Conference on Applied Statistics in Agriculture 1997 - 9th Annual Conference Proceedings

SOME FACTORS LIMITING THE USE OF GENERALIZED LINEAR MODELS IN AGRICULTURAL RESEARCH

Walter W. Stroup

SOME FACTORS LIMITING THE USE OF GENERALIZED LINEAR MODELS IN AGRICULTURAL RESEARCH

by Walter W. Stroup
Department of Biometry
University of Nebraska-Lincoln

ABSTRACT

The generalized linear model (GLM) is a "hot" topic in statistics. Numerous research articles on GLM's appear in each edition of all major journals in statistics. GLM's are the subject of substantial numbers of presentations at most statistics conferences. Despite the high level of interest and research activity **within** the statistics community, GLM's are not widely used, with some exceptions, by biological scientists in the statistical analysis of their research data. Why? Reasons include 1) many statisticians are not comfortable with GLM's, 2) the biological research community is not familiar with GLM's, and 3) there is little in introductory statistics courses as currently taught to change (1) or (2). Whether or not this is a real problem is unclear. This paper looks at some of the factors underlying the current state of GLM's in statistical practice in biology.

1. INTRODUCTION

The 1997 East North American Regional meetings of the International Biometric Society featured a session entitled *Impact of Generalized Linear Models on the Agricultural and Environmental Sciences.* As I had done some work on various agricultural applications of GLM's, I was asked to give a presentation. When asked, my initial reaction was, "What impact?" As I discussed this with colleagues and with the session organizers, it became clear that any session dealing with impact ought to consider not just disciplines using GLM's and disciplines pushing the further development of GLM's, but also disciplines which routinely use response variables that seem to call for GLM's but which for the most part do not use them.

My immediate reaction was based largely on impressions through the consulting and collaborative work that I do in the Institute of Agriculture and Natural Resources at the University of Nebraska, and through reading the journals of the disciplines of those with whom I consult frequently. With a handful of exceptions, statistical procedures used in agricultural research seem to be almost completely untouched by GLM's. Two notable exceptions among the disciplines in which I work with some frequency are animal breeding and genetics and some areas of ecology. In animal breeding and genetics, there is a *lot* of research concerned with "threshold" models, that is, generalized linear *mixed* models for ordinal categorical data. These are cumulative link models overlaid with a substantial amount of animal breeding jargon (which unfortunately sometimes keeps animal science graduate students from seeing their relation to other linear models!). Probit links are commonly used because the assumed underlying unobservable normal process lends itself to well-known quantitative genetic theory. A major focus is on the development of estimation techniques for very large data sets through sparse matrix procedures, Gibbs sampling, and related

procedures. In ecology, response variables are often binary (present/absent) or counts. There is increasing acceptance of logit and log-linear models as well as active interest in trying to understand how counts (e.g. of insects, weeds, fish, etc.) are actually distributed in field studies and how best to model them.

These are the exceptions. To check the validity of my impressions I spent some time in the library, looking over research articles from several dozen journals from a number of disciplines in the agricultural and environmental sciences. I focused on journals identified by my consulting clients or by clients of my colleagues in the Biometry Department as important in their disciplines. Disciplines included agronomy, animal science, agroforestry, botany, entomology, fisheries, food science, genetics, horticulture, irrigation technology, molecular biology, plant growth regulation, plant nutrition, range science, veterinary science, virology, weed science, wildlife ecology, and zoology. While the majority of these articles dealt with continuous response variables where the assumption of normality is likely reasonable, a substantial number used response variables that **should** be likely candidates for GLM's. Many articles dealt with treatment effects on percent occurrence/nonoccurrence (e.g. seed germination, adcission, leaves affected, surviving animals, species present). Others dealt with counts (e.g. number of weeds, number of fish, number of virus). Many articles addressed treatment comparisons by evaluating "percent of control," a ratio of two random variables.

In most of the articles I surveyed, data were analyzed - without regard to plausibility of the response variable's normality - using 2-treatment-at-a-time t-tests, or, in more sophisticated cases, ANOVA F-tests. Occasionally one sees a transformation, such as arc-sin [square root(percent)] or log(count+1). Genetics researchers make frequent use of chi-square tests for goodness of fit to evaluate relative genotype frequencies. Other than the examples mentioned in the preceding paragraph, GLM 's are very rare. Even in animal genetics, despite the high level of sophistication in development of GLMM's, they are seldom used in practical analysis of animal breeding trials, even when the response variable is categorical.

Clearly, despite the high level of interest in GLM's among biometricians, GLM's are not among the statistical tools in use by most agricultural and environmental researchers. Is this a problem? If it is a problem, how serious is it? If it is a serious problem, what factors are responsible and what would it take to correct them? Let us consider each of these questions.

2. IS THERE A PROBLEM?

The majority of research that depends on non-normal response variables is currently analyzed using simple t-tests, ANOVA, or regression methods, as if the variables were normal. Transformations are unusual. GLM's are rare - unheard of in many disciplines. If all these analyses were redone using appropriate GLM's chosen using thorough model checking procedures, how much would the accuracy, power, efficiency, validity , etc. of the results change? Aside from the more blatant and egregious abuses of statistics (which are usually a result of poor **design** more than poor choice of data analysis method), it is not clear that we really know.

Many consulting statisticians who are quite well-trained and familiar with GLM's are nonetheless reluctant to recommend their use in practical situations. Part of this stems from the fact that their clients are often unfamiliar with GLM's. But even if their clients were more aware, many of my colleagues say they would still hesitate. Their reluctance stems from the fact that confidence intervals and hypothesis tests with GLM's require asymptotic statistics whose small sample properties, they argue, are insufficiently documented. Ordinary least squares may not be pretty, but it ***is robust***. Better, as John Maynard Keynes once said, "to be approximately correct than to be precisely wrong."

Clearly, from the above comments, many statisticians are not convinced that infrequent use of GLM's is much of a problem. I did a fairly extensive, but by no means exhaustive, study of how much research actually exists comparing GLM's to alternative methods. In the Current Index of Statistics, there were 191 articles between 1972 and 1994 on GLM's. Most were either theoretical in nature, or reported a new model to do this or that. A handful contained an empirical study of the small sample characteristics of the methods associated with the new model presented. None was explicitly concerned with a comparison of GLM's to competing non-GLM alternatives. Herein, I believe, lies a problem that the statistics community must address.

Although I am convinced that GLM's are seriously underutilized agriculture and ecology, I have to admit that my more skeptical colleagues have a good point. For example, if a client of mine has binary data, I can argue that a logit or probit model will keep estimated probabilities in the parameter space. I can argue that odds or odd-ratios have more satisfactory interpretation than simple differences among proportions (or for a geneticist, I can appeal to the advantages of the probit model mentioned earlier). However, can I produce hard evidence that my client's ultimate conclusions will be surely and seriously affected? Or that the power or efficiency benefit from a GLM will be substantial? This is far less clear. Advocates of greater awareness and use of GLM's by researchers cannot really expect this to happen unless studies of GLM's small-sample behavior clearly establish them to be either better - more accurate or more efficient - than non-GLM alternatives, or at least equal statistically, but more amenable to sound scientific interpretation.

3. WHAT FACTORS ARE RESPONSIBLE?

Researchers use the tools they are taught and the tools that are available. To understand why t-tests and F-tests are so widely used, and are usually applied without transformations regardless of the response variable, one needs to look at the statistical curriculum for agricultural and environmental researchers in training and the statistical software that is most available through training and infrastructure.

First, consider the curriculum. A student getting a M.S. in an agricultural or environmental discipline will typically take at most a year (two semesters) of statistics, usually using Snedecor & Cochran (1989) or a Snedecor & Cochran clone as a text. They will be exposed to t-tests, F-tests, correlation, and regression, all for normal random variables. They **may** be exposed to contingency tables. Most of what they are taught will be in the context of orthogonal experimental designs (CRD, RCBD, Latin Square). A Pd.D.

candidate may take one or two additional courses (a more advanced regression or design course, a course in survey sampling methods). The majority of these third and fourth courses involve more sophisticated methods for normally distributed data, but rarely mention GLM's, much less treat them in any depth. Researchers use what they are taught.

If one looks at the curriculum for graduate students in statistics, the situation is similar. Particularly for students preparing for careers as consultants, courses are heavily weighted toward Snedecor & Cochran-style methods. There is more depth and substantially more underlying theory, but the M.S. student graduates prepared to use the same basic set of methods taught to their biologist colleagues-in-training. Most will have a superficial exposure to GLM's. Many will have no exposure at all. Most linear models courses persist in teaching "the general linear model" exclusively, even though contemporary linear model theory ***is*** generalized linear model theory and the normal errors "general" (specific?) linear model is but a special case.

Snedecor & Cochran-style methods have been enormously successful. Their robustness is well-documented and time-tested. In many agricultural disciplines, these methods have been institutionalized since the 1920's or 1930's. They are considered "standard methods." Journal editors expect to see these methods used. Members of the disciplines are comfortable providing peer review when these methods are used. Advisors are most comfortable when their students use these methods. Using "non-standard " methods requires extra effort on everybody's part. This usually includes close collaboration with a statistician. Many universities are set up to reward faculty statisticians for this kind of effort, but many others clearly are not. The case for using something other than what is taught in standard statistics courses has to be pretty strong.

Second, consider statistical software. Until fairly recently, software for using GLM's was not available in a form that was marketable to non-statisticians. In the 1970's, software such as SAS PROC ANOVA and PROC REGR (later PROC GLM - not to be confused with *generalized linear model!*) became available. These programs were widely integrated into Snedecor & Cochran-style courses. Their impact on agricultural research is obvious. Until the 1990's, computer software for GLM's has not been at the same level of accessibility. Now, PROC GLM-like software, e.g. PROC GENMOD, is available for GLM's. However, it is not integrated into the curriculum as is PROC GLM. Perhaps if it were, things would begin to change.

4. WHAT WOULD IT TAKE TO BRING GLM'S INTO WIDESPREAD USAGE?

There are two requirements for GLM's to have more of an impact on biological research. First, comparative research must establish the value of GLM's relative to alternative statistical methods. If nothing else, the sheer inertia of over a half-century of "standard practice" works against alternatives that lack compelling arguments in their favor. Assuming such a case is made, then the second requirement for bringing GLM's into widespread usage is education - and a lot of it!

If we truly believe that GLM's should be among the standard tools for agricultural

and environmental researchers, then statistical curriculum has to be amended accordingly. Roger Mead (1988), in his *Design of Experiments* text, commented that despite radical advances in computing technology between the 1930's and the present, there was little evidence in current statistical methods texts that these advances had ever occurred. He made this comment in 1988 - the texts are still little-changed but look at the changes in computing since then!! A current introduction to statistical methods course is not much different from the course laid out in Snedecor's first edition in the early 1930's, with its heavy emphasis on statistical arithmetic (e.g. t-values and simple ANOVA tables) and on designs simple enough and response variables well-behaved enough to permit the use of simple arithmetic. The main evidence of the computer revolution is that we have computerized the t-test and the ANOVA table. I have colleagues who argue persuasively in favor of the traditional methods course curriculum. However, if GLM's are essential tools for modern researchers, then the unmodified traditional methods course is clearly a disservice to our clientele.

Some have suggested leaving the introductory courses alone, but having students who work with categorical data or counts take an additional course. For Ph.D. students that may work, but it is unrealistic for M.S. students. I work with a number of M.S. students who have thesis research that revolves around non-normal data. There is no room in their schedule for a third statistics course. It is out-of-touch, and I think a bit arrogant, for us as a profession to respond to their needs in this manner. We have to assume we have these students for two semesters. Period. If we truly believe GLM's are an important tool for agricultural and environmental researchers, we have to adjust these classes. There is one text that reflects an attempt to do just that - Lindsey's (1995) *Introductory Statistics*. Though written for social science students and probably unsuitable, as is, for graduate students in the biological sciences, it does at least illustrate that change is not inconceivable.

Similar comments apply to core courses for graduate programs in statistics. Although the theoretical **components** of GLM's are present in most mathematical statistics and linear models courses, these courses currently tend to emphasize providing theoretical underpinning for traditional t-tests, ANOVA, and regression. A different emphasis is required to prepare these students to work with and advise others on GLM's.

A concerted program of continuing education would also be required. As with the education of students, this works on two levels: education **within** the statistics community and education **by** statisticians **to** members of the various biological disciplines. A fair amount of continuing education is already occurring within the statistics community through activities such as short courses at ASA and Biometrics meetings and in-house training of statisticians employed by agriculturally or environmentally oriented industries. Education **by** statisticians **to** members of the various biological disciplines is another matter.

Most of us work collaboratively or as consultants with agricultural and environmental researchers. Many of us offer seminars and workshops to our colleagues in agriculture and environmental science. Some of us belong to professional societies such as the Agronomy Society of America. Some are associate editors for their journals. Some of us write feature articles on statistical methods for their journals. These are all ways to "get the word out." It is not very controversial to say that these things are important and we all ought to consider

them part of our professional responsibility. We do have to make sure that our employers, our administrators, our fellow statisticians, the Biometric Society, the American Statistical Association, etc., understand the importance of these activities - and the time and energy they take - and do their part to ensure that they are adequately rewarded.

Returning to the original question, how much impact have GLM's had on agriculture and environmental science? The answer is "not much," although they have had an impact on what statisticians who consult in these areas talk about with one another. Clearly, it would take a concerted educational effort to bring GLM's into more widespread use. Is such an effort justified? No consensus exists, and the question itself may be premature. What I have called, for want of a better term, "standard methods" are deeply entrenched in statistical curriculum and statistical practice. Comparative research needs to do be done to establish the tangible benefits, if any, GLM's offer. A consensus among statisticians (about anything!) is unlikely, but agricultural and environmental researchers do expect (justifiably) a certain amount of consistency in the curriculum of our methods courses and in the procedures we recommend for various situations. If a strong case for GLM's is made, we *must* make room for them - and not just in an odd course here and there.

REFERENCES

Lindsey, D.K. (1995) *Introductory Statistics: A Modelling Approach.* Oxford University Press, Oxford, UK.

Mead, R. (1988) *The Design of Experiments.* Cambidge University Press, New York.

Snedecor, G.W. and W.G. Cochran (1989) *Statistical Methods, 8th Edition*, Iowa State University Press, Ames, IA.