Teaching Excel the General Linear Model for Student Access and Accuracy

John D. Morris **Mary Lieberman**

Florida Atlantic University

Due to the inaccessibility of university labs during the Pandemic, we created a powerful Excel program that performs most General Linear Model analyses, allowing delivery of a statistical methods course, beginning or more advanced, with Excel. All students are comfortable with and have immediate access to Excel. We also took this opportunity to embed contextual guidance toward correct understanding of analyses and to point out where commercial packages err. To promote use of quantitative methods, all relevant output is automatically generated in "camera-ready" APA cut-and-paste tables. We continue to develop the software and find that students gain more from it than the commercial package we introduce. We offer its availability, our experiences, and a student evaluation here.

There are issues in commercial packages that we would like to remedy. Other than the ilk, these are too numerous to enumerate here. Some are clerical but lead to students clinging to software lingo rather than understanding concepts; an example of that regards portraying a probability as "Sig" (SPSS) rather than "p," and in turn, rendering a "Sig" of ".000," that students then often convert to "p<.000." Other than wryly noting that a negative probability is **really** significant, we would like to avoid such problems. Some issues also regard expediency at the expense of understanding.

Commercial software, to make analyses automated and thus "easy," avoids many important decision points for users. An example would be, upon encountering a matrix that is singular, when an inverse or eigenanalysis is required, such software arbitrarily reduces the rank of the matrix by eliminating a variable without consulting the user regarding what is almost always a conceptual mistake. We believe that such decisions must be brought to the user for action before results are permitted. In our experience, the common practice in commercial software of eliminating a variable with zero variance is also a mistake for the same reason. The student needs to be confronted with the situation to net understanding of the problem and then be allowed to seek a solution. "**You have posed a problem that can't be answered; until you fix the data, no results will be rendered.**" Is our stand. This does make an extra step in respect to commercial software taking automatic action to "fix" the problem, but ease of use at the expense of understanding is unsatisfactory.

Programming Objectives

We also wanted control over what is displayed and how; we want to embed contextual instruction with what the student reads in the "output" so that we can educate while students calculate. Also, we sought to do a better job at instructional integration of methods than do typical software packages that compartmentalize analyses, making no attempt to integrate procedures that are inherently mathematically linked. Herein, that is least-squares and maximum likelihood. As well, we wanted students to gain some understanding of the power of resampling, and to get some notion of the potential losses in the use of least-squares or maximum likelihood by considering their deficits and some alternatives.

We found that we could introduce students to resampling herein through Allen's (1971) PRESS estimator and Lachenbruch's (1967) U or equivalently Huberty's (1994) **L**eave-**O**ne-**O**ut (L-O-O) classification estimator. Alternatives to our mathematical models could be illustrated with ridge-regression as an alternative to least-squares, and for logistic regression thus maximum likelihood, Firth's (1993) penalized logistic regression as an alternative.

We are interested in "maximizing" the likelihood of students using quantitative methods in their research, including in a dissertation. Therefore, it was decided that all production output had to be automatically produced as APA "camera-ready." This then means that the APA table format must automatically "wrap-around" differing numbers of variables. All such tables are produced as "**Table 1**" – the student can change the numeral as appropriate😊.

The Spreadsheet -- Criterion

The Excel Workbook is divided into various spreadsheet "tabs" that execute various desired analyses; each with extensive annotations embedded that are used to teach in context and to provide clarification regarding often encountered interpretive errors. As so many theoretical linkages are available therein, this

Excel spreadsheet is focused on relating predictors to a criterion and is thus named "Criterion." A **Read Me** sheet regarding goals, use, and the various other tabs used to access different sheets is first (Figure 1).

Some salient analysis sheets will be displayed in Figures here with two data sets. A multiple regression data set (Kerlinger & Pedhazur, 1973, p. 292), and a group contrast/classification data set (Morris, Ehren, & Lenz, 1991). Detailing all of these analyses will simply not fit in a proposal.

Input (Figure 2): To render the analysis, one only needs to enter the 1) N, 2) p (the number of variables), 3) α (we believe it is important for students to make a choice), 4) variable names, and 5) data (with the **C**riterion first in column **C** as shown). This is the first data set wherein GPA is predicted from GRE-Q, GRE-V, MAT and Average Faculty Rating.

APA R (Figure 3) and **APA R Bon** (Figure 4): The APA style **R** matrix is under the "**APA R**" sheet. The "**APA R Bon**" sheet offers the same matrix but with a Bonferroni type correction for per-hypothesis error, with a description of alternatives to the conservative Bonferroni.

Scatterplot (Figure 5): Any two variables may be selected by inserting their ordinal number into cells B8 (X) and B9 (Y), with those numbers included in a legend on that page; a scatterplot as well as histograms and comprehensive univariate statistics is automatically rendered.

APA Model (Figure 6): An APA table including the usual statistics for a regression model, with the addition of Cohen's (1988) f^2 effect size is here. As it is often seen as appropriate to include the model information in text rather than in a table, an alternative suggested skeleton sentence with data specific to the analysis is also included in **bold**.

APA Coefficients (Figure 7): A typical APA table of results for individual coefficients is included. We also have tried to ameliorate the seemingly continual misinterpretation of tests of partial slope inappropriately as bivariate tests of predictor-criterion relationship by including the ΔR^2 due to each variable, pointing out that, in the sense of relationship, this is what is being tested.

APA Generalizability (Figure 8): There are many details discussed here, but our fundamental purpose is to help students understand the distinction between the modeling goals of **explanation** (estimation of the population ρ^2 from the sample R^2) and **prediction** (estimation of the cross-validated correlation between the predicted and actual criterion, ρ_{cv}^2, in the population).

R^2 Increment (Figure 9): Here, the user can select which variables form a subset for which a test of contribution to the full model R^2 is sought including an "X" in the row just below the listed variable names. Then, the analysis will be produced and change as each new X is entered or removed.

Moderation (Figure 10): Moderation tests are accommodated with constituent predictor variables (except dichotomous) centered before analysis. The same variable selection method (designating in the row below the predictor variable names) for model testing is used, but in this case "X" selects a variable and "M" selects a moderator. For significant effects, "split case" plots and correlations are rendered.

Mediation (Figure 11): Using the same mechanism for designation of variable roles as X or M**,** Sobel's test with an SE consistent with Aroian, 1944/1947, is used.

Ridge (Figure 12, 13, and 14): For k [0,1], ridge weights ($\boldsymbol{\beta}$*), the squared correlation of the ridge predicted score with the criterion [$R^2_{y.\hat{y}(ridge)}$], the variance of those weights [Var($\boldsymbol{\beta}$*)], and the correlation of those weights to $\mathbf{r_{xy}}$ [r($\boldsymbol{\beta}$*,$\mathbf{r_{xy}}$)] are included.

Data Set 2 (Figure 15): The remainder of the sheets discussed herein are only triggered when a dichotomous criterion is entered; the program can detect that. This is the second data set in which dropout is predicted from 8^{th} grade data is included.

MANDDA (Figure 16): This sheet is named such as it has **MAN**ova as well as **D**escriptive **D**iscriminant **A**nalysis (**DDA**, Huberty, 1994) results. To the left are the multivariate effect test via usual methods (Hotelling T^2, Wilks' Λ, Bartlett-Pillai, and Roy), as well as Box's test of homogeneity of covariance matrices. As well, all known multivariate effect sizes are included. To the right are univariate tests with HOV tests and highlighting to show students which t-test to use (Gosset or Welch) and bolding of significant effects (in respect to the α on the **Input** sheet).

PDA Casewise (Figure 17): The intent here is prediction, thus "**P**redictive **D**iscriminant **A**nalysis" (Huberty, 1994) is used, and several types of linear models (Regression, Regression PRESS cross-validated, PDA, PDA L-O-O cross-validated, and PDA L-O-O cross-validated with priors and unequal costs of misclassification) are used to classify subjects including the ability to manipulate prior probabilities and costs of misclassification "real-time."

PDA Summary (Figure 18): Tests of hit-rate accuracy in respect to both the proportional and maximum chance criterion and the "I" index (Huberty, 1994) for all classification models included on the **PDA Casewise** sheet are here.

PDA Increment (Figure 19): The same increment to prediction accuracy question addressed in the **R^2 Increment** page is accomplished here in respect to classification accuracy with the method suggested by Morris and Huberty (1995).

LR (Figure 20): We now change to the maximum likelihood mathematical model by applying Logistic Regression to this same two-group data. Newton-Raphson iteration with input tolerance (default = .001) for weight convergence is used. Unlike all commercial packages, with the exception of Stata, no answer is given if convergence is impossible (complete or quasi-compete separation). All usual tests and Psuedo-R^2s are given, as well as classification accuracy.

LR Firth (Figure 21): Firth's (1993) penalized logistic regression is accomplished. As maximum likelihood is a large sample technique, Firth's method is intended to aid ML with small N. A side benefit is that iteration from Firth's LR version always converges, regardless of complete or quasi-complete separation.

LR Increment (Figure 22): The same Full vs. Restricted model test of classification accuracy increment as done in **PDA Increment** is accomplished here but using logistic regression for the two models. A more traditional test of the difference in Log Likelihoods between models is also provided.

PDA LR Comp (Figure 23): The same correlated proportions test between hit rates as in the **PDA Increment** page is conducted here, but between mathematical models (PDA vs. LR) rather than Full and Restricted models.

References (Figure 24): All relevant references are provided for students in Criterion.

Significance

In a recent multivariate statistics class that had students from Business, Psychology, Environmental Sciences and Education, a five item Likert-type scale ("5" was the most favorable opinion) was distributed through Google-Forms for anonymous response. It asked students to rate Criterion's 1) ease of use, 2) ease of interpretation of results, 3) ease of export of results, 4) protection from statistical errors, and 5) comprehensiveness. As well students were asked to provide narrative comments. Fourteen of 19 students responded.

For all items, responses were very positive. For Ease of Use all students responded "5." For ease of interpretation, protection from statistical errors, and comprehensiveness, all students except one responded a "5" with the remaining student responding "4." For ease of export, all students responded "5" except for one who responded "4" and one who responded "2." In the latter case, we are confident that the issue regarding use of a web interface that we subsequently phased out in favor of using OneDrive.

The narrative comments were even more informative; the program was very helpful for students in accomplishing assignments, but even more importantly, in understanding statistical concepts. See Table 1 for all comments (with identifying narrative deleted from one response).

If you care to take a look, an anonymized "voiceless" secure video demonstration of the analysis of these data with Criterion is here in a Google Drive account (make your browser full-screen and play in 1080p, or download, else the image may be blurry):
https://drive.google.com/file/d/1_29j0vyIF0dx3iBM_dpkml01awY-HxOi/view?usp=sharing

References

Allen, D. M. (1971). *The prediction sum of squares as a criterion for selecting predictor variables*. Technical Report *23*, Dept. of Statistics, University of Kentucky.

Aroian, L. A. (1944/1947). The probability function of the product of two normally distributed variables. *Annals of Mathematical Statistics*, *18*, 265-271.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173-1182.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society*, *4*, 268-282.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcola delle probabilita, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Browne, M. W. (1975). Predictive validity of a linear regression equation. *British Journal of Mathematical and Statistical Psychology*, *28*, 70-87.

Cohen J. E. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, *69*, 161-82.

Darlington, R.B. (1978). Reduced variance regression. *Psychological Bulletin, 85*, 1238-1255.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, *69*, 161-82.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association, 70*, 892-898.

Ezekiel, M. (1930). *Methods of correlational analysis*. New York: Wiley.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*, 27-38.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179-188.

Hayes, A. F. (2013). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. New York: Guilford Press.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55-67.

Huberty, C. J, (1994). *Applied discriminant analysis*. New York: Wiley.

Huberty, C. J, & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin*, *105*, 302-308.

Huberty, C. J, & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. *Educational and Psychological Measurement*, *40*, 101-112.

Jeffreys, H. (1946). An invariant from for the prior probability in estimation problems. *Proc. R. Soc. A*, *186*, 453-462.

Kerlinger, F. N. & Pedhazur, E. J. (1973). *Multiple regression in behavioral research*. New York: Holt.

Lachenbruch, P. A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics*, *23*, 639-645.

Lawless, J. F., & Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics, A5*, 307-323.

Longley, J. W. (1967). An appraisal of least-squares programs from the point of view of the user. *Journal of the American Statistical Association, 62*, 819–841.

Lord, F. M. (1950). *Efficiency of prediction when a regression equation from one sample is used in a new sample*. (Research Bulletin No. 50-40) Princeton, N. J.: Educational Testing Service.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58*, 593-614.

McNemar, Q. (1947).Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrika*, *12*, 153-157.

Morris, J. D., & Lieberman, M. G. (2021, April). *Pandemic Package for the Social Sciences – An Ongoing Effort Toward Access and Accuracy*. Paper presented at the annual meeting of the American Educational Research Association, Virtual Conference.

Morris, J. D., & Lieberman, M. G. (2018). Multicollinearity's effect on regression prediction accuracy with real data structures. *General Linear Model Journal*, *44*, 29-34.

Morris, J. D., & Lieberman, M. G. (2015). Prediction, explanation, multicollinearity, and validity concentration in multiple regression. *General Linear Model Journal*, *41*, 29-35.

Morris, J. D., & Huberty, C. J (1995). Full versus restricted model testing in predictive discriminant analysis. *Journal of Experimental Education*, *63*, 161-165.

Morris, J. D. (1986). Selecting a predictor weighting method by PRESS. *Educational and Psychological Measurement*, *46*, 853-870.

Morris, J. D. (1981). Updating the criterion for regression predictor variable selection. *Educational and Psychological Measurement*, *41*, 777-780.

Morris, J. D., Ehren, B. J., & Lenz, B. K. (1991). Building a model to predict which fourth through eighth graders will drop out in high school. *Journal of Experimental Education, 59,* 286-293.

Nicholson, G. E. (1948). The application of a regression equation to a new sample. Unpublished doctoral dissertation, University of North Carolina. [Condensed in G. E. Nicholson, Prediction in future samples. In I. Olkin et al, (Eds.), *Contributions to probability and statistics*. Stanford: Stanford University Press, 1960.]

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, *29*, 201-211.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*, 879-891.

Quirk, T. J. (2011). Excel 2010 for Educational and Psychological Statistics: A guide to solving practical problems. Springer, NY.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlation: A clarification. *Psychological Bulletin*, *85*,1348-1351.

Sobel, M. E. (1982). Asymptotic intervals for indirect effects in structural equations models. In S. Leinhart (Ed.), *Sociological Methodology 1982* (pp.290-312). San Francisco: Jossey-Bass.

Stein, C. (1960). Multiple regression. In I. Olkin et al. (Eds.), *Contributions to probability and statistics*. Stanford: Stanford University Press.

Student. (1908). The Probable Error of a Mean. *Biometrika*, *6*, 1–25.

Welch, B. L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika***,** *34***,** 28–35.

Wherry, R. J. (1975). Underprediction from overfitting: 45 years of shrinkage. *Personnel Psychology*, 28,1-18.

Wherry, (1975). Underprediction from overfitting: 45 years of shrinkage. *Personnel Psychology*, *28*,1-18.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics*, *2*, 440-451.

Send correspondence to:	Mary Lieberman Florida Atlantic University Email: mlieberm@fau.edu