Journal of Methods and Measurement in the Social Sciences,
Vol. 6, No. 2, 60-69, 2015

Enhancing the Teaching of Statistics by Use of the Full GLM

Richard L. Gorsuch
Fuller Theological Seminary
UniMult, Inc.

Increased elegance in math and science is by the use of more comprehensive, easier to understand, and easier to use models. Increasing elegance allows courses to cover more material in greater depth. While the GLM is more elegant than the traditional ANOVA / Regression models, it has in practice been just one more topic added to already filled statistics courses and has had little impact on day-to-day statistical analyses. Introduced in the 1960s - 1970s, its impact has been delayed because it has been necessary to produce a new generation that knew the GLM but could also converse with the pre-1970 generation. When considering a possibly more elegant model, the "full" GLM includes not only GLM -- and therefore ANOVA and regression -- but also chi square contingency table analyses as well as multivariate analyses, and uses the *F* as the hypothesized variance divided by the error variance in all cases. Given advancing technology, the computations are now readily done and typically easier than the traditional ANOVA/Regression programs, allowing more focus on issues ranging from how the information is encapsulated so as to best test the hypotheses (by logs, logits, interactions, polynomials, repeated measures as slope or covariates, etc.) to meta-science and the principles of meta-analysis needed to use research literature. And, as expected, the more elegant full GLM means explicitly GLM software can be easier to learn and use.

Keywords: General Linear Model, Unimult, univariate statistics, multivariate statistics

Teaching, Elegance, and the GLM

Currently, the usual statistics course sequence introduces the student to four models of statistics: (1) contingency table ("chi square", often assumed to be adequately covered in a prior introductory course), (2) multiple regression, (3) ANOVA, and (4) the GLM (general linear model). Could our students be better served if they had only to learn one model instead?

There are always multiple models which may aid in answering a question, models for which a decision must be made as to which is a better model to include (given the reality constraints in pedagogy). A simple example in early statistics was the several measures proposed to measure variability. These included two ways of measuring variability beyond the simple ones: the average deviation (*AD*) and the standard deviation (*SD*).

The latter is the one we use currently; the former is defined in Guilford's first edition of his text (Guilford, 1942) as

> "$AD = \Sigma|x| / N$ (1)
> where *AD* = average deviation, |x|, with the vertical bars embracing it = absolute value of x, *i. e.,* disregarding algebraic sign." (Guilford, 1942, pg.59)

AD has some advantages. It is not influenced as much by extreme scores as the *SD*, it provides a "fairly reliable" estimate without the computational time for *SD*, and the *SD* can be estimated from it (if the distribution is "fairly normal", *SD* = 1.253 *AD* (Guilford, 1956, pp. 117f))

In later editions, *AD* was not even in the index (Guilford & Fruchter 1978, 6th edition). It was dropped because it had been superseded by *SD*. *SD* was more elegant because it could be related more directly to the normal curve and was used in various formulae. The course was more elegant just using the *SD*, and the time could be better spent than teaching *AD* as well.

If we used the GLM only, would it not simplify our courses? A few examples similar to the *AV/SD* example will illustrate some simplifications possible using the full GLM as the primary model for statistics such as ANOVA and regression.

The correlation coefficient, *r*, and eta (also known as the correlation ratio) are both introduced as the square root of the proportion of hypothesized (model) variance (with sign added for the correlation coefficient) and then used for all cases of the full GLM.

In addition to having one definition of effect size, this effect size can be used in a universal significance test that applies to all three of the separate statistical models: the *F* test, taught as the ratio of two estimates of variance. The universal *F* test is:

$$F = H / E \quad (2)$$

where *H* and *E* are the variance estimates from the hypothesized model and from the errors (or residuals) and which can be computed from *R* as $((R^2/df1)/((1-R^2)/df2)$, *F* is Fisher's *F* Ratio, *df1* is the *df* of Xs times the *df* of Ys, and *df2* is the number of cases minus *df*1 minus one. Variants, such as used in model comparisons of a full model and a reduced model, are then added as special cases.

Another major simplification is in language. All that is needed is the list of variables (including transformations for interactions, curvilinearity, etc. as needed) for the X(s) and for the Y(s) to describe the analysis. Labels such as "regression analysis" and "ANOVA" are redundant.

> "As has always been the case in theory, the type of analysis is a function of the type of variables submitted. There are two major types of variables: nominal, a variable with a set of mutually exclusive variables, and non-nominal variables, for which cases vary on a scale (technically includes ordinal, interval, and ratio levels of measurement). (Dichotomous variables are a special case of either nominal or non-nominal variables depending on how they are conceptualized). Table 1 gives the classical names of analyses among these variables. For example, if several nominal variables are included along with their interactions, then that is a classical ANOVA so long as Y is non-nominal. If all the Xs are non-nominal and the Y is interval, then it is a multiple regression analysis. If nominal and non-nominal variables are included, it is, for lack of a traditional name, called here a least squares analysis. Hence the label for the analysis arises out of the level of measurement of the variables." (UniMult, 2015)

Students need to memorize this table for reading the pre-GLM literature.

Table 1
Classical Univariate Labels as a Function of Level of Measurement

	Y Variable	
X Variable(s)	Nominal	Non-Nominal
Nominal	Contingency Table	ANOVA
Non-Nominal	Classification Analysis	Multiple Regression
Mixed		GLM

(Adapted from Table 3.0.1, UniMult Guide, Chapter 3)

The GLM normally taught is univariate. That includes ANOVA and regression but not any analysis requiring a nominal Y, whether it be a contingency table analysis or a discriminant analysis. Knapp (1978) noted that, however, these are included in a full GLM which uses the multivariate generalization of the GLM (which can be taught in an advanced course). Variations on this full GLM model include canonical analysis (Thompson, 1984) but also includes variations that allow a multitude of applications (e. g., Harris, 2001, Heese, 2011) and this uses the general *F* ratio formula given above (Heese, 2011, pg. 105). This means that contingency tables can be analyzed (which gives the exact same *p* as the classical cross-tab analysis) with the other possible analyses. (The terminology for the general multivariate model is not yet standardized; some, such as Knapp (1978) and Figueredo (personal communication) define "canonical analysis" to include the variants. The current approach considers canonical analysis to be a variant which includes canonical variants.) In the full GLM, canonical variates which have always been

difficult to interpret are not needed. The generalization to the full GLM is surprisingly easy for those who have been introduced to the GLM, which means later courses can build upon the research sequence using the full GLM as its model.

There are classical labels for analyses computed under the full GLM.

> "As has always been the case in theory, the type of analysis computed depends upon the level of measurement of the Xs and of the Ys. So when you identify your variables, UniMult™ then computes the most powerful analysis possible for those variables based on whether they are dichotomous, nominal, or non-nominal (the term "non-nominal" includes ordinal, interval, and ratio levels of measurement).
>
> Exchanging the X list of variables for the Y list produces the same overall test of significance, but some of the other output varies. To guide you in deciding which variable(s) to list as Xs and which as Ys when the research design is not clear on this, see Table 2. It summarizes the type of analysis as a function of the nature of the variables. Perhaps more important, it gives the classical names for the special cases. Just as is the case with Xs, the names are not needed to run analyses nor when presenting the results, but they are still essential to be able to read the literature. Analyses mixing nominal with other types of variables are fine if logically appropriate for your problem" (UniMult, 2015).

Students need to memorize this table for reading the pre-GLM multivariate literature.

Table 2
Classical Multivariate Labels as a Function of Level of Measurement

	Y Variables	
X Variable(s)	Nominal	Non-Nominal
Nominal	Contingency Tables	MANOVA
Non-Nominal	Discriminant/ Classification Analysis	Multivariate Regression
Mixed		MANCOVA Full GLM

(Adapted from Table 5.0.1 UniMult Guide Chapter 5)
Note: When both X and Y contain mixed levels of measurement, it is a Full GLM analysis.

Why Do We Teach As If There Are Multiple Models?

The GLM model just referenced was, with only a few exceptions, established by 1975. Here is a partial list of major contributions from that

period which introduced us[1] to the GLM:

- Bottenberg and Ward (1963) was an Air Force technical document which introduced model comparison analysis using nominal variables with the categories coded zero and one along with non-nominal variables and interactions as a part of regression analysis.
- Cohen (1968) introduced psychology at large to the GLM using the language of regression analysis.
- Kelly, Beggs, and McNeil (with Eichelberger and Lyon) (1969) taught the GLM as a multiple regression extension.
- Tatsuoka (1971; 1975) gave a concise introduction to ANOVA as a GLM special case.
- Ward and Jennings (1973) published a true GLM text, *Introduction to Linear Models*, using neither regression analysis nor ANOVA terminology.

By the rational in the *AV-SD* example, and as both ANOVA and multiple regression are special cases of GLM, it is more elegant to teach the GLM as the basis for all analyses and use the time freed up for other important topics[2].

By 1975, the GLM had become well known. But it did not become the standard paradigm but, rather, was added as an additional model. So we are in the embarrassing position of teaching models that are over 30 years out of date. What happened? I have seen no clear data and am not a historian, but several suggestions can be made.

---

[1] Guilford concluded his text with material on reliability and validity. While there are several assumptions for the GLM, the most crucial for a proper conclusion is that the variables be reliable and valid. As noted in the last table in Gorsuch and Lehman (2012), their impact on statistics is more dramatic than any other factor. Today, I vote that the time saved by teaching one model instead of four models in a statistics course be the principles of meta-analysis (although Black, Garcia, & Figueredo (2014) and Garcia, Black, & Figueredo (2014) make the case for meta-science); students will spend much more of their career reviewing literature than they will running analyses. And research studies can only be properly combined in even a casual review if there is an understanding of meta-analysis, which brings together the principles of research/evaluation design and statistics.

[2] As Cohen (1968) points out, the GLM with regression and ANOVA as special cases has been long known among statisticians. Tatsuoka (1975) notes that Fisher seems to have known this -- he was very active in correlational methods before introducing ANOVA -- and suggests Fisher did not present ANOVA as GLM because the computational procedures were too difficult; assumptions he imposed on ANOVA allowed simplification of the calculations. By this reasoning, ANOVA as a separate model should have been dropped with the introduction of computers in the 1960s -- which seems to have been an unacceptable conclusion to many. Perhaps note should also be taken of Fisher's assumption that everyone could see statistics as he did -- as geometry in his head that needed no further explaining -- and his antipathy towards Pearson whose regression model resembled GLM.

A factor reported by colleagues is the resistance of other faculty to changing the courses. They have been doing their analyses by the older models and wanted their students to know the analyses they currently use. As all of that language is a special case of the GLM that is easily accomplished once the GLM has been recognized. In the 1980s, many faculty had taken their statistics courses prior to the GLM, but the GLM has generally been taught in graduate classes at least since 2000 so that should now be of lesser importance. Their interests can be met by materials such as Tables 1 and 2.

A complicating factor may have been that the GLM was easiest, in the initial years, to formulate as a regression model. Of the references listed above, only Tatsuoka (1971, 1975) was obviously ANOVA. (Ward and Jennings (1973) was not a regression model but referenced neither regression analysis nor ANOVA and so seemed a different model.) Given the fact that those doing non-experimental research used regression as their paradigm while those doing experimental research used ANOVA and did not recognize the regression-based GLM as ANOVA, hostilities between these two groups -- which developed in part from the antagonism between Fisher (ANOVA) and Pearson (correlation) -- led to a defensive entrenchment that prevented statisticians from shifting to the GLM. People protected their own paradigm.

Perhaps the common statistical courses should be labeled Research Methods I and II (rather than ANOVA and Regression) which could at first be taught as a continuation of the courses as currently taught. The courses could then be transitioned to GLM as the professors of these courses are able to appropriately re-arrange their lecture materials, treating both regression analysis and ANOVA in detail but as special cases.

The difficulties shifting to the GLM are attributed to, by others I have asked, the desire to continue teaching from the text the new instructor used when a graduate student. It is true that introductory textbooks that give a prominent place to the GLM are few (Thompson, 2006; Vik, 2014; Figueredo in development, personal communication), although there are many for advanced courses, including Allison (2009), Dunteman and Ho (2006), Fox (1997), Gill (2001), Hardin and Hilbe (2007), and Stevens (2007). But there would be more first year graduate texts if there was demand.

Another factor mentioned in my informal survey is, to be frank, laziness on the part of instructors. They do not wish to do a major revision of what they have been teaching. And if they are new, they want to use the notes from the class they took as students as the basis for their lectures. Sharing of GLM presentation slides could help answer this need.

Software Support

Personally, it has seemed to me that a major problem is that most statistical packages have their routines almost frozen in place -- and they use ANOVA and regression terms (any GLM program has generally been just added to the rest and has often been difficult for students to use). These programs are like the statistical package I developed at Vanderbilt University in the 1960s. They were built on hand calculations and had a separate routine for named special cases of the GLM that were commonly used and taught. Special case programing was used with each routine -- which followed the short cuts developed for hand calculations or for calculators -- but it is still part of the GLM and so the GLM was in essence programmed in each routine. The only major improvement has been adding an over-arching master program to obtain the data and pass it to a called subroutine stored on a disk -- which is just the traditional 1960's program with improved interface. This means that variations on the GLM are programmed multiple times into most statistics packages. Indeed, the packages have often become more complex due to the overhead of managing the multiple subroutines.

The resulting packages can be difficult for students to master. In talking with clinical psychology students using one of the most widely used programs, they report a steep learning curve, copying script and using it verbatim without understanding. They then need to force their data to fit one of the several procedures for which they have the script, instead of adapting the analyses to fit their problem. They have little confidence that they can apply it to their master and doctoral theses without help so extensive that it amounts to the helper running the analyses for them. When a class was asked to give three adjectives to describe the package, the most common were terms such as "confusing" and "frustrating". This type of program locks teacher and student alike into pre-1975s paradigms.

As the lack of a true full GLM computer program is a strong reason for still teaching multiple models, a new program has been developed from the ground up knowing everything is a special case of the full GLM. Due to the greater elegance of the GLM as opposed to having contingency table, ANOVA, and regression separately programed, such a program would also be simpler to use. The major simplification from running the GLM on everything is that there is no need for the user to specify ANOVA or regression or contingency table analysis. All the program needs to know is whether a variable is nominal (so that nominal category coding can be set up) or not and whether the variable is an X or a Y. So the program just asks for which variable(s) are X(s) and which are Y(s) (see UniMult walk-through and Guide, available at UniMult.com).

And why should one need to tell the program that it is a multivariate analysis? Can not the computer count the number of Ys submitted? With

the computations using the full GLM there is no need to specify univariate or multivariate. The program just cycles through for the number of Ys. If there is only one, it stops after that one. If there is more than one Y, it continues to cycle for each.

I am a firm believer that the purpose of computer software is to make our lives easier, and we have worked to make that so with UniMult. In addition to using the most elegant statistical model available -- the full GLM -- we have used another basic principle: recognition is easier than recall. A major problem with any software that requires a special script is that the script must be learned until you can recall it. That's fine -- except for students and for those of us who do only three or four analyses a year; in other words, it is fine for the professional statistician but not for most of us. Recognition is much easier.

With UniMult, all of the possible next steps are in the current window. It is a just a matter of looking over the options until you have an "ah ha!" experience. And if you are still unsure? Pick one that might be helpful and give it a try. Does it give you the answer you need? Still in doubt? The *Guide* discusses every option if you prefer written directions (these can also be useful supplements to your text). As elegance is the quality of being easy for us humans to use, you could say that we are working not only for more elegant statistics but for more elegant software as well.

The student learns that the gist of most variations in data analysis lay in the information submitted, so understanding of the basic principles is needed to run the GLM for different research designs. For example, you can include an interaction or leave it pooled in the error term, can process repeated measures by a transformation which gives the slope/change from one occasion to the next or can use exponents to include possible curvilinearity effects. Thus the emphasis is not on a script to write but on what information will be most useful.

Running the GLM means that one is not limited to classical procedures. The classical ones came from the days of hand calculation and therefore were for the popular analyses. But the popular analysis may not be the needed analysis. One may, for example, need to interact a three category nominal variable with the linear and quadratic transformations of another X for the major Y with a prior Y control variable partialled out. With UniMult, the appropriate transformations are requested and then listed as Xs and Ys. That process can be done almost quicker than it has taken me to write this paragraph.

Greater elegance has the delightful quality of making statistics easier. This is not only the case with teaching statistics but also in computations. In addition to a spreadsheet window for data entry and transformations, UniMult has only one other major screen. It just requires entering the Ys and the Xs (including any interactions or other variations such as for curvilinearity and repeated measures). There are, of course, a variety of

simple special screens and transformations to process variants such as factor analysis. The student's focus is on choosing the variants that best test the model being addressed, not on scripts or other memorized particulars of the package.

Student input has been quite positive. The first version of UniMult was in 1990, and it was those students that encouraged a new version that takes advantage of the many improvements since 1990 -- such as each student being able to run the program from their personal computer or from any other computer with internet capability regardless of location. (And it has been good to hear a student in their first in-depth experience with any statistical package spontaneously say "This is easy!" The greater sophistication and elegance of the full GLM makes the program easier.)

Students report it takes less than an hour to learn the program. All other time in labs can then be spent on learning what it means to, for example, have repeated measures.

Users of the first version have been publishing the UniMult results with ready acceptance from editors and reviewers. At most, they are occasionally requested to provide more information on a more unusual analysis that is proper GLM but does not fit easily into the older models although it is the best test for the question at hand.

Conclusion

We now have the ingredients to bring the elegance of the 1975 full GLM into teaching and doing statistics. It would be nice to have students learn the late 20th century paradigm for use in the 21st century.

Author Notes: An earlier draft of this paper, "Enhancing the Usefulness of the Full GLM for Teaching Statistics", was presented at the American Psychological Association 2014 Annual Meeting, Washington, D. C. Richard L. Gorsuch is the majority stock holder of UniMult Inc.

References

Allison, P. D. (2009). *Fixed Effects Regression Models*. Thousand Oaks: Sage.

Black, C., Garcia, R. & Figueredo, A. (2014) Meta-Scientific Foundations of Statistics. Presented at the 2014 Annual Meeting of the Association for Psychological Science.

Bottenberg, R. A. & Ward, J. H. (1963). *Applied Multiple Linear Regression*. Lackland Air Force Base: Personnel Research Laboratory.

Cohen, J. (1968) Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70, 426-443.

Dunteman, G. H. & Ho, M. R. (2006). *An Introduction to Generalized Linear Models*. Thousand Oaks: Sage Publications.

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks: Sage.

Garcia, R., Black, C. & Figueredo, A. (2014) Meta-Scientific Applications of Statistics. Presented at the 2014 Annual Meeting of the Association for Psychological Science.

Gill, J. (2001). *Generalized Linear Models: A Unified Approach*. Thousand Oaks: Sage.

Gorsuch, R. (1991). *UniMult: For Univariate and Multivariate Data Analysis*. Altadena: UniMult.

Gorsuch, R. L. & Lehmann , C. (2010). Correlation Coefficients: Mean Bias and Confidence Interval Distortions. *Journal of Methods and Measurement in the Social Sciences, 1*, 52-65.

Guilford, J. (1942 1st edition, 1950 2nd edition, 1956 3rd edition) *Fundamental Statistics in Psychology and Education*. New York: McGraw-Hill.

Hardin, J. W. & Hilbe, J. M. (2012). *Generalized Linear Models and Extensions*. (3rd ed.). College Station: Strata Press.

Harris, R. J. (2001). *A Primer of Multivariate Statistics*. (3rd ed.). Mahwah: Lawrence Erlbaum Associates.

Heese, Richard F. (2011). *Multivariate Linear Models*. Los Angles, Sage.

Kelly, F. J., Beggs, D. L., & McNeil, K. A. (with Eichelberger, T. and Lyon, J.). (1969). *Multiple Regression Approach: Research Design in the Behavioral Sciences*. Southern Illinois University Press.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. *Psychological Bulletin, 85*, 410-416.

Stevens, J. P. (2007). *Intermediate Statistics: A Modern Approach*. (3rd ed.). New York: Taylor & Francis Group.

Tatsuoka, M. (1971). *Multivariate Analysis*, 2nd edition. NY: John Wiley & Sons.

Tatsuoka, M. (1975). The General Linear Model: A "New" Trend in Analysis of Variance. Champaign, IL: IPAT. A republication of the identically named chapter in Tatsuoka (1971).

Thompson, B. (1984). *Canonical Correlation Analysis: Uses and Interpretation*. Newbury Park: Sage.

Thompson, B. (2006). *Foundations of Behavioral Statistics: An Insight-Based Approach*. New York: The Guilford Press.

Vik, P. (2014). *Regression, ANOVA, and the General Linear Model: A Statistics Primer*. Thousand Oaks: Sage Publications.

Ward, J. H. & Jennings, E. (1973). *Introduction to Linear Models*. Englewood Cliffs: Prentice-Hall.