Some Historical Notes on Statistical Data Analysis

Joe Ward

The historical notes below form a basis for concluding that combining a cell-means prediction-model approach with modern computers can empower data analysts to:

-- Analyze many different-appearing data analysis procedures with one general approach, which reduces the amount of material to be learned.

-- Create models that are more appropriate to the problems of interest, rather than forcing problems into packaged algorithms that may not answer the questions.

-- Reduce the risk of unknowingly obtaining answers from statistical software that are unrelated to the research questions of interest.

-- More easily and correctly specify the computational requirements to the computer.

-- Simplify communicating results of the analyses, since the models are developed from natural language concerns of the researcher.

1951 -

Joe Ward began working at the Air Force Personnel and Training Research Center (AFPTRC) at Lackland Air Force Base to move data analysis from desk calculators to IBM punched card machines. The first task was to implement an iterative algorithm for solving least squares equations that was not sensitive to linearly dependent predictors.

1953 - 1963

Bob Bottenberg and Joe Ward collaborated in enhancing research capabilities at AFPTRC by exploiting the power of Regression Models (Linear Models) made possible through the use of high speed computers. Many experiences combined to bring about a new perspective in research analysis at AFPTRC. While studying at Stanford University, Bottenberg was influenced by Z.W. Birnbaum, Albert H. Bowker, Meyer A. Gershick, George Polya and others. And while attending several Southern Regional Education Board Summer Institutes at the U. of Florida, North Carolina State, and Virginia Polytechnic Institute, Ward had valuable perspectives from association with Richard Anderson, Gertrude Cox, David Duncan, George Nicholson Jr., Lowell Wine and others. Also, of prime importance was the influence of Harry M. Hughes of the Air Force School of Aerospace Medicine.

During the 1950's most of the personnel at AFPTRC were PhD Research Psychologists who had received their statistics education prior to the availability of high speed computers. This meant that techniques of analysis did not involve the use of approaches to analysis that required a large amount of computing. During the late 1950's Bottenberg and Ward developed a Statistics course for personnel at AFPTRC. The plan was to provide a sequence of background concepts that would "eventually" lead to the exploitation of regression models and the computer for analysis. However, the participants were anxious to get on to the highly publicized promises that they would be able to create models appropriate to the research questions of interest and the course contents were adjusted accordingly. Unfortunately, little has changed in many one-semester, required college statistics courses. So much time is spent on the "assumed background prerequisites" that the students are rarely given the opportunity to realize the data analysis capabilities that are readily at their command.

During the AFPTRC course it became apparent that a "Top Down" approach was the way to go for persons who were interested in seeking answers to practical research questions. This implies starting with the problem stated in "natural language" and creating models that fit the problem rather than trying to fit the problem into an easily computable (possibly inappropriate) algorithm. This approach also suggests that concepts be introduced AS NEEDED, rather than spending time on topics which might have been assumed to be prerequisites for creating models to answer questions of interest. In situations where the participants are already indoctrinated with the pre-computer algorithms, it may be useful to relate the regression model approach to the older methods.

This need to empower researchers to create their own models was recognized by Raymond Christal and others at AFPTRC and as a result Bottenberg and Ward were encouraged to develop and document their ideas. This resulted in publication in March, 1963 of "Applied Multiple Linear Regression" by Robert A. Bottenberg and Joe H. Ward, Jr., PRL-TDR-63-6, which is available as AD413-128 from the Clearinghouse for Federal Scientific and Technical Information. For several years after this document was published it was among the highest volume sales from the Clearinghouse.

1964 -

In the summer of 1964 Bottenberg and Ward led a two-week National Science Foundation training session for a group of social sciences university faculty members. This session was directed by Earl Jennings and used the computing and dormitory accommodations at the University of Texas at Austin. The instructional activities focused on the use of regression models and computers in research data analysis. The participants were shown that it was now possible to solve the systems of simultaneous equations that are sometimes required for statistical models. And it wasn't (and still isn't) really necessary to have "equal or proportional n's" that were required BC (Before Computers). Furthermore, even if a researcher has NO OBSERVATIONS in some categories of an "Analysis of Variance" model, the problem can be readily analyzed by stating meaningful hypotheses about the population "cell means" for which there ARE OBSERVATIONS. With model creation skills it may be possible to create a defensible model that produces estimates of population means in cells in which there are no observations and to test hypotheses about the means of those cells.

1967 - 1975

During this period a series of Presessions were led by Bottenberg, Jennings, and Ward at the annual meetings of the American Educational Research Association. These sessions provided an opportunity for practitioners of educational research to become aware of the power of the regression models approach in the computer age. The large number of "graduates" of these Presessions stimulated the creation of the special interest group within AERA, SIG/Multiple Linear Regression. This MLR SIG has an informal publication, "Viewpoints", that provides communication among its members.

1973 -

After many years of teaching about and using regression models and computers, Ward and Jennings collaborated on a book that was to be included in the Prentice-Hall Series in Educational Measurement, Research, and Statistics. Specifically, the book was designed as a supplement to the Gene Glass and Julian Stanley book, "Statistical Methods in Education and Psychology". Englewood Cliffs, NJ: Prentice-Hall, 1970. The book (ILM) by Ward and Jennings was titled "Introduction to Linear Models", Englewood Cliffs, NJ: Prentice-Hall, 1973.

The book was an attempt to provide the reader with fundamental notions that would enable them to create models to answer research questions of interest. The ILM book developed the linear models approach in the traditional sequence presented in the Glass and Stanley book. That sequence was Inferences About the Mean, Difference Between Two Means, One-Factor Analysis of Variance, Two- Factor Analysis of Variance,....

1989 - 1994 >

From 1989-1992 Joe Ward served as a member of the American Statistical Association-National Council of Teachers of Mathematics (ASA-NCTM) Joint Committee on the Curriculum in Statistics and Probability. Ward continues to keep in close contact with the activities of the Committee and continues work with secondary schools through the "Adopt a School" program of the ASA. Ward started working with high school students and teachers in the use of computers in 1958. While the emphasis during those early years was on introducing computers into the secondary schools, Ward took the opportunity to introduce a few high school students to the combined power of regression models and computers. He now works with high school students and teachers in the San Antonio area who wish to enhance their data analysis skills. Ward teams with Laura Niland, a statistics teacher at MacArthur High School and the 1988 Texas Presidential Awardee in Secondary Mathematics, in workshops for high school teachers and students. Ward has taught Problem Solving Using Data Analysis to high school students in the Prefreshman Engineering Program (PREP) of the University of Texas at San Antonio.

Teaching both high school and college students who have had no previous introduction to Data Analysis has lead to the conclusion that a "TOP-DOWN" approach to Data Analysis will allow students to make practical use of their Data Analysis

experiences before they become "turned-off". Notice the use of the term "Data Analysis" in place of "Statistics". The use of a "different name" for the course allows more freedom to start with real-world problems, introduce the use of regression models and computers and apply these techniques to the data analysis requirements. Topics that are frequently taught as prerequisites are introduced when needed in the data analysis process.

1951 - 1994

Ward has interacted with a wide variety of researchers who call themselves by different labels. These include Research Psychologists, Educational Researchers, Operations Researchers, Economists, Statisticians, Computer Scientists, Mathematicians, Sociologists, Management Scientists, Engineers, etc. Fortunately many of these researchers learn -- while on the job -- to create models to fit their problems and to use the computer to "crunch the numbers". However, observations of newly trained researchers and the books used for their training indicate that much time is spent learning the "pre-computer" approaches to data analysis. Those authors that do show the student some examples of the general linear model approach to analysis do little to empower the student to create their own models. It is not clear why many classical texts include so many special computational formulas that were necessary in earlier years. There may be a belief that a learner acquires a stronger degree of understanding if they know how to do the pre-computer arithmetic. Many of the statistical software packages emphasize the use of the computerized versions of the "pre-computer" algorithms. And these packaged programs can occasionally provide answers to uninteresting questions that are different from the hypotheses that the data analyst thought were being tested.

There still remains a great need to develop instructional approaches that will allow researchers to create their own models as required and to use the computer to handle the computational burden. It seems that a good approach to introducing students to model development is to begin with a problem that is of interest to them and to use concepts that are familiar. The use of "averages" of collections of data are a great way to start since learners of all ages have heard the term and have been subjected in school to the use of "averages" as performance indicators. Most learners can talk easily in "natural language" about comparing "averages" among categories (e.g., batting averages, shooting average, etc.). Then these ideas can be expressed in more formal prediction models of forms such as :

DEPENDENT VARIABLE = PREDICTION + ERROR,
DATA = FIT + RESIDUAL,

DATA = MODEL + ERROR, or

$Y = XB + E.$

The important idea is to provide learning experiences that will eventually allow students to create models relevant to the questions of interest. The solutions to these models are now feasible by high-speed computers.