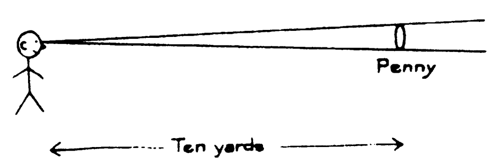
Brahe tried to use his observations to piece together a theory of the solar system. He made all the planets except the earth revolve around the sun--while the sun went around the earth.

Johannes Kepler (1571-1630)
From the Wolff-Leavenworth Collection
George Arents Research Library, Syracuse University

Now I introduce Johannes Kepler (1571-1630). Kepler went to Prague in 1600 to join Brahe, who had just moved there from Denmark. Kepler spent many years trying to fit Brahe's data on Mars by means of circular orbits. But even with the best such orbit, the theoretical position of Mars on a certain date proved to be eight minutes of arc away from the position observed by Brahe. Now eight minutes of arc is a very small angle. It is the apparent size of a penny held at a distance of ten yards from the observer. As in figure 3, a penny held ten yards from the eye covers both the actual position, and the theoretical position of Mars computed from the circular orbit. However, Kepler knew that Brahe was very unlikely to have made a measurement error that large. (Brahe's observations of the planetary positions were accurate to within four minutes of arc or so.) So this small difference forced Kepler to break with the tradition of circular orbits, and in time led him to discover the true shape of the planetary orbits. As Kepler said (Dreyer, 1953, p. 385),

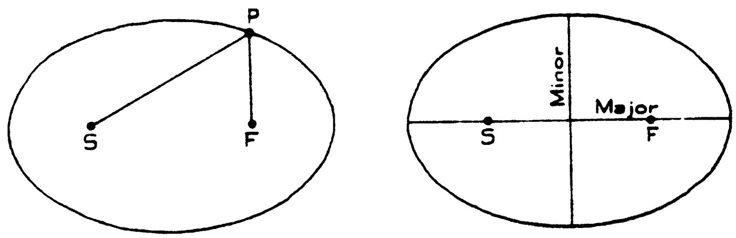
> To us Divine goodness has given a most diligent observer in Tycho Brahe, and it is therefore right that we should with a grateful mind make use of this gift to find the true celestial motions.

Figure 3. Eight minutes of arc is the apparent size of a penny held ten yards from the eye. The figure is not drawn to scale, and the angle shown is about 100 minutes of arc.



What Kepler found was that the planets really move in elliptical orbits. As figure 4 shows, an ellipse can be drawn by tying a loop of string around two nails, tucking a pencil into the loop, and tracing a curve with the string held taut. Each nail is a focus of the ellipse. The <u>major axis</u> goes through the foci of the ellipse; the minor axis passes half-way between the foci, being perpendicular to the major axis and somewhat shorter (figure 4). The ellipse was discovered by mathematicians in ancient Greece, while they were investigating cones ; before Kepler, there was absolutely nothing to connect this curve to the paths of the planets in the sky.

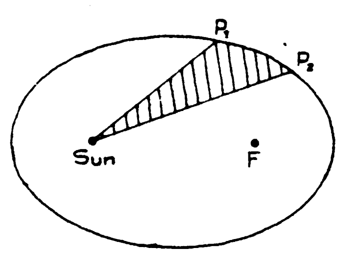
Figure 4. The ellipse. The points S and F are the foci of an ellipse; as P moves around the ellipse, the sum of the distances PS + PF remains constant.



Kepler's first two laws can be stated as follows:

- Each planet (including the earth) moves around the sun in an elliptical orbit, with the sun at one focus.
- A planet moves in its orbit at varying speeds, in such a way that if it is joined to the sun by an imaginary line, this line will sweep out equal areas in equal times (figure 5).[1]

Figure 5. Kepler's first two laws. Mars moves in an elliptical orbit with the sun at one focus. The shaded area is swept out by the radius vector as the planet moves from P_2 to P_1. Equal areas are swept out in equal times.



[1] The first law governs the shape of the orbit; the second, the rate of motion along the orbit. The third law is that the square of the period of rotation is proportional to the cube of the average distance from the sun.

A theory assuming circular orbit for Mars causes only a small discrepancy between predictions and observations, because the elliptical orbit of Mars is in fact nearly circular: the ratio of the minor axis to the major is 199 to 200, compared to a ratio of one for a perfect circle.[1]

Kepler's attitude belongs to the great age of science. He was looking for the true path of Mars in space, rather than a device for computing apparent positions. Indeed, he seems to have been the first astronomer to define the problem that way. He was willing to spend years on the search and in the end go against centuries-old physical doctrine, because he was certain that the true shape of the path would be accessible to the human mind. And his theory had to explain all of Brahe's observations on Mars. Not some of them, or most of them, all of them. He was guided by faith that some natural law would be found to govern the shape of the orbit, if he could only see what that shape was.

I sometimes have a nightmare about Kepler. Suppose a few of us were transported back in time to the year 1600, and were invited by the Emperor Rudolph II to set up an Imperial Department of Statistics in the court at Prague. Despairing of those circular orbits, Kepler enrolls in our department. We teach him the general linear model, least squares, dummy variables, everything. He goes back to work, fits the best circular orbit for Mars by least squares, puts in a dummy variable for the exceptional observation--and publishes. And that's the end, right there in Prague at the beginning of the 17th century.

---

[1] Using an eccentric circle, with speed regulated by an equant, makes the circular motion even closer to the elliptical one--down to the eight minutes of arc.

Appendix I. Discussion of the Joreskog-Sorbom Paper in this Volume

My major criticism of this paper has to do with the underlying model, which the authors do not spell out in anything like enough detail. Once stated clearly, the main assumptions may seem quite implausible. If so, the analysis is without adequate foundation. Let me now state the assumptions explicitly, following the Joreskog-Sorbom notation as closely as possible. We have children indexed by c, each child belonging to some cohort g; tests i are administered at times t. Thus, y^g_{cit} denotes the score obtained by child c in cohort g on test i at time t. There are four cohorts, three tests, and three times.

The model is summarized in two equations:

$$(1) \qquad y^g_{cit} = \nu^g_{it} + \lambda^g_{it}\eta^g_{ct} + \varepsilon^g_{cit}$$

$$(2) \qquad \eta^g_{ct} = \alpha^g_t + \beta^g_t\eta^g_{ct-1} + \delta^g_{ct}$$

Here, η^g_{ct} is a latent variable, intended to represent child c's "intelligence" at time t. The coefficients ν^g_{it} and λ^g_{it} in (1) relate the scale of the latent variable η to the scale of the observable test scores y. The coefficients are assumed constant across children within cohorts. For the bulk of the paper, they are assumed constant across times and cohorts as well. Equation (2) states that η evolves in an autoregressive way. Again, the coefficients are assumed constant across children.

To finish specifying the model, it is necessary to spell out the assumptions governing the "errors" ε and δ in (1) and (2). I believe that the assumptions needed to justify the statistical manipulations are the following:

(3) Children are independent, and within cohorts identically distributed, i.e., the vectors

$$\{y^g_{cit}, \eta^g_{ct}, \varepsilon^g_{cit}, \delta^g_{ct}: \quad i,t = 1,2,3\}$$

are independent across c's, with a multivariate distribution dependent only on g.

(4) The ε's are independent of the δ's.

(5) The ε's are independent across tests; however, dependence is allowed within child and test across times: so ε^g_{cis} and ε^g_{cit} have non-zero covariance which depends on g, i, s and t, but not on c.

(6) The δ's are independent across time; the variance of δ^g_{ct} is allowed to depend on g and t, but not c.

(7) All variables are jointly gaussian -- that is, multivariate normal.

(8) The ε's and δ's all have mean 0.

These assumptions are not subject to direct empirical verification, mainly because η is unobservable. However, they are inherently implausible. To begin at the beginning, why should a child's intelligence be representable as a single number? Many specialists in factor-analytic theories of intelligence would reject this idea, leaving η without much appeal as a construct. And whatever η may be, why would it obey equation (2), especially with all the other assumptions on the errors?

The major assumption in the paper is perhaps (3), that children are independent, and identically distributed within cohorts. Without independence, the statistical computations reported by Joreskog and Sorbom have little meaning. Of course, their computer package LISREL will do the arithmetic in any case. If the independence assumption is wrong, however, the estimates may be biased--or meaningless.

Joreskog and Sorbom do not discuss the sample design. With e.g. a conventional cluster sample, some of the children in the study must have known each other, played together, gone to school together, or even come from the same family. Under such circumstances, independence is most unlikely. Ignoring these inter-child correlations biases the estimates. Building them into the model makes it under-identified.[1]

Next, consider the hypothesis in (3) that within cohorts, children are identically distributed. Casual empiricism and psychological doctrine alike suggest that different children evolve in different ways -- so the coefficients in (1) and (2) as well as the covariances of the errors may really depend on the child in question. Again, if the model is modified to allow this kind of person-to-person variation, it becomes under-identified. But without this modification, the parameters may lose their meaning.

For the sake of argument, let us set all this aside for a moment, and look at assumption (5), that the "measurement errors" ε in test scores are uncorrelated across tests. This assertion too is quite implausible. Suppose, for instance, that a child is depressed one year: this could easily lower all the test scores -- producing correlated ε's. Other sources of correlation are easy to imagine. Building inter-test correlations into the model makes it under-identified: leaving them out biases the estimates.

---

[1]Joreskog and Sorbom concede, in a footnote, that the independence "may be questionable."

Finally, consider assumption (7), that the variables are jointly gaussian. Like (3), this is a strong assumption, and it is not stated explicitly in the paper. But in the absence of this assumption, the log-likelihood χ^2-tests in the paper have no scientific foundation.

These log-likelihood tests are the principal mode of statistical analysis used by Joreskog and Sorbom, so they are worth considering in some detail. To begin with, such tests cannot give any absolute, overall check on the fit of the model to the data. The reason is that log-likelihood tests are always nested, with some general hypothesis H_{gen} and a more specific hypothesis H_{spec}. The log-likelihood test takes H_{gen} for granted, and asks whether the data are relatively more or less likely under H_{spec}. To spell out the mechanics a bit, let $L_\theta(x)$ be the likelihood of the (multivariate) data x if the vector of parameters is θ. The test statistic is

$$T = 2 \log \sup_{\theta \in H_{gen}} L_\theta(x) - 2 \log \sup_{\theta \in H_{spec}} L_\theta(x) .$$

And this is a measure of relative likelihood. It is a well accepted part of the statistical folklore that asymptotically, on the null hypothesis, T follows the chi-squared distribution. This can be proved rigorously when

- the likelihood function is smooth
- H_{gen} and H_{spec} are open subsets of Euclidean spaces
- the parameter vector corresponding to the null hypothesis is an interior point of H_{spec}.

See Cox and Hinkley (1974, pp 331 and 355); or Kendall and Stuart (1961, Vol. II, p. 231).

Joreskog and Sorbom do not cite any specific theorems whose assumptions are satisfied in their application. Even if T is asymptotically chi-squared on the basis of some unspecified theorem, there still are some formidable technical problems. Each data point (the test scores for one child) is nine-dimensional: 3 tests by 3 times. So there are 9 means, 9 variances, and $\frac{1}{2} \cdot 8 \cdot 9 = 36$ covariances to estimate for each cohort, from about 200 children per cohort. This only works out to 4 data points per parameter. The asymptotic theory, with sample sizes going to infinity, may not offer any very reliable guide to the sampling distribution of T in the present application. Finally, the joint distribution of scores can hardly be exactly jointly gaussian. Minor departures from normality may have a major impact on the distribution of T.

Now let me waive all such technical objections to the log-likelihood tests. The substantive interpretation in the Joreskog-Sorbom paper still presents real difficulties. For instance, consider testing the hypothesis H_{spec} that the "factor loadings" ν_{it}^{g} and λ_{it}^{g} in (1) are constant across cohorts g and occasions t. According to Joreskog and Sorbom,

> The test gives $\chi^2 = 198.6$ with 104 degrees of freedom. Although the outcome is significant at conventional levels of significance the residuals are generally small and in view of the sensitivity of χ^2 the fit must be regarded as a reasonably good one. We shall therefore consider the measurement model to be invariant over occasions as well as over cohorts. [p. 14, Draft of Feb/81]

In fact, the p-value of this test is 10^{-7}: i.e., if H_{spec} were right, and the asymptotic theory were valid, we would have only one chance in ten million of getting a χ^2-value as big as or bigger than the one reported by Joreskog and Sorbom. The only reasonable conclusion is that H_{spec} is wrong: the factor loadings depend on cohort and occasion.

Social scientists often argue that with large sample sizes, the χ^2-test is bound to reject; so χ^2-points per degree of freedom is used as a measure of goodness of fit. I am somewhat sympathetic to this argument, when the model has any prior claim on our sympathy. However, the Joreskog-Sorbom model is not simple, elegant and useful. It does not have any theoretical justification. The only possible defence is empirical: it fits the facts. However as the χ^2-test shows, H_{spec} does not fit the facts--even if we take H_{gen} for granted.

I now wish to sum up the statistical part of the discussion. The model used by Joreskog and Sorbom is very incompletely specified in their paper. When completely specified, its major assumptions turn out to be untestable and in important ways implausible. If these assumptions are wrong, the parameter estimates and associated standard errors may be severely biased -- and perhaps meaningless. Even if the assumptions are right, presently available statistical theory does not justify the paper's principal mode of analysis: log-likelihood χ^2 tests. Clearly, Joreskog and Sorbom have gone off on their own. A final summary point: waiving all technical questions about these χ^2-tests, they flatly contradict the specific hypothesis of the "measurement model," namely, constant factor loadings: the p-value is about 10^{-7}. Thus, the entire analytic apparatus used by Joreskog and Sorbom is faulty, from top to bottom.

Having summarized the statistical discussion, I want to make a non-statistical comment. Nothing seems to ride on the statistical analysis, because all the inferences are about unobservables: the parameters governing the postulated stochastic distribution of the unobservable variable η . The paper does not say anything about any quantity that could be observed. Not only is the technical foundation of the analysis somewhat shaky, but its scientific value remains questionable--until the unobservables are tied into a theory which is open to empirical testing.

Appendix II. Discussion of the Markus-Converse Paper in this Volume.

The main intuitive idea seems to be that partisanship develops in a dynamic way. The basic equation is a first-order autoregression:

(1) $$S_t = \alpha + \beta S_{t-1} + u_t$$

where S_t measures partisanship in period t and u_t is a disturbance.

The justification for (1) is that under certain circumstances, a first-order autogression will have some qualitative properties in common with the data. However, many other equations would satisfy the same criteria. One such is

(2) $$S_t = \alpha + \sum_n \beta^n S_{t-n} + u_t$$

Another is

(3) $$S_t = \alpha(S_{t-1})^\beta \; u_t$$

These alternative specifications have quite different dynamical properties from (1), and might lead to different substantive conclusions. However, Markus and Converse make no effort to justify the choice of specification (1), rather than (2) or (3) or the infinite number of other possibilities. Indeed, in the present state of the art, it is very unlikely that any such effort could succeed. The theory is too crude to make such discriminations, and so are the data. The authors concede this point, a bit obliquely, in their footnote 3.

Period effects are crucial to the argument in the paper. These are incorporated into (1) by allowing α to depend on t:

(4) $$S_t = \alpha_t + \beta S_{t-1} + u_t$$

At time t, data is available for many different cohorts, which will be indexed by i. So the model becomes

(5) $$S_{i,t} = \alpha_t + \beta S_{i,t-1} + u_{i,t}$$

Here, $S_{i,t}$ is the average partisanship score of cohort i at time t.

The parameters in equation (5) are

- the period effects α_t
- the coefficient β --"stability"

These parameters are unknown, and must be estimated from the data. No justification is given for the functional form in (5). Indeed, the paper does not derive the autoregressive model from theory, or test it against the data: instead, the model is simply assumed.

In the paper, almost nothing is said about the u's. However, their behavior is crucial to the statistical argument, so a close look is in order. It will be helpful to say clearly what the u's aren't and this involves some preliminary discussion. The parameters α_t and β can be estimated by a procedure called "least squares;" estimates will be denoted by corresponding roman letters a_t and b. These estimates are found by minimizing the sum of squares

(6) $$\sum_{i,t}(S_{i,t} - a_t - bS_{i,t-1})^2$$

Once a_t, b have been computed, $S_{i,t}$ can be "predicted" from the independent variables. The predicted value is denoted with a hat:

$$\hat{S}_{i,t} = a_t + bS_{i,t-1} \tag{7}$$

There is a discrepancy between the observed value $S_{i,t}$ and the predicted value $\hat{S}_{i,t}$. This discrepancy is called a residual, and denoted $\hat{u}_{i,t}$: so

$$S_{i,t} = a_t + bS_{i,t-1} + \hat{u}_{i,t} \tag{8}$$

Equation (8) involves perfectly definite, tangible numbers, which can all be computed from the data. I will call this equation the computer model for the data. The computer model is different from the stochastic model (5). In particular, notice the hats on the residuals in (8). This is to signal that the residuals in the computer model are different from the disturbance terms in the stochastic model (5).

The arithmetic behind (8), namely the minimization of the sum of squares (6), can always be done. A computer neither knows nor cares about the stochastic model (5). However, with some assumptions about the u's, minimizing (6) gives very sensible parameter estimates.[1] With other assumptions, the same arithmetic leads to complete nonsense. Furthermore, the arithmetic does not check itself to make sure it applies. The rule is caveat emptor. In order for the regression to make sense, something must be assumed about the stochastic disturbance terms $u_{i,t}$ in (5). Unless these assumptions are made explicity, the stochastic model must be regarded as incompletely specified. The stochastic model in the Markus-Converse paper is incompletely specified. So are the stochastic models in all the other papers presented at the conference.

---

[1] Markus and Converse use a more complex estimation procedure, designed to compensate for measurement errors in the data. I will focus on ordinary least squares, to simplify the exposition. For Markus and Converse, different cohorts may be in the study for different periods.

There is a standard set of assumptions to make about the stochastic disturbance terms $u_{i,t}$, which justify the least squares estimation procedures. In statistical jargon, these assumptions can be stated as follows: the errors $u_{i,t}$ are independent of one another, and are identically distributed with mean 0. This translates into quite a definite story: a story not about the data, but about the mechanism which generated the data.

This mechanism is assumed to be like the following hypothetical procedure, in which:

- The variables $S_{i,t}$ are public--in the data base.
- The parameters α_t, β are hidden, not known to the investigator.
- There is still another hidden object: a box of tickets, each ticket bearing a number; these numbers average out to 0.

I do not wish to seem patronizing, but to make the stochastic assumptions as vivid as possible, I will introduce a fictitious character called "the MC" (for master of ceremonies). The MC generates the data base, one period at a time. Suppose we have gotten through period t-1. Focus on one cohort, say cohort #1. Now the average partisanship score for this cohort is period t-1 has some value, $S_{1,t-1}$. To generate $S_{1,t}$, the MC draws a ticket at random from the box, and makes a note of the number on it. Then the ticket goes back in the box, for the future use, and $S_{1,t}$ is generated by the rule

$$S_{1,t} = \alpha_t + \beta S_{1,t-1} + u_{1,t}$$

Exactly the same procedure is followed for the other cohorts: the errors

$u_{i,t}$ are always drawn (at random, with replacement) from the same box: so they are independent of one another, and are identically distributed. In particular, the u's show no trend or pattern. These are the crucial assumptions about the mechanism for generating the data.

The MC and the box of tickets are purely fictitious--as are the parameters α_t, β. All the investigator gets to see are the data: the $S_{i,t}$. If the process which generated the data is like the one just described, least squares is a good way to estimate the parameters α_t, β. If the assumptions are wrong, a computer package can still be used to generate the least-squares estimates a_t,b. But these estimates may be biased, e.g., if there is serial correlation in the u's. Or the "standard errors" computed by the package may be off, if e.g. there is correlation in the u's across cohorts. Or there may not be any parameters around to estimate. So if the assumptions are wrong, least squares can be an intellectual disaster. Do the standard assumptions hold for the process in question? Markus and Converse do not face this issue squarely. Their footnote 12 acknowledges, again obliquely, that the assumptions are open to some question.[1]

The stochastic disturbance terms are unobservable. When you come right down to it, there is only one way to show that the stochastic disturbance terms satisfy the standard conditions: by argument a priori. This involves developing some theory to show where the disturbance terms come from.[2]

---

[1]Markus and Converse do not discuss the assumptions behind the Wiley estimation procedure used for β; nor the legitimacy of using that procedure for β followed by ols for the α's, with β held fast at its estimated value: this is not standard procedure. The ols estimate for β is .71: see footnote 11. The Wiley estimate is .95: the choice of statistical procedure matters a lot--and therefore assumptions are crucial to the analysis.

[2]For technical discussions of the impact of assumptions on performance of statistical procedures, see e.g. Breiman and Freedman (1983), Freedman (1983), Freedman and Peters (1983).

One response to this kind of criticism is crudely empirical: "But the model fits." So let us consider the fit. The model is autoregressive: partisanship this time is explained in part by partisanship last time. Autoregressions are expected to fit quite well, because most time series evolve smoothly: ordinarily, tomorrow will be rather like today. The Markus-Converse model includes almost 20 period effects, one for each election: even if some of the elections depart from the autoregression, the "period effects" should bring them back into line. The model is run on aggregate data; this too usually promotes very high correlations. Even so, the r^2 is only 0.7. Waiving all questions about the assumptions, surely the crude empirical conclusion is that equation (5) does not fit: lots of variance gets away.[1]

Another issue to consider is aggregation. Let us suppose, for the sake of argument, that the Markus-Converse model (5) is correct--for individuals:

$$S_{n,t} = \alpha_{n,t} + \beta_n S_{n,t-1} + u_{n,t} \tag{11}$$

Here, $S_{n,t}$ is the n'th person's partisanship score in period t. For the sake of argument, let us even stipulate to the standard assumptions about the $u_{n,t}$: they are independent of one another and have a common distribution with mean 0 and finite standard deviation.

The parameters in (11) were deliberately subscripted: person "n" has his or her own personal "period effect" $\alpha_{n,t}$ and "stability coefficient" β_n. If these parameters are in fact constant across people, then they can be estimated either from aggregate data or from panel data.

---

[1]A draft version of the paper gives the r^2 on microdata as 0.3.

But the basic assumption seems quite unlikely. To test it, separate regressions would have to be run for each person in the panel study; and the variability in the resulting parameter estimates would have to be analyzed. Markus and Converse do not test this basic assumption, or even mention it.[1] If this assumption is false, the specification (5) is likely to be quite wrong, and serious bias in the estimate of the stability coefficient is likely; see Freedman (1981).

There is another way of looking at regressions. Suppose, for example, that there was a stable population of voters in the United States over the period of the Markus-Converse study. Index these voters by n: so $S_{n,t}$ represents the partisanship score of individual n at time t. If $S_{n,t}$ was known for all periods and voters, it would be possible to run the regression

(12) $$S_{n,t} = \alpha_t + \beta S_{n,t-1} + \delta_{n,t}$$

That is, the parameters α_t, β can be defined by the requirement that they minimize the sum of squares

(13) $$\sum_{n,t}(S_{n,t} - \alpha_t - \beta S_{n,t-1})^2$$

This definition shortcircuits any question about the errors $\delta_{n,t}$: they are defined by (12). The coefficients α_t, β may be construed as statistics describing the population of voters.[2] So far, so good. The catch is that we cannot do the minimization in (13), because

---

[1] In a draft version of their paper, they argue the weaker assertion, that estimates from aggregate data and panel data are in reasonably good agreement.

[2] Are these descriptive statistics any good? To find out, we have to consider the same issues of linearity and homogeneity raised before, so this reconstruction may only move the difficulty to another place.

we do not have data on all the voters, but only on the ones in the sample. So it is impossible to compute the parameters α_t, β. However, it is possible to run the regression on the sample, hoping the coefficients a_t, b obtained that way will be good estimates of the population values α_t, β, hoping too that the standard errors printed out by the computer package will indicate the accuracy of these estimates.[1]

Unfortunately, it all depends on how the sample was chosen and how the residuals in (12) are related to the variables. The conventional assumptions of simple random sampling and homoscedasticity are embedded in all the computer packages. But the design of the sample used by Markus and Converse involved a series of repeated cross-sections. And each cross-section presumably involved a multi-stage cluster sample. This design is far from a simple random sample, so the standard errors computed by the conventional formula can be off by a large factor. Markus and Converse do not face the issues created by sample design or heteroscedasticity. Neither did any other paper presented at the conference.

To summarize, the form of relationship among the basic variables of the Markus-Converse paper is unknown, so there is no basis for the proposed model or the estimation procedure. Therefore, the conclusions drawn by Markus and Converse do not have adequate scientific foundation.

What are the alternatives? There is no easy, mechanical answer. In particular, I think it would be wrong to introduce still more technique (causal modeling, latent variables, two-stage least squares). I believe it is necessary to begin much closer to the beginning. This means figuring out what

[1]Of course, some of the basic assumptions needed to bring off this sampling-theory justification are quite wrong. The population of voters over the study period was not stable. Most of the regressions were run not on individuals but on cohorts.

the basic variables are, and how to measure them. It means collecting good data. It means developing some theory and some ways of looking at the data which will bring out the fundamental laws connecting the variables. Finally, it is necessary to test the theory by making nontrivial predictions about the future, and seeing whether they come true.

REFERENCES

[1] L. Breiman and D. Freedman (1983). How many variables should be entered in a regression equation? Technical Report No. 1, Department of Statistics, University of California, Berkeley. To appear in JASA.

[2] H. Butterfield (1949). The Origins of Modern Science. London.

[3] C. Christ (1975). Judging the performance of econometric models of the United States economy. International Economic Review, Vol. 16, pp 54-74.

[4] D. Cox and D. Hinkley (1974). Theoretical Statistics. Chapman and Hall, London.

[5] J. L. E. Dreyer (1953). A History of Astronomy from Thales to Kepler. Dover, New York.

[6] A. Franklin (1980). Historical Studies in Physical Science. Vol. 11, Part II.

[7] D. Freedman, R. Purves, R. Pisani (1978). Statistics. Norton, New York.

[8] D. Freedman (1981). Some pitfalls in large econometric models: a case study. Journal of Business, Vol. 54, pp 479-500.

[8a] D. Freedman (1983). A note on screening regression equations. American Statistician, to appear.

[9] D. Freedman and S. Peters (1983). Bootstrapping a regression equation: some empirical results. Technical Report No. 10, Department of Statitistics, University of California, Berkeley. To appear in JASA.

[10] D. Freedman, T. Rothenberg, and R. Sutch (1983). On energy policy models. Journal of Business and Economic Statistics, Vol. I, No. 1.

[11] M. Friedman (1953). Essays in Positive Economics. University of Chicago Press.

[12] O. Gingerich (1971). Sky and Telescope, Vol. 42, pp 328-333.

[13] H. Grosser (1962). The Discovery of Neptune. Harvard University Press.

[14] J. Hausman and D. Wise (1982). Problems in the design and analysis of social and economic experiments. N.B.E.R. Working Paper.

[15] D. Hendry (1979). Econometrics - alchemy or science? Inaugural lecture, London School of Economics.

[16] G. Holton (1978). Subelectrons, Presuppositions, and the Millikan-Ehrenhaft Dispute. Historical Studies in the Physical Sciences. ed. by R. McCormmach et al, Vol. 9, Johns Hopkins University Press, Baltimore.

[17] H. F. Judson (1979). The Eighth Day of Creation. Simon and Schuster, New York.

[18] S. Karlin (1979). Comments on statistical methodology in medical genetics. pp 497-520 in Genetic Analysis of Common Diseases: Application to Predictive Factors in Coronary Disease. Alan R. Liss, Inc., New York.

[19] M. G. Kendall and J. R. Stuart (1961). The Advanced Theory of Statistics. Griffin, London.

[20] J. Kiefer (1979). Comments on taxonomy, independence, and mathematical models (with reference to a methodology of Machol and Singer). Mycologia, Vol. LXXI, pp 343-378.

[21] A. Koestler (1963). The Sleepwalkers. Grosset and Dunlap, New York.

[22] A. Koyré (1973). The Astronomical Revolution. Cornell University Press.

[23] T. S. Kuhn (1969). The Copernican Revolution. Random House, New York.

[24] T. S. Kuhn (1970). The Structure of Scientific Revolutions. University of Chicago Press.

[25] D. Livingston (1973). Master of Light. Scribners, New York.

[25a] R. E. Lucas and T. J. Sargent (1978). After Keynesian macro-economics. In After the Phillips curve: Persistence of High Inflation and High Employment, pp. 49-72, Conference Series No. 19, Federal Reserve Bank of Boston.

[25b] O. Morganstern (1963). On the accuracy of economic observations. 2nd ed. Princeton University Press.

[26] J. Neyman (1977). Frequentist probability and frequentist statistics. Synthese, Vol. 36, pp 97-131.

[27] E. Rosenberg (1979). Cell and Molecular Biology. Holt, Rinehart and Winston, New York.

[28] J. L. Russell (1964). Kepler's laws of planetary motion, 1609-1666. British Journal of the History of Science, Vol. 2, pp 1-24.

[28a] C. A. Sims (1980). Macro-economics and reality. Econometrica, Vol. 48, pp. 1-48.

[28b] C. A. Sims (1982). Scientific standards in econometric modelling, paper presented at the symposium on the developments in econometrics and related fields. Netherlands Econometric Institute.

[29] L. Swenson (1972). The Ethereal Aether. University of Texas Press, Austin.

[30] C. Wilson (1968). Kepler's derivation of the elliptical path. Isis, Vol. 59, pp 5-25.

[31] C. Wilson (1972). How did Kepler discover his first two laws? *Scientific America*.

[32] V. Zarnowitz (1979). An analysis of annual and multiperiod quarterly forecasts of aggregate income, output, and the price level. *Journal of Business*, Vol. 52, pp 1-34.

[33] H. Zeisel (1982). Disagreement over the evaluation of a controlled experiment. *American Journal of Sociology*, Vol. 88, No. 2, pp 378-389.