

# The Secrets Behind Producing Meaningful Quantitative Research: What Every Foundation Official Really Needs to Know

Fred J. Galloway  
University of San Diego

January 2004

## Introduction

For the last thirty years, the rapid growth in information technologies has led to an absolute explosion in the production and diffusion of statistical software. With drop-down menus replacing meticulously written computer code, it now seems that literally *anyone* can estimate a series of hierarchical regression models or structurally decompose a time series. Despite the fact that a nuanced understanding of techniques like these takes years of theoretical study and empirical practice, aspiring quantitative researchers can now perform these procedures in a matter of minutes on such popular types of statistical software as SAS and SPSS. As a result, there has been a growing mismatch between what many quantitative researchers are *able to do* and what they are *able to understand*.

Although this growing mismatch has been felt in such social scientific disciplines as political science, anthropology, and sociology, nowhere has the effect been more pronounced than in the field of educational research, which has traditionally produced some of the most poorly designed quantitative research studies of all. While some researchers have argued that this is due to the messy nature of educational research itself, others have argued that the transition from classroom teacher to quantitative researcher is inherently problematic and simply too difficult to be left to most schools of education. As noted Ed School scholar David Labaree argues in the most recent edition of *Educational Researcher*:

“Under such circumstances of great complexity, vast scale, uncertain purpose, and open choice, researchers are unlikely to establish valid and reliable causal claims that can be extended beyond the particulars of time, place and person. As a result, research claims in education tend to be mushy, highly contingent, and heavily qualified, and the focus is frequently more on description and interpretation than on causation.” (Labaree, 2003)

Of course, from the perspective of a foundation that actively supports research in higher education, Labaree’s words provide little reassurance that their research dollars will be spent in ways that actively contribute to the policy debate surrounding issues of access and persistence. Since every foundation, including Lumina Foundation for Education, seeks to extract the “biggest bang” from their limited research dollars, this goal of this paper is to help Foundation officials distinguish between quantitative research that offers inferentially rich insights and policy relevant conclusions and the sort of quantitative research that, because of an flawed design or inappropriately applied analytical

technique, actually detracts from the body of knowledge with its analytically unsupported conclusions. This will be first accomplished by discussing the basics of research design and causality, followed by an overview of the databases that support them. After setting out the basics of research design, the next section of the paper will concentrate on the most widely used analytical technique in quantitative analysis, multiple regression analysis, and will help Foundation officials understand why this technique is so often used *and* misused. A brief discussion of the most common types of data problems and their solutions follows, and the paper then concludes with some insight into how the results of the regression models can be presented in a clear and insightful manner.

## **Research Design and Causality**

The most important part of any quantitative study involves the research design and methodology used by the researcher. Of course, if the research design is solid and the correct analytical techniques used, then the researcher has the ability to produce research that is inferentially robust and generalizable. But what exactly constitutes a solid research design? And what about making casual inferences? Although both of these questions will be addressed in the next few paragraphs, the theme of this section can be stated quite simply – what is bungled by design *cannot* be fixed by analysis. In other words, unless the correct research design is selected for the study, even the most sophisticated analytical techniques will not produce meaningful results. With this important thought in mind, the discussion now turns to the relationship between design and causality.

As mentioned above, the process of selecting a particular research design for a project has profound implications for both causality and generalizability, and the design choice should not be taken lightly. In fact, it is often said by methodologists that causality is purely a design issue, so that the ability to make casual inferences is essentially determined before any of the analysis is done. And interestingly enough, there are actually various “shades” of causality, ranging from the sort of pure causality that might emerge from a study of identical twins or triplets where siblings receive different treatment to the somewhat weaker type of causality that emerges from studies using randomized assignment. However, these forms of causality should not be confused with the sort of pseudo-causality that emerges from correlational studies, where two variables are strongly related, and the researcher attempts to portray one in a causal manner, which is, of course, fallacious.<sup>1</sup>

To help understand why there are actually various degrees of causality, it may prove useful to understand why the purest form of causality emerges from studies of identical twins or triplets. Since each set of individuals share exactly the same genetic material, and presumably, were raised under similar environmental conditions, these individuals are so alike that researchers can administer a particular treatment to one of the identical siblings and not the other, and then attribute any difference in outcomes to the treatment itself. Of course, the more differences there are between the individuals *before* the treatment is administered, the more difficult it is to make a causal inference regarding the treatment since any of the other differences between the individuals might be causing the observed outcome. In fact, this notion of attempting to hold everything else but the

treatment constant between individuals is at the heart of what it means to assign causality to a particular event or treatment.<sup>2</sup> In other words, the more demographic and environmental things that can be controlled for between groups (as with identical twins or triplets), the stronger the type of causality that emerges.

As such, the notion of random assignment, where individuals are randomly assigned to two groups – typically called the control and treatment group – is viewed as the next strongest form of causality. Although the strength of the causality that emerges from random assignment is determined by a number of things, the most important is the actual level of randomization. For example, if the names of all study participants are arranged alphabetically and every other person is assigned to the treatment group, this might appear to be perfectly random, but what if as a result of this process, the treatment group ends up being 25 percent female and the control group 75 percent female? Does this mean that any sort of causative statement derived from the analysis is less valid? Of course, the answer depends on whether the outcomes of interest are functionally related to gender, but if the two groups had been perfectly matched in terms of gender, the causative statements that emerged from the analysis would have been stronger than with the non-matched groups.

For this reason, the actual degree of randomization that occurs helps determine the strength of the causative statement that emerges from the design. However, because it is impossible to achieve perfect randomization for all the factors that might be related to the measured outcome, this form of causality is always tempered, and as a result, it usually takes a number of studies uncovering the same finding before the result is accepted as fact. In fact, randomization is really only a necessary condition for causality, since often times other factors can contaminate the study, like the physician that knows which patients are receiving the real medicine and which ones are receiving placebos, and unknowingly tips off the patients, which in turn, effects the outcome for several of the patients. Although outcomes like this can be solved by using double-blind studies where neither the physician nor the patient knows who is getting the real medicine, researchers clearly need to pay careful attention to all facets of design to strengthen the type of causality that emerges from their controlled experiment.

This notion of attempting to control for differences among participants in a study extends to non-experimental studies as well. In fact, most of the quantitative research in education uses non-experimental data, and to complicate matters, researchers are often forced to work with data that was gathered by others for expressly different purposes. However, when this occurs, researchers do have the ability to use statistical techniques that in essence, attempt to control for differences among individuals so that researchers can tease out the effects of a single variable on the outcome of interest while holding the effects of all the other variables in the model constant. Although a variety of techniques are used to accomplish this task, multiple regression analysis is perhaps the most popular, and while causality cannot be implied with non-experimental data, at least the effects of a particular variable on the outcome of interest can be estimated, and the degree of error associated with this estimate measured as well.

Although many of the issues surrounding multiple regression analysis will be discussed later in this paper, it is again important to note that unless experimental data is used, causality cannot be inferred from the results of any regression analysis. Despite the fact that almost all of the data gathered in postsecondary education is non-experimental, there are enough important things that can be done with multiple regression analysis to warrant a thorough understanding of the strengths and weaknesses of this popular methodology. However, before beginning this discussion, the next section presents an overview of the different types of databases that can be used with this popular technique.

### **Types of Databases Available in Higher Education**

To make sense of the different types of databases available for quantitative research in higher education, it's helpful to recall the difference between a population and a sample, since databases can appear in either form. When a database contains *every* single element in a group, like all of the students admitted to a school, or all the student loans held by a particular organization, then the database is said to contain the entire population. Examples of databases like this in postsecondary education are typically institutional -- like all those students admitted to the University of San Diego for academic year 2002-03, or all of those students whose loans are currently held by Sallie Mae. Databases like this can be contrasted with those that contain only samples taken from the population, like the National Postsecondary Student Aid Survey (NPSAS), which samples both schools and students every five years, or the National Longitudinal Study of Youth, which sampled 12,686 students between the ages of 14 and 22 in 1979, and then conducted annual follow-up interviews for almost fifteen years.

Since databases that contain samples drawn from populations are the most common type of databases in education, a further distinction needs to be made in how the samples are drawn. With *probability samples*, the probability that each member, or element of the population, is included in the sample is known by the researcher, where in *judgment samples*, personal judgment plays a major role in determining which elements of the population are selected, and this probability is not known. For example, if a list of every Historically Black College and University (HBCU) were compiled and every other one selected, then this would be a probability sample, with the probability of selection given by .5. However, if another sample was selected based solely on whether the researcher had personally visited a particular HBCU, then this would be a judgment sample and the probability of selection would be unknown.

The distinction between probability and judgment samples is an important one because researchers working with probability samples have the ability to estimate how large their sampling error is, however with judgment samples there is no way to tell how "far off" a sample result is likely to be from the true population value. For this reason, most databases are constructed from probability samples, and estimates of the sampling error provided, making it easier to discuss how much confidence the researcher has that the results they found in the sample also occur in the population. This ability to generalize from the sample to the population is perhaps the most fundamental and necessary condition for any well-designed quantitative study, and all researchers should be willing

to address this issue in the methodology section of their proposal. Since this condition can never be met by judgment samples, they are to be avoided unless both financial and logistical concerns necessitate their use.

Although there are many types of probability samples, including stratified random samples, systematic samples, and cluster samples, all of them are characterized by an element of randomness in the actual selection of the sample. Whether the randomness appears as selecting every *k*th element on a list of all elements in the population, or simply dividing a population into strata and then taking a random sample from each of the strata, the notion of random selection is what probability samples are all about. And since the probability of selection in probability samples is always known, researchers have the ability to “weight up” the responses from individuals in the sample to represent the entire population on such key dimensions as gender, race/ethnicity, income, and location.

Although discussed in more detail in the companion paper, the notion that weighting can be used to match samples demographically to their respective populations is a powerful analytic tool that can help researchers make more informed judgments about how a population behaves. Since this sort of weighting is impossible without knowing the probability of selection, probability samples are to be used wherever possible. Of course, there may be times when it is impractical or impossible to completely describe a particular population, as might be the case when studying a Tribal College that can't provide an accurate enrollment count because of insufficient or outdated records, or a proprietary school that might not wish to disclose enrollment information for a variety of reasons. Although such exceptions do occur, whenever possible, quantitative research proposals should be based on databases formed from entire populations or probability samples drawn from populations.

From this class of admissible databases, a further distinction can be made based on whether the researcher is looking at things at one point in time or looking at things over time. If the former is true, the database is said to be *cross-sectional* in that the research focus is on looking at things at one particular point in time, while if the latter is true then the database is said to be *time-series* in that it is looking at how things have changed over time. If both conditions are true, the database is said to be *panel* or *longitudinal*, in that it may contain observations on thousands of individuals, each observed at numerous points in time. For example, the 1993-94 NPSAS database is cross-sectional in that it looks at sources of student financial support at one point in time, that being academic year 1993-94, but when the degree completers in this database were then followed over time, the resulting database, *Baccalaureate & Beyond*, became longitudinal. This can be contrasted with a pure time-series database such as the number of undergraduates admitted each year to Harvard College over the last century, where one can plainly see that the same individuals are not being followed over time. Although all three types of databases have their place in empirical research, the longitudinal databases have the greatest potential for powerful inferences, since they provide variation both at individual points in time, as well as over time.

Of course, no matter what type of database is used in the analysis, researchers need to explicitly specify their research questions and choice of analytical techniques in their proposal. Although there are a number of important analytical techniques available to quantitative researchers in higher education, multiple regression analysis is perhaps the most general since it can be applied towards a wide range of problems, ranging from those involving the degree of relationship among variables to those involving the search for some underlying structure or the prediction of group membership. To provide Foundation officials with an overview of this important methodology, the next section of this paper will concentrate on the ways in which multiple regression analysis can be used to produce quantitative research that is both methodologically sound and generalizable, helping to inform the work of state, federal, and institutional policymakers.

### **Multiple Regression Analysis**

Although regression analysis can be used to address several different types of research questions, its most common application is to questions involving the degree of relationship among variables. As such, the emphasis in this section will be on how regression analysis can be used to estimate the degree of relationship among variables, although several paragraphs will be devoted to how this popular technique can also be used to answer questions involving the search for underlying structure and the prediction of group membership. The discussion begins with perhaps the two most popular tools of regression analysis, bivariate and multivariate correlation, followed by the techniques of multiple regression analysis, including hierarchical multiple regression analysis and logistical regression. Throughout the discussion, the focus will be on the methodological insights necessary to produce inferentially meaningful work, and should not be viewed as a substitute for a basic understanding of the underlying mathematical statistics.

In assessing the degree of relationship among variables, the most fundamental tool is the bivariate *correlation coefficient*, which measures the linear association between any two variables. This measure is unit-free and bounded between  $-1$  and  $+1$ , so that a correlation close to  $+1$  means that the two variables tend to move in the same direction together (i.e. both variables are either above their means, or below their means), while a correlation close to  $-1$  means that the variables tend to move in opposite directions (one is below its mean while the other above its mean). In a similar manner, the multivariate correlation coefficient measures the association between a dependent variable and an optimally weighted combination of two or more independent variables.<sup>3</sup> However, unlike the simple bivariate correlation coefficient, the multivariate correlation coefficient is only bounded between  $0$  and  $+1$ , where a value of zero means that no relationship exists with the independent variables and a value of  $+1$  indicating a perfect relationship.

Although correlation coefficients are helpful in assessing the degree of relationship among variables, they are less helpful when it comes to predicting values for the dependent variable or estimating the relative contribution of each of the independent variables in predicting the dependent variable. Fortunately, *multiple regression analysis* solves these problems by allowing researchers to select, or specify, a set of independent variables that the researcher believes may help explain why a particular dependent

variable behaves in the way that it does. For example, multiple regression analysis might be used to explain why some colleges have higher graduation rates than others, or why some students have higher grade point averages than others.

Although the mechanics of regression analysis combine the calculus of optimization theory with the power of inferential statistics, the basic notion behind regression analysis is that linear combinations of independent variables can be used to not only predict the dependent variable, but to explain the variance in the dependent variable as well. Of course, while developing these multiple regression models, researchers must be extremely careful in their selection of independent variables since the omission of any statistically relevant variable typically leads to biased estimates of the contribution that each independent variable makes to the dependent variable.<sup>4</sup> This problem, called “specification error”, is one of the most pervasive problems in regression analysis and often times occurs when researchers are forced to select their variables from data already gathered by others. To see how a problem like this might occur, consider the example of why some colleges have higher graduation rates than others and imagine a thoughtful researcher trying to decide what independent variables to include in their model. One can easily imagine selecting at least three sorts of variables -- those relating to the quality and preparation of the student body, those describing the way that institution behaves towards students, and finally those variables that describe the net price that students pay at their respective college.<sup>5</sup> Although the information for the first two types of variables might be readily available from a database like the Integrated Postsecondary Education Data System (IPEDS) or from institutional records, the third bit of information might be very difficult, or even impossible to locate, since many financial aid offices may not keep data in this form. Of course when this occurs, the researcher has no choice but to limit their modeling effort to including only the first two types of variables and acknowledge that their model may suffer from specification error.

Although specification error is one of the most common problems in regression analysis, there is a simple two-step method for dealing with it. First, when actually specifying the variables to be used in the model, researchers need to use theory to identify what sorts of effects might be present, and then every effort needs to be made to include all of the variables that represent these effects. Second, when data limitations prevent researchers from including all of the candidate variables, they need to explicitly state that several variables that might have proven significant were not included in the analysis, and to the extent that these missing variables are correlated with any of the variables in the model, their estimated coefficients may be biased. However, the specification of a model’s independent variables is also the place where, as magicians say, “the rabbit goes into the hat” in that the effects or variables that appear significant in a model are largely a function of the variables that were initially included for testing in the model. In fact, in his 1983 article, “Let’s Take the Con Out of Econometrics”, Edward Leamer argues that through judicious selection of a model’s candidate variables, researchers with very different backgrounds were all able to claim that their particular hypotheses are supported, despite the fact that in many cases these hypotheses were mutually exclusive. The bottom line is that unless researchers are both willing and able to discuss the

*inferential fragility* of their results, in other words, how sensitive their results are to the inclusion of other variables, their results should simply not be believed or accepted.

This notion of inferential fragility is indeed a powerful concept that applies not only to how models are specified, but also to the size of the effects produced by the variables in a regression model. Although many researchers spend lots of time talking about how statistically significant some of their variables are, the more interesting and relevant question is what sort of effect is produced by each of the significant variables in a particular model. Within this context, the notion of inferential robustness suggests that instead of just producing a point estimate of the effect size for each of the variables, a range of effect sizes that are associated with slightly different specifications of the regression model may be more appropriate. For example, if there are three potential ways of measuring a student's financial capital while in college (total income, disposable income, and parental income) in a particular regression model, the best way to gauge the size of the effects that *other* variables in the model may have on the dependent variable is to run three different regression models, each one containing a different measure of income. As the different models are run, of course, different effects are produced for all of the model's variables, and a range can be established that contains all three of the estimated effect sizes for a particular variable. As expected, the narrower the range for the effect sizes, the more stable, or robust, is the inference, while a wider range for the effect sizes suggests that the inference is more fragile and less likely to be believed.

By paying attention to the problems of specification error and inferential fragility, aspiring researchers can use the tools of multiple regression analysis to produce methodologically sound quantitative research. However, when attempting to disentangle the amount of variation in the dependent variable explained by the various types of independent variables in a particular model, *hierarchical multiple regression* analysis is more appropriate. In this type of analysis, variables are loosely grouped into several general categories and the contribution of each category of variables is then estimated by running a series of nested models, with each successive model adding another category of variables. For example, in our earlier example of why some colleges have higher graduation rates than others, recall that three sorts of variables were selected for analysis -- those relating to the quality and preparation of the student body, those describing the way that institutions behave towards students, and finally those variables that describe the net price that students pay at their respective college. Through the use of hierarchical multiple regression analysis, the contribution that each one of these three groups makes in helping to explain variation in graduation rates can be assessed by first regressing graduation rates against those variables relating to the quality and preparation of the student body, then running a second model where graduation rates are regressed against variables relating to the quality and preparation of the student body *and* variables describing the way that institutions behave towards students, and finally, running a third model where all three categories of variables are included. Methodologically, this allows the researcher to net out the effects of each of the three types of variables, so that the contribution that each of these categories makes can be easily estimated. And of course this can be critically important for policy matters, since if the quality and preparation of the student body explains twice as much of the variation in graduation rates as does the

way that institutions behave toward students, then increasing graduation rates can be as simple as increasing admission standards, rather than concentrating on ways in which the institution can provide more for their current students. For this reason, it's not surprising that hierarchical multiple regression analysis has become an increasingly popular tool for serious quantitative researchers.

In addition to using multiple regression analysis to examine the degree of relationship among variables, these techniques can be used in conjunction with select nonlinear techniques to predict group membership from a set of independent variables. For example, instead of examining why some colleges have higher graduation rates than others, researchers might want to focus on why some individual students graduate while others dropout. Although a number of other analytical techniques could be used to address this problem,<sup>6</sup> when the independent variables used in the analysis are a mix of continuous and discrete, then *logistic regression analysis* represents the best way to predict group membership. In essence, these models represent a hybrid of standard multiple regression analysis and the logit form of multiway frequency analysis, and provide estimates of the marginal contribution that each independent variable makes to the probability of membership in each group, which can be used to predict overall group membership for every individual in the database. These models can also be easily extended to more than two choices – like whether a student has graduated, is still in school, or dropped out. However, for those researchers that insist on using regular multiple regression analysis, a host of statistical problems awaits – including predictions that lie outside of the unit interval, the introduction of a non-constant error variance into the model, and a non-normal error terms that invalidates many common tests of statistical inference.

The techniques of multiple regression analysis and logistical regression analysis can also be combined with factor analysis for researchers interested in the search for some underlying structure in the data. These techniques, collectively known as *structural equation modeling*, are extremely useful when estimating the set of relationships between one or more independent variables and one or more dependent variables.<sup>7</sup> The generality of this collection of techniques has made structural equation modeling probably the hottest analytical technique of the past decade, resulting in some incredibly insightful pieces of quantitative research as well as some of the most poorly designed studies imaginable. To help distinguish between the really good and the really bad, remember that the techniques of structural equation modeling are confirmatory, rather than exploratory, and should only be used to test the appropriateness of various theories. So rather than searching for the model that seems to fit the data the best, researchers need to pay careful attention to what theory has to say about the relevancy of particular variables as well as the hypothesized relationships among the variables included in the model. This point is absolutely crucial, because in some circles, structural equation modeling has already developed a bad reputation as a result of researchers spending too much time using these techniques in an exploratory manner, typically testing a variety of different modeling specifications.

## Common Data Problems

Since data problems are a fact of life for those engaged in quantitative research, a necessary condition for producing meaningful research is that all data problems be handled in the most inferentially robust manner possible. In other words, this is another place where, as magicians often say, the rabbit goes in the hat, so Foundation officials are urged to pay close attention to the way in which each and every data problem is resolved. Although there are a number of different types of problems, perhaps the three most pervasive and serious facing quantitative researchers are handling missing data, dealing with unusual observations, and what to do when two or more variables in a database are highly correlated. To help Foundation officials understand how these problems can affect the generalizability of the results from a particular study, this section will provide an overview of the problems and their solutions beginning with several solutions to the missing data problem and their respective statistical properties.

Without a doubt, the problem of *missing data* is the most pervasive problem in all of quantitative research, and at one time or another, all researchers have been forced to deal with the issue. The problem can occur either systematically or randomly, and of course the former is much worse than the latter, since if the same survey question regarding household income is continually left blank by those with high levels of income, this may introduce a bias regarding the effect of income in your model, but if survey questions appear to have been left blank randomly, the only modeling cost is a loss of efficiency or precision.<sup>8</sup> In dealing with these sorts of problems for the independent variables in a multiple regression model, there are really three very different sorts of solutions, including dropping the missing observations, using the sample mean in place of the missing observations, and finally, using sophisticated techniques that involve using other models to forecast the missing observations. Of course, if the missing observations are for the dependent variable in the model, the researcher has no choice but to drop the missing observations from the analysis.

Of the three different correction procedures, the easiest and most convenient method is for researchers to simply *drop any observations that they may be missing data for*. Under this correction procedure, if the researcher were collecting thirty bits of data from each individual and if *any* one of the thirty bits were missing, then the individual would be completely dropped from the analysis. As mentioned in the preceding paragraph, if the data were missing randomly then the analysis would not be as precise as if you had all of the data; however, if the data were missing systematically, then the analysis would be biased. In terms of trading off between biasness and loss of precision, so long as you still have enough observations to provide relatively precise estimates, then a small loss in precision is much better than introducing a systematic bias into your analysis.

The second type of correction procedure is called the *zero-order correction procedure* and simply involves substituting the sample mean in for the missing observations. From an operational standpoint, this means calculating the sample mean for the variable with the missing observations, and then using it in place of the missing observations. Despite the seeming non-sophistication of this approach, this correction procedure has desirable

statistical properties in that it produces unbiased parameter estimates when using regression analysis with only a small loss in precision.<sup>9</sup> Unfortunately, the more variation in the data, the greater the loss in precision, so when researchers are facing cross-sectional databases with lots of variation, this technique may not be as favorable as the first-order correction procedure, discussed below.

The final procedure, known as the *first-order correction*, involves actually estimating other models to predict or forecast the missing observations. Under this procedure, the researcher looks for independent variables that are correlated with the variable that has missing observations but relatively uncorrelated with each other, and then a regression is run on all of the complete cases with the variable with the missing observations serving as the dependent variable, and the correlated variables serving as the independent variables. Sometimes, the predicted values from the first round are inserted in place of the missing observations and then all cases are used in a second regression. This process can then be repeated until the predicted values from each subsequent regression converge. Although this approach may indeed introduce a small amount of bias into the model, the gain in precision can be large, especially if a number of observations are missing and the regression used for forecasting solid in terms of goodness-of-fit.

In addition to the problem of missing observations, another pervasive data problem concerns what to do about unusual observations, often called *outliers*. These values typically occur for one of four general reasons; incorrect data entry; failure to correctly specify missing value codes so that the codes are incorrectly being read as data; a legitimate value for someone that should not have been sampled for the study (like a staff member at a college university that inadvertently fills out a faculty survey and lists their highest degree as associates); or a legitimate value that just seems unusual (like a 12 year-old college student). Although the first three of these cases are easy to deal with, the case of the 12 year-old college student presents real problems since it appears to be a legitimate observation and may have more influence on the results than appears to be warranted.

In fact, the presence of outliers can be a real problem in regression analysis since the typical regression line is fit by minimizing the sum of squared errors, which means that when an observation is far away from the other data points, its contribution to the regression line is huge because of the squared distance measure. Given the problems that outliers can cause, it comes as no surprise that several regression diagnostic measures have been invented that can help in the identification of outliers, and serious researchers need to investigate these measures thoroughly before discussing the robustness of their results. These diagnostic measures are called *case statistics*, meaning that there will be one value for each of these statistics for each observation in the database.

The first of these case statistics is called *leverage*, and measures how unusual the case is in terms of its values on all of the independent variables. In other words, this measure gets at how far the observed values for each case are from the mean values on the independent variables. The second type of case statistic is called *discrepancy*, and is a measure of the squared distance between the predicted and observed values for the

dependent variable. The third and final case statistic is *influence*, which reflects the amount that the regression coefficients would change if the outlier were removed from the database. For large databases where visual inspection of the data is impossible, researchers are urged to calculate all three case statistics, order them from lowest to highest, and then print out the highest values for examination.

In addition to the use of case statistics, examination of frequency distributions and boxplots may also be helpful in identifying the presence of outliers. However, once an outlier is identified, a determination needs to be made as to whether the outlier was caused by some sort of error in the data recording process, or whether the value is legitimate and needs to be dealt with. Of course, if the former is true then the observation can simply be corrected or dropped from the analysis, but if the latter is true then there are a limited number of possible solutions. The most popular of these is to redefine the sample, if possible. For example, the presence of a 12 year-old college student in a sample of undergraduates might not be a problem if the researcher redefined the sample to include only those undergraduates that are at least 18 years old. However, if redefining the sample is not an option, then the researcher might consider using some non-linear modeling specification so that the regression line “naturally bends” toward the outlier, rather than pulling the entire regression line towards it, or perhaps even using a least absolute deviation estimator instead of the ones produced through minimizing the sum of squared errors. Regardless of the technique used, researchers need to pay careful attention to the presence of outliers since they have the ability to significantly distort the results of any quantitative study, in effect, limiting the power and generalizability of the study.

In addition to the problems that arise from specific cases within the database, problems can also occur between within specific independent variables, most notably the problem of *multicollinearity*, defined as a strong correlation between two or more independent variables. To understand why multicollinearity is a problem, recall that in regression analysis the researcher is interested in the unique, individual contribution that each particular independent variable makes to the dependent variable, and this contribution can be perfectly measured when the independent variables are completely uncorrelated, or orthogonal, with each other. In fact, when we interpret regression coefficients we always have to add the caveat, “holding the other variables constant”, which of course cannot be done if a particular independent variable is correlated with another. However, since all independent variables in a model are at least somewhat correlated with each other, the problem of multicollinearity is one of degree, and as such there is considerable debate about exactly when it becomes a problem.

Although there may be some debate as to when multicollinearity becomes a problem, fortunately there is no debate over the consequences of multicollinearity. When multicollinearity occurs between two or more variables, all of the estimated coefficients for these variables have larger standard errors than they otherwise would, which reduces the value of their t-stats and makes it less likely that the researcher will find these variables significant in explaining variation in the model’s dependent variable. The two most common ways to test for this problem are the simple bivariate correlation coefficient, which was discussed earlier in this paper, and a measure called the *variance*

*inflation factor* (VIF), which provides an estimate of how much smaller the variance of each of the model's estimated coefficients would be if all of the variables were perfectly uncorrelated. Although there are no precise statistical tests associated with either of these measures, the rule of thumb for the VIF is that any value over 10 provides evidence of serious multicollinearity.

When multicollinearity does occur, there are several accepted methods for dealing with the problem. The simplest and most popular method is simply to drop the offending variable or variables from the model, providing of course that the researcher is not introducing any bias into the analysis by dropping a theoretically important variable. Another simple method is simply to collect additional data, since the larger the sample size the smaller the variance surrounding the estimated regression coefficients. When neither of these techniques can be used, researchers have the option of using the method of principal components to combine all of the multicollinear variables into one or two super variables, which by construction will be perfectly uncorrelated with each other. However, when there is an extremely high amount of correlation within a dataset, the researcher may feel the need to turn towards ridge regression, which decreases the standard errors surrounding the regression coefficients by adding a constant to the variance of each independent variable. Taken together, these four correction techniques offer researchers a choice of analytical procedures that can be used to increase the precision and generality of their estimates.

### **Content and Presentation Issues**

Although this section may seem like an afterthought, in many respects it is the most important part in the paper since methodologically sound educational research typically has policy implications, and if policy-makers can't understand the research themselves, then its value is significantly diminished. To help reduce this problem, this section will first discuss what types of information need to be presented so that the results of the regression analysis are complete and transparent. After Foundation officials know *what* information should be presented, the final part of this section will concentrate on *how* the information should be presented so that both practitioners and scholars can easily follow the story line and the results.

The question of exactly what to report from an analysis has for years confounded many quantitative researchers in education. On the one hand, some researchers feel that every single exploratory model must be reported on, while others are content to report only on their final regression model. Clearly, the solution lies somewhere in the vast middle, but unless the researcher has discussed the inferential robustness of their results, then as Leamer (1983) argues, "all concepts of traditional theory utterly lose their meaning by the time an applied researcher pulls from the bramble of computer output the one thorn of a model he likes best, the one he chooses to portray as a rose...the consuming public is hardly fooled by this chicanery".

Fortunately, the solution to the problem of inferential robustness is rather simple to implement, although for many researchers it requires a major attitudinal shift. Instead of

focusing on a single model, researchers need to use theory to specify sets or families of models that incrementally use different measures and different specifications, including linear and nonlinear ones, to estimate a range of effect sizes. In this manner, rather than producing point estimates for each of the variables in a single model, researchers can report a range of effect sizes for each of the categories of variables used in the family of models. For example, if theory suggests that there should be an income effect in a particular model, and if there are three compelling ways of measuring income, then all three should be used separately and a range of effects reported, not just for the income variable but for the other variables in the model as well. In this manner, the robustness of any inference regarding income can be easily seen, and if it turns out that only one measure of income was significant in only one model, then the inference is not robust but fragile, and any generalizations inappropriate. However, if all three measures were significant in almost every model run then the researcher can have confidence in the existence of an income effect, and the size of the effect reported in a range, rather than as a point estimate.

In addition to reporting a range of effect sizes, the notion of inferential robustness also extends to the way in which researchers handle some of the data problems discussed earlier in this paper. For example, if missing data is a serious problem then researchers might want to try several of the correction procedures to see if their results vary significantly across the different procedures. And of course if they do, then their results are clearly a function of the particular correction technique used and as such, are not generalizable. In much the same way, instead of just trying one solution to the outlier or multicollinearity problem, researchers are urged to explore the full range of analytical solutions and then report on the range of results. In this manner, both researchers *and* Foundation officials can see how sensitive the results of a study are to slightly different correction procedures and modeling specifications, making it ultimately easier for decision-makers to base policy on.

If these simple suggestions are followed, methodologists and other researchers will have an easy time making sense of the importance of a particular piece of research. And when researchers combine the notion of inferential robustness with a sound design and correct set of analytical techniques, the resulting research has the potential to influence policymakers at all levels, provided of course that they can *understand* the research that has been presented. To help researchers and Foundation officials in this regard, the discussion now turns to the question of *how* to present quantitative research.

Although many educational researchers may feel that they already know how to present their research in a clear and insightful manner, there are some informal, general rules worth reviewing since the stakes are clearly high for both researchers and Foundation officials. Since these rules involve the substance *and* flow of printed information, the discussion will begin with the flow of information and then move to the specifics of such things as using tables and reporting modeling results. However, before turning to the flow of information there is one over-arching rule for all researchers to remember – the importance of writing in a clear and jargon-free manner. Although at one level this may seem obvious, experienced researchers know that it takes real work to express

sophisticated analytical concepts in plain, simple English, but for those that do, the dissemination rewards are great.

In addition to writing in a clear and jargon-free manner, the paper must also flow logically and intuitively from beginning to end. To ensure that this happens, researchers are urged to begin with an introductory section that provides a general overview of the issues surrounding the study and demonstrates the significance of the problem. This section should then be followed by a review of the literature that effectively grounds the study in a larger body of research and provides a clear rationale for the study. The researcher can then articulate the study's research questions and choice of analytical techniques, being sure to discuss exactly why a particular analytical technique was chosen. And of course, if the researcher is using data gathered from a sample, all of the particulars of the sample and the instrument need to be discussed as well. In the results section that then follows, researchers can discuss the inferential robustness of their findings, being sure to focus on the generalizability of their results so that Foundation officials and policymakers will understand exactly what message they can take away from the study. For that matter, researchers are urged to address any policy implications from their study directly in the last section of their paper where they offer their conclusions and suggestions for future research.

By following this general format, which of course parallels the format for a quantitative dissertation, researchers can be assured of a logical flow for their work. However, when working with families or sets of models rather than a single model, legitimate questions arise about using tables and reporting modeling results. Fortunately, the answers are fairly simple as well as intuitive and suggest that instead of simply providing point estimates for everything from estimated coefficients to goodness-of-fit measures like R-squared and adjusted R-squared, intervals or ranges should be provided that contain the results of each successive family of models. In fact, if separate tables like this are constructed for every set of models, then several "super" tables can be constructed that effectively roll-up these individual tables into summaries of all the sets of models, so that readers can get a visual sense of how generalizable the results truly are.

Although these tables may take some time to design and construct, Foundation officials can encourage researchers to think carefully about the exact order that the tables will appear in *before writing*, so that the results section has been already organized by table before any writing begins. That way, the layout of tables serves as a sort of advance organizer for the results section, forcing the researcher to think carefully about exactly how the notion of inferential robustness will be addressed. And once the tables have been constructed, researchers are reminded that each table needs to be fully discussed in the text, so that readers know how each table is organized and what the main findings from the table are.

So taken together, these suggestions offer both Foundation officials and quantitative researchers some quality control guidelines for presenting their research. By concentrating on the basics – simple, jargon-free writing, an organizational structure that really allows the writing to flow, and an emphasis on the inferential robustness of their

results -- aspiring quantitative researchers can produce work that not only speaks to the need of both Foundation officials and policymakers, but to the broader research community as well.

## **Conclusions**

The broad purpose of this paper has been to provide Lumina Foundation officials with a methodological overview of the analytical techniques and strategies required to produce quantitative research that offers inferentially rich insights and policy relevant conclusions, rather than the sort of research that, because of a flawed design or inappropriately applied analytical technique, actually detracts from the body of existing knowledge with its analytically unsupported conclusions. To provide officials with the necessary intuition to distinguish between these two types of research, this paper has focused on four critical issues – research design and causality, database design, common data problems, and content and presentation issues – and one overarching analytical technique – multiple regression analysis. This researcher’s hope is that through increased awareness of these critical methodological issues, Foundation officials will be able to more efficiently allocate their scarce research dollars, resulting in an increased return on their investment in research, both for internal as well as external audiences.

And finally, although these suggestions for improving the quality of quantitative research have been directed at Foundation officials, they could just as easily have been targeted at the entire educational research community. However, given the increasingly important role that Lumina Foundation has been taking in both state and federal policymaking, the importance of the Foundation producing methodologically sound research cannot be overstated. And hopefully, as more quantitative researchers in education pay attention to these critical methodological issues, when the National Academy of Science decides to once again look at the state of research in education they won’t find what they did in 1992 -- “methodologically weak research, trivial studies, an infatuation with jargon, and a tendency towards fads with a consequent fragmentation of effort” (Atkinson and Jackson, 1992).

## Endnotes

---

<sup>1</sup> An example of this might be the researcher that notices that every time the ice cream truck comes, children wear shorts. Therefore, the researcher concludes, the ice cream truck causes children to wear shorts. Of course, the real cause of both events is another variable, defined as “summer” or “hot weather”, but this example demonstrates that just because two variables are highly related that does *not* mean that one variable is causally linked to the other.

<sup>2</sup> The notion of holding everything equal between individuals while letting one thing vary is what economists commonly refer to as “*ceteris paribus*”.

<sup>3</sup> The multivariate correlation coefficient can also be thought of as a bivariate correlation coefficient between the dependent variable and the composite variable created by the independent variables.

<sup>4</sup> As long as the missing variable is correlated with any of the other variables in the model, then the coefficients of those variables are biased and inconsistent.

<sup>5</sup> Since students typically receive federal, state, or institutional grants, this means that the net price that they pay the college or university has been reduced by the amount of the grant aid received. So for a student facing a sticker price of \$20,000 that receives a \$5,000 institutional grant, their net price is really \$15,000.

<sup>6</sup> For example, when the independent variables are all continuous, multiway frequency analysis is the optimal technique; when the independent variables are all discrete then discriminant function analysis represents the best choice.

<sup>7</sup> Structural equation modeling is also known in the literature as causal modeling, simultaneous equation modeling, analysis of covariance modeling, path analysis, and confirmatory factor analysis.

<sup>8</sup> This loss occurs because there are fewer observations to estimate the relationship under study, effectively reducing the precision, or efficiency of any inference from the data.

<sup>9</sup> This surprising result follows from the fact that, by construction, the regression always passes through the point of means, so that when the sample mean is used in place of the missing observations, the slope of the regression line doesn’t change and the estimated coefficients remain unbiased.

---

## Bibliography

Atkinson, R.C., and Jackson, G.B. (1992). *Research and Education Reform: Roles for the Office of Educational Research and Improvement*. Washington DC: National Academy of Sciences.

Leamer, E.E., "Let's Take the Con Out of Econometrics," *American Economic Review*, March 1983, 73, 31- 43.

Tabachnick, B.G., and Fidell, L.S. (2001). *Using Multivariate Statistics*. Boston: Allyn and Bacon.

Thompson, B. (1999). "Common Methodological Mistakes in Educational Research, Revisited", Invited address presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

*Fred J. Galloway is currently associate professor in the School of Education at the University of San Diego, where he has also served as associate dean and director of strategic programs. Prior to joining the university faculty, he was project director for the national Direct Student Loan Evaluation project at Macro International, as well as director of federal policy analysis at the American Council on Education, where he represented the interests of the higher education community before the Executive and Legislative branches of the federal government. Dr. Galloway received his bachelor's and master's degrees from the University of California, San Diego, and his doctoral degree in the economics of education from Harvard University.*