Multiple Linear Regression Viewpoints

A Publication sponsored by the American Educational
Research Association's Special Interest Group on
Multiple Linear Regression: The General Linear Model

MLRV

Volume 31 • Number 1 • Fall 2005

Table of Contents

Multiple Linear Regression Viewpoints

Editorial Board

Multiple Linear Regression Viewpoints (ISSN 0195-7171) is published by the AERA Special Interest Group on Multiple Linear Regression: General Linear Model through the **University of Alabama at Birmingham**.

Subscription and SIG membership information can be obtained from:
Jeffrey B. Hecht, MLR:GLM/SIG Executive Secretary
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL 60115-2854.
jbhecht@niu.edu

MLRV abstracts appear in CIJE, the ERIC system, and microform copies are available from University Microfilms International, 300 North Zeeb Road, Ann Arbor, MI 48106. *MLRV* is listed in the *EBSCO Librarians Handbook.*

An Application of Panel Regression to Pseudo Panel Data

Jeffrey E. Russell **John W. Fraas**
Ashland University

This article illustrates how, in the absence of true panel data, multivariate regression analysis can be used in conjunction with a pseudo panel data set to identify variables that were related to the increase in the proportion of two-income spouses in the United States between 1940 and 2000. We present the procedures used to form the pseudo panel data set, construct and estimate the various models used to analyze the pseudo panel data, and interpret the results produced by those models. Our analysis revealed an inverse relationship between the proportion of two-income spouses and the presence of young children as well as an increasing trend across generations in the proportion of two-income spouses.

This article provides an illustration of how researchers can apply panel regression analysis techniques to pseudo panel data when true panel data are not available. The analysis conducted in this study was designed to identify variables that were related to the dramatic increase in the proportion of two-income, married couples in the United States between 1940 and 2000. Ben-Porath (1973) suggested that such investigations are best addressed through the use of panel data, and Baltagi (1995) described the substantial advantages provided by panel data analysis relative to cross-section or time-series data. However, due to the extensive time period covered by the data in our study, we were unable to form an appropriate panel data set. Thus, we constructed a pseudo panel data set to act as a substitute for a true panel data set.

The remaining sections of this article present the techniques we employed to construct and analyze this pseudo panel data set. Specifically, the first section of this article shows how a pseudo panel data set was constructed using cross-section data from the Decennial United States Census collected from 1940 to 2000. The second section describes the specifications of the various models used to analyze the pseudo panel data. The third section discusses the procedures followed to interpret the results produced by the various models. The final section summarizes the findings and the procedures used to produce those findings.

Construction of a Pseudo Panel Data Set

Panel data and pseudo panel data sets are obtained by pooling comparable cross-section data collected repeatedly over time. To maintain comparability, both true panel data and pseudo panel data should be based on responses to similar questions collected in a similar manner. True panel data also needs to be repeatedly collected from the same individuals across time to ensure comparability. The formation of a true panel data set is usually not a significant problem if individuals are defined to be a relatively small number of entities such as the member countries of the United Nations Security Council, and the questions are unambiguous (e.g., What is the population of each country?). In these situations, panel data covering an extended period of time may be constructed and pseudo panel data are usually not needed as an alternative.

Comparability over time becomes a more significant issue for true panel data if an individual is defined to be *individual* people or households and the number of individuals is very large. The Panel Study of Income Dynamics (PSID) from the Survey Research Center at the University of Michigan, and the National Longitudinal Surveys of Labor Market Experience (NLS) from the Center for Human Resource Research at The Ohio State University are two examples of this type of large, individual panel data. These high-quality data sets are very careful to pose consistent questions to the same individuals across time. Nonetheless, continued comparability becomes increasingly difficult over time as data are lost. A loss of data can occur due to individuals (a) failing to answer some questions in one or more time periods, (b) failing to respond at all in some years, or (c) dropping out of the data set because of death, migration, or deciding to no longer participate in the survey.

When the loss of data is non-random, researchers are faced with potential problems of bias that become increasingly problematic over time, even in top quality panel data. Since the likelihood of non-random data loss increases as the time period covered by the panel data increases, large panel data sets usually cover a relatively short period of time. To answer long-term individual behavioral questions, such as the ones we are addressing in this article, pseudo panel data can be used as a substitute for the unavailable true panel data.

Deaton (1985) demonstrated that a pseudo panel data set has the advantage of a less stringent requirement. That is, the data can be repeatedly collected from random samples drawn from the same time-stable cohort of individuals rather than repeatedly from the same specific individuals. Pseudo panel data are constructed by first defining cohorts using individual characteristics that are stable over time. If the size of each cohort is sufficiently large, successive surveys will generate successive random samples of individuals from each of the cohorts. For every cohort, the mean value for each variable is then calculated for each time period. These mean values become the observations in the pseudo panel data. As noted by Deaton, this procedure allows pseudo panel data to be constructed from any series of cross-section data that includes variables that can be used to identify stable cohorts.

In addition to filling gaps in the availability of true panel data, Deaton (1985) identified four additional advantages of pseudo panel data. First, data from different sources can be combined into a single set of pseudo panel data if comparable cohorts can be defined in each source. Second, attrition problems often found in true panel data are minimized. Third, the problem of the individuals' response errors is smoothed by the use of cohort means and can be explicitly controlled by using errors-in-variables methods. Fourth, inconsistencies between micro and macro analysis can be analyzed by moving from individual data to ever larger cohorts to one macro cohort.

Source of Data and Data Issues

As previously discussed, it was necessary to identify a source of successive surveys for the 1940–2000 time period. For our analysis, the successive surveys were the one percent public use microdata samples available through the United States Census Bureau for the seven census years beginning with 1940 and ending with 2000 (Ruggles, Sobek, Alexander, Fitch, Goeken, Hall, King, & Ronnander, 2004). Prior to forming our pseudo panel data set from this information, three data issues were addressed: (a) cohort stability over time, (b) measurement error bias, and (c) differentiation between age, period, and cohort effects.

Establishing the stability of cohorts over time. Even if awkward variables result, time-constant cohort definitions must be used with pseudo panel data. We defined cohorts using race, gender, and generation to prevent the movement of individuals between cohorts over time. Because we were investigating the work behavior of married couples, we first considered marital status as an additional cohort definition. While this would allow the straightforward calculation of the proportion of married couples that are two-income couples, marital status cannot be used as a cohort definition because individual marital status is not necessarily constant over time. To create a dependent variable of interest while maintaining cohort stability we calculated the proportion of the generation-race-gender cohort that consisted of working individuals married to working individuals. While the proportion of the cohort that consists of working individuals married to working individuals is not as straightforward as the simple proportion of married couples that are two-income couples, this type of variable definition was necessary to maintain cohort stability.

Addressing errors in measurement. As with true panel data, observations that are measured with a systematic error may need to be eliminated from the data to avoid biased results. The possible gains in obtaining more interpretable results produced by this technique must be weighed against the potential for bias due to a systematic elimination of a non-random group of individuals. For example, our study uses data from questions posed to individuals by the United States Census Bureau regarding their work behavior. The questions are not designed to reflect a farmer's work pattern. Consequently, a high degree of error in the responses of individuals engaged in farming exists. To eliminate this source of error in the data, we followed the practice suggested by Coleman and Pencavel (1993) and eliminated all observations from individuals living on farms prior to the calculation of the pseudo panel data cell means. This decision limits the applicability of the results to non-farmers. More importantly, this decision implicitly assumes the migration between census years of individuals off the farm and into the population used to calculate the cohort means was a random event and introduced no systematic bias. Immigrants to the United States were also eliminated from the data to maintain cohort stability.

Differentiating between age, period, and cohort effects. When data contain observations on many individuals over an extended period of time, observed variance can be attributed to three functionally related effects: (a) differences between cohorts, which are labeled the cohort effect; (b) differences associated with different points in the life cycle, which are labeled the age effect; and/or (c) differences associated with different periods, which are labeled the period effect. The problem that must be addressed regarding these three effects is that they cannot be simultaneously identified because only one time dimension and one individual or cohort dimension exists. More specifically, the functional relationship between all three effects causes perfect colinearity when all three effects are fully specified (Fienberg & Mason, 1985; Ryder, 1965). For example in our data set, if a regression model includes a cohort variable of 1910 for the 1906-1915 birth cohort, and a mean age variable of 40 for that cohort using the 1950 census, the 1950 period variable cannot be specified because it is already defined by the cohort and age variables (e.g., 1910 + 40 = 1950).

The question of how best to solve this identification problem has generated controversy, especially among sociologists (Rodgers, 1982; Smith, Mason, & Fienberg, 1982). If a linear restriction is imposed on any pair of age, period, or cohort variables (e.g., the membership in the cohort born 1906-1915 is no different from membership in the 1916-1925 cohort, thereby restricting the cohort variables to be equal for this pair), then the results are identifiable. However, Rodgers shows that such a restriction must be made on strong a priori grounds, and the researcher should know the restriction can easily distort the results.

An alternative solution is to recognize that the three accounting variables are proxies for substantive characteristics associated with age, period, and cohort. If one of the accounting measures can be replaced with a direct measure of a characteristic, the identification problem is solved (Feinberg, et al., 1985). For example, if the accounting variable for period is replaced with a substantive measure of the unemployment rate for each year, the age and cohort effects can be identified. The weakness of this strategy, however, is the inherent assumption that the substantive measure fully captures all aspects of the effect. In other words, the use of the unemployment rate implicitly assumes there are no other substantive period effects such as military conflicts or high rates of inflation.

We addressed the age, period, and cohort identification problem by using a linear restriction that all period effects are equal and are included in the constant term. This assumption allows a set of mean age dummy and cohort dummy variables to exactly identify the cohort and age effect in the regressions. An assumption that the period effect is a linear trend would also solve the identification problem. It should be noted, however, that the regression results produced by using a trend specification are more difficult to interpret because the cohort and age coefficients would then measure deviations from the trend.

Formation of the Cohorts

Stable cohorts were defined by race, gender, and generation (i.e., year of birth). The race characteristic was restricted to Caucasians and African-Americans only to maintain sufficient sample size for each of the two cohorts. The gender characteristic consisted of a male cohort and a female cohort. And the generation characteristic consisted of seven cohorts with each cohort representing a ten-year span. The first and seventh generation cohorts contained individuals born between 1906 and 1915, and between 1966 and 1975, respectively.

The race (2), gender (2), and generation (7) cohort definitions describe 28 potential (2*2*7 = 28) cohorts. Repeated over the seven census years, there was a potential of 196 cells of cohort mean data. However, to reduce the impact of schooling and retirement on the decision to work, individuals younger than 25 or older than 64 were excluded from the data. Consequently, beginning with the generation born in the years from 1946 to 1955, complete working age life-cycle data were not available because individuals in these later generations were less than 55 in the 2000 census. Similarly, as the oldest cohorts reach the age of 65, they no longer contribute data. As a result, the number of cells with data was reduced to 88 cells of sample mean data drawn from 28 distinct cohorts. Table 1 lists the actual set of 88 data cells that define the 88 cohort observations. The numbers listed in the cells indicate the number of individuals contained in the cohorts each census year. A review of Table 1 reveals the secular movement of younger cohorts into the data set and older cohorts out of the data set which reduced the number of useable cells to 88.

Table 1. Total Number of Observations in Each of the 88 Cohort Cells

	Census years						
	1940	1950	1960	1970	1980	1990	2000
White male born:							
1906-1915	21,943	23,208	80,242	72,874			
1916-1925		26,066	94,786	92,855	86,610		
1926-1935			89,373	93,129	92,353	89,550	
1936-1945				101,325	105,428	107,078	97,284
1946-1955					152,274	155,418	146,964
1956-1965						170,077	164,808
1966-1975							131,788
Black male born:							
1906-1915	1,897	2,299	7,865	6,957			
1916-1925		2,463	9,446	9,144	8,338		
1926-1935			9,539	9,926	9,684	7,516	
1936-1945				10,805	11,685	9,648	9,511
1946-1955					18,202	15,272	15,986
1956-1965						18,152	20,151
1966-1975							17,398
White female born:							
1906-1915	23,373	25,246	84,353	82,503			
1916-1925		29,139	100,706	99,122			
1926-1935			94,274	97,827	98,293	99,286	
1936-1945				106,583	108,219	111,627	105,490
1946-1955					155,083	159,867	151,485
1956-1965						174,432	169,314
1966-1975							134,886
Black female born:							
1906-1915	2,459	2,678	8,886	8,379			
1916-1925		3,240	11,338	10,799	10,138		
1926-1935			12,208	12,660	12,431	10,221	
1936-1945				14,065	14,852	12,304	12,079
1946-1955					22,416	19,071	20,126
1956-1965						23,611	25,833
1966-1975							23,256
Mean	12,418	14,292	50,251	51,907	62,821	73,946	77,897

Variable Formation

Three different types of variables can be used to represent the various characteristics of the pseudo panel data cohorts. A given characteristic can be represented by (a) a continuous variable, (b) one or more dummy variables, or (c) one or more proportional variables. The type of variable or variables formed to represent a given characteristic is, for the most part, dictated by the type of individual information collected in the surveys and its relationship to the cohort definitions.

Continuous variable. Some of the information used to form a pseudo panel data set may reflect a continuous type of measurement, such as income. Information of this nature would be used to form a continuous variable in the pseudo panel data set. For example, a continuous income variable in the pseudo panel data set would be formed by calculating the mean income for the individuals in each cell (i.e., cohort). It would have been possible for us to form an income variable in this manner, but we did not because income and educational level variables were highly correlated. We included only educational level variables in our analysis. One reason for selecting educational levels rather than income was to avoid the problems caused by truncated income data for individuals who were not working because their income was too low. Thus, it should be noted that we did not form any continuous variables to represent cohort characteristics.

Dummy variables. Other types of information used to form a pseudo panel data set could simply reflect the presence or absence of a specific characteristic for a given person. For certain cells, a characteristic was possessed by everyone in the cell or by no one in the cell. Variables formed from this type of information are true dummy variables. That is, these variables contain only values of zero or one. Dummy variables were used to represent four characteristics identified in our pseudo panel data set: (a) gender, (b) race, (c) generation, and (d) age.

Since gender and race consisted of only two categories, only one dummy variable was required to represent each of these characteristics. A value of zero was assigned to every cell that contained only males, while a value of one was assigned to every cell that contained only females. For example, the cohort of Caucasian males who were born between 1906 and 1915 and who responded in 1940 contained only males. Thus, the value for the gender variable was set equal to zero for this cohort. In the same fashion, a value of zero was assigned to every cohort that contained only Caucasians, while a value of one was assigned to every cohort that contained only African Americans. To facilitate the interpretations of the models used to analyze the pseudo panel data the gender and race variables were named for the groups assigned the value of one. Thus, the gender and race variables were labeled female and African American, respectively.

The generation characteristic specified whether an individual was or was not a member of a given generation. Unlike the gender and race characteristics, however, the generation characteristic consisted of more than two categories or levels. The generation characteristic consisted of the following seven levels: (a) Born 1906-1915, (b) Born 1916-1925, (c) Born 1926-1935, (d) Born 1936-1945, (e) Born 1946-1955 (f) Born 1956-1965, and (g) Born 1966-1975. Thus, seven dummy variables, with names corresponding to the cohort labels, were constructed to represent this generation characteristic.

While the information related to age would have allowed the calculation of a continuous mean age variable, we instead used four dummy mean age variables defined by the cohort's birth year and the census year to represent four levels of age. The four age levels were (a) Mean Age 30, (b) Mean Age 40, (c) Mean Age 50, and (d) Mean Age 60. Each variable contained a zero or a one value. We used this set of dummy variables with names corresponding to the cohort labels rather than a single, continuous age variable to avoid a linear restriction on the impact of age on the proportion of two-income spouses. Another reason for not using a continuous variable is that the calculation of a mean age every ten years for a group that is evenly distributed over ten possible birth-years results in very discontinuous mean age values that cluster tightly around the mean ages.

Proportional variables. Even though some information may indicate the presence or absence of a specific characteristic for each person in a cohort, the cohort will not be uniform regarding that characteristic. That is, the cohort will contain both individuals with the characteristic and individuals without the characteristic. For such variables, which are called proportional variables, a value equal to the proportion of individuals in the cohort with the characteristic was assigned to that cell for the variable. Our pseudo panel data contained four characteristics that required the formation of one or more proportional variables.

The dependent variable for this study, which was named two-income spouse, was a proportional variable. The 88 values formed for this variable were equal to the proportion of individuals identified as working and married to a working spouse in each of the 88 cells. To illustrate, since 40% of the individuals in the cohort containing male Caucasians born between 1906 and 1915 who responded to the survey in 1940 had a working spouse, the value for that cell in the dependent variable was 0.40. This value indicates the probability is 0.40 that an individual in that cohort will be a two-income spouse.

Proportional variables were constructed for three additional characteristics: (a) young children, (b) marital status, and (c) education level. The proportional variables formed for these characteristics were identified as independent variables. One proportional variable was constructed for the young children characteristic. Each value contained in this variable, which was named young child present, indicated the proportion of the individuals in a given cohort who had at least one child less than five years of age. The marital characteristic was also represented by one proportional variable. Each value recorded for this variable, which was labeled married, represented the proportion of individuals in a given cohort who were married. The educational characteristic reflected four levels of education: (a) less than high school, (b) high school graduate, (c) more than high school but less than four years of college, and (d) four or more years of college. Four proportional variables formed to represent these four education levels were named (a) less than HS, (b) HS graduate, (c) more than HS, and (d) four or more years of college. Every value for each of these variables was the proportion in the cohort with that level of education. For example, the 0.60 value recorded for the cohort containing male Caucasians born between 1906 and 1915 who responded to the 1940 survey indicated that 60% of these individuals had an education level less than high school.

Specification and Estimation of the Pseudo Panel Linear Regression Models

Since panel data can vary over both time and individuals, variables in a panel data regression model typically have a double subscript as follows:

$$y_{it} = \alpha + \beta x_{it} + u_{it} \qquad i = 1, \ldots, N; \quad t = 1, \ldots, T \tag{1}$$

where *i* represents the cross-section dimension (e.g., individuals, households, firms, countries, etc.) and *t* represents the time series dimension. α is a scalar, β is a vector of *K* explanatory variables, and x_{it} is the *i*th observation from time *t* on *K* explanatory variables.
Most panel data analyses use the following *one-way* error component model:

$$u_{it} = \mu_i + v_{it} \tag{2}$$

where μ_i represents unobservable, individual specific effects that do not change over time and v_{it} represents the remaining unobserved effects that vary over both individuals and time. Combining Equations 1 and 2, the one-way model is fully described as follows:

$$y_{it} = \alpha + \beta x_{it} + \mu_i + v_{it} \qquad i = 1,\ldots,N \quad t = 1,\ldots,T \tag{3}$$

The term *one way* refers to the decomposition of the error component in only the one dimension of time-constant, individual specific unobserved effects. The following *two-way* error component model is also possible:

$$u_{it} = \mu_i + \lambda_t + v_{it} \tag{4}$$

where the symbol λ_t represents unobservable, time-specific effects that do not change over individuals. An example of this type of effect would be different levels of funding in different years for a school district that impact all individual students in a similar, yet unobservable manner.

As with true panel data, a set of *T* independent cross sections represented by Equation 3 is pooled in pseudo panel data. Unlike true panel data, however, with pseudo panel data, *N* is a new, and most likely different set of individuals sampled in each census. To construct pseudo panel data, a set of *C* cohorts is defined such that any individual *i* sampled from the population will always be in the same, unique cohort every year. For example, in the data used in our analysis, an African-American male born in 1930 would be included in the African-American, male, 1926 -1935 cohort if that person was sampled in the 1960 census, and that person would be included in the same cohort if that person happened to be included in the 1980 sample.

Taking the mean value of each cohort's sample in each time period results in:

$$\overline{y}_{ct} = \overline{x}_{ct}\beta + \overline{\mu}_{ct} + \overline{v}_{ct} \qquad c = 1, \ldots, C \quad t = 1, \ldots, T \tag{5}$$

In this equation $\overline{y}_{ct}$ is the average of y_{it} over all individuals belonging to cohort *c* at time *t*. Unlike μ_i obtained from the true panel data Equation 3, $\overline{\mu}_{ct}$ retains the *t* subscript to indicate that each period's cohort mean is calculated from a new, and most likely different set of individuals. This results in a potentially different $\overline{\mu}_{ct}$ value for each period. In practice, if the number of individuals in each cell is large, as is the case for the data used in this article, the assumption is made that ($\bar{\mu}_{ct} = \bar{\mu}_c$) for every t and the fixed cohort effect ($\bar{\mu}_c$) is treated like a fixed individual effect (μ_i), resulting in the basic pseudo panel equation:

$$\bar{y}_{ct} = \bar{x}_{ct}\beta + \bar{\mu}_c + \bar{\nu}_{ct} \qquad c = 1, \ldots, C \quad t = 1, \ldots, T \qquad (6)$$

Additionally, if the cell size is large, random individual fixed effects will tend to be eliminated in the process of estimating the cell mean, leaving only the cohort fixed effect.

Much of Deaton's (1985) seminal work on pseudo panel data focuses on the availability of variances and covariance obtained in the construction of the cohorts' sample means which can then be used to weight the analysis of the pseudo-panel data using an "errors-in-variables" technique. Baltagi (1995) notes that as the average cohort size (number of cohorts/sample size) tends to infinity, measurement errors as well as their estimates tend to approach zero. Consequently, as is the practice followed by many applied researchers (e.g., Pencavel, 1998), the analyses presented in this article ignore the measurement error problem and simply weight the analysis using cell-size to address heteroscedasticity arising from the different levels of precision for cell means with different numbers of observations.

The Random Effects Model

The analysis of the pseudo panel data set begins with the estimation of the Random Effects Model using Equation 6. This model assumes that the $\bar{\mu}_c$ error term, which represents possible bias from unobserved, fixed cohort heterogeneity, is identically and independently distributed (IID) with a mean of zero (Baltagi, 1995). Baltagi also notes that this assumption allows the Random Effects Model to support inference for the population, assuming the sample is representative of the underlying population. Consequently, the Random Effects Model is preferred when analyzing either panel or pseudo panel data sets.

In our Random Effects Model the vector of cohort variables, x_{ct}, included the following:

1. The gender and race characteristics were represented by dummy variables named female and African American, respectively.

2. The characteristics of whether individuals had at least one child less than 5 and their marital status were represented by proportional dummy variables named young child present and married, respectively.

3. Since the four dummy variables used to represent the four levels of the age characteristic were linearly dependent, only three of the variables were included in the model: (a) mean age 40, (b) mean age 50, and (c) mean age 60. The Mean Age 30 age level served as the reference group for the coefficients estimated for these three variables.

4. Since the six dummy variables used to represent the generation characteristic were linearly dependent, only five of the variables were included in the model: (a) born 1916-1925, (b) born 1926-1935, (c) born 1936-1945, (d) born 1946-1955, (e) born 1956-1965, and (e) born 1966-1975. The Born 1906-1915 cohort level served as the reference group for the coefficients estimated for these five variables.

5. Since the four dummy variables used to represent the four levels of the education characteristic were linearly dependent, only three of the variables were included in the model: (a) less than HS, (b) more than HS, and (c) four or more years of college. The HS Graduate education level served as the reference group for the coefficients estimated for these three variables.

As previously mentioned, the use of the Random Effects Model relies on the assumption that $\bar{\mu}_c$ is IID with a mean of zero, that is, significant fixed effects do not exist. Thus before we begin to interpret the results of the Random Effects Model we must determine if significant fixed effects do, in fact, exist. The first step in this testing procedure is to construct and estimate a Fixed Effects Model.

The Fixed Effects Model and Testing for Fixed Effects.

The Fixed Effects Model, which is also called a *Least Squares Dummy Variable* (LSDV) model (Green, 1993), is estimated as follows when using pseudo panel data:

$$\overline{y}_{ct} = \overline{x}_{ct}\beta + \overline{\mu}_c + \overline{v}_{ct} \qquad (7)$$

This is identical to Equation 6. However, x_{ct} now includes a set of C cohort dummy variables. This assumes the impact of each cohort contains an estimable component that is fixed across time and these cohort components are significantly different. This is in contrast to the Random Effects Model which assumes these cohort components are IID and are simply included in the $\overline{\mu}_c$ error term. Unlike the Random Effects Model, inference from the results of the Fixed Effects Model is limited to the type of cohorts included in the analysis.

The Fixed Effect Model includes the same variables used to represent the characteristics of age, young children, and education as those included in the Random Effects Model. However, the dummy variables used to represent the gender, race, and generation characteristics are fully defined by the 28 cohort dummy variables. Consequently, the variables used to represent these characteristics, which are included in the Random Effects Model, are not specified in the Fixed Effects Model.

As previously mentioned, the preferred Random Effects Model can only be used if there are no significant fixed effects. To test for significant fixed effects the random effects estimation is compared to the fixed effects estimation. Specifically, a test for joint significance of the individual fixed effect dummy variables is calculated as follows (Baltagi, 1995):

$$F_0 = \frac{(\mathrm{RRSS} - \mathrm{URSS})/\ (\mathrm{N}-1)}{\mathrm{URSS}/(\mathrm{NT} - \mathrm{N} - \mathrm{K})} \sim F_{\mathrm{N}-1,\ \mathrm{N}(\mathrm{T}-1)-\mathrm{K}}$$

In this equation RRSS is the restricted residual sum of squares obtained from the random effects estimation and URSS is the unrestricted residual sum of squares obtained from the fixed effects estimation. *N* represents the total number of individuals, (*N* = *C* = 28 cohorts for our analysis), while *T* is the number of time periods (7 census years for our analysis). *K* represents the number of independent (non-cohort) variables in the x_{ct} vector of the Fixed Effects Model, (K = 8 for our analysis). If the panel is balanced, *C*T* will result in the total number of observations used in the regressions.

When the data do not contain information on all cohorts in all time periods, as is the case for our pseudo panel data set, *CT* overstates the number of observations and the associated degrees of freedom. For example, in our data set, *CT* = 28*7 = 196. However, only 88 observations are actually available due to the life-cycle nature of the data. Consequently, when calculating the F value *n* for our pseudo panel data set, the *NT* is equal to 88 and the denominator's degrees of freedom becomes 52 (88 - 28 - 8 = 52).

Data Transformation Models

If significant fixed effects exit, the Random Effects Model cannot be used. One alternative to using the Random Effects Model is to use the Fixed Effects Model. The Fixed Effects Model, however, may result in an undesirable loss of degrees of freedom due to the addition of a large number of cohort dummy variables. In these situations, the Within and the First-Differenced Models, which use transformed data, provide attractive alternative techniques to eliminate fixed effects without a large decrease in degrees of freedom. It should be noted that even if no significant fixed effects are present, the Within Transformation Model and the First-Differenced Model, along with the Between Transformation Model, provide additional insight into the core results of the Random Effects Model.

Researchers should be aware that when the Fixed Effects Model or the Within Transformation Model and First-Differenced Model are used, they do not eliminate bias from unobserved cohort heterogeneity that changes over time. In addition, the data transformations we employed for the Within and First-Differenced Models eliminate all observed as well as unobserved time-constant variables from the regressions. Despite these limitations, transformed panel data can offer a powerful rebuttal to criticisms that conclusions based on observed variables are actually just the result of correlation with unobserved variables.

The Within Transformation Model. The within transformation allows estimation of an equation where the bias from unobserved fixed cohort effects is *swept* from the equation, along with observed fixed effects. The within transformation controls for cohort fixed effects by calculating each variable's mean value across time for each cohort, then subtracting that mean from all observations. With pseudo panel data, this transformation first requires the calculation of each cohort's time mean values using the set of the cohort's mean values found in the data cells. Specifically, time mean values for the equation to be estimated are calculated as follows:

$$\bar{y}_c = \alpha + \beta \bar{x}_c + \mu_c + \bar{v}_c \qquad (8)$$

Equation 8 is identical to Equation 6, except the *t* subscript has been eliminated to indicate a mean value across time as well as across cohorts. The μ_c error term represents the unobserved fixed cohort effect and consequently is unchanged between Equations 6 and 8. The within transformation is obtained by subtracting Equation 8 from Equation 6 as follows:

$$y_{ct} - \bar{y}_c = \alpha - \alpha + \beta(x_{ct} - \bar{x}_c) + \mu_c - \mu_c + v_{ct} - \bar{v}_c \qquad (9)$$

The intercept term (α), as well as the cohort fixed effect (μ_c), do not change over time. Consequently, they are already time means by definition. If μ_c is assumed to sum to 0 across all cohorts, the within transformation is estimated as follows (Baltagi, 1995):

$$y_{ct} - \bar{y}_c = \beta(x_{ct} - \bar{x}_c) + (v_{ct} - \bar{v}_c) \qquad (10)$$

Data for the four race-gender cohorts from the youngest generation were also eliminated from the estimation of the Within Transformation Model. This was necessary because the youngest generation had observations in only the 2000 Census. Consequently, the time mean equaled the actual 2000 Census observations and the within transformation resulted in a set of zero values for all variables. Thus the number of observations for the Within Transformation Model was reduced to 84.

Variables used to represent characteristics that do not change over time (e.g., gender, race, and generation) are not included in the within transformation data set because the transformation of the values contained in these variables caused them to equal zero. The same set of proportional variables contained in the Random and Fixed Effect Models are also included in the Within Transformation Model.

The within transformations of true dummy variables that vary with time (e.g., the dummy variables for the age characteristic) cause a problem with the interpretation of the results produced by the Within Transformation Model. To allow us to interpret the coefficients for such variables the values that are generated by the within transformation are replaced in the transformed data set by their original 0 and 1 values.

If true panel data are used, the residual sum of squares (RSS) for the Within Transformation Model will be identical to the RSS for the Fixed Effects model. This relationship allows researchers to test for significant fixed effects in large data sets (e.g., the PSID) where software limitations on matrix size preclude estimation of a Fixed Effects Model with thousands of dummy variables. Unfortunately, when the Within Transformation Model is used with pseudo panel data the RSS produced by the model is not identical to the Fixed Effects RSS. This inconsistency is caused by the cell-size weighting used in pseudo panel estimations. As a result, when the transformed pseudo panel data are analyzed with the Within Transformation Model its RSS cannot be used to test for fixed effects.

Fortunately, the potential problem of too many dummy variables in the Fixed Effects Model can be addressed with pseudo panel data by deciding how narrowly to define the cohorts. For example, if we had defined generations on a one-year basis rather than a ten-year basis, an unmanageable 280 cohorts, and thus 280 additional dummy variables, would be required. In that case we would not have been able to practically test for the presence of significant fixed effects. Since our pseudo panel used ten-year generations, only 28 additional dummy variables were needed to estimate the Fixed Effects Model.

The Between Transformation Model. We also estimated the Between Transformation Model, which used data transformed by Equation 8. It should be noted that the application of this transformation procedure to the pseudo panel data set produces values for the set of dummy variables used to represent the age characteristic that do not vary. Hence that set of variables cannot be included in the Between Transformation Model. With the exception of the dummy variables used to represent the age characteristic, the Between Transformation Model includes the same set of variables used in the Random Effects Model. Rather than eliminating unobserved fixed cohort heterogeneity, the Between

Transformation Model isolates this heterogeneity and provides useful insights in the interpretation of the results of the Random Effects Model.

The First-Differenced Model. If a hypothesis involving a trend is to be tested and significant fixed effects from unobserved cohort heterogeneity is a concern, a first-differencing transformation can be used to sweep away the fixed effect and retain the trend. Calculating the first differences results in fewer degrees of freedom in the First-Differenced Model than exist in the Random Effects Model. This difference is due to the loss of the oldest observations for all 28 cohorts. As a result, the transformed pseudo panel data set used in conjunction with the First- Differenced Model contains 60 observations rather than 88. First-differenced data are obtained by subtracting each cohort's variable values from the prior year's values as follows:

$$y_{ct} - y_{c,t-1} = \alpha - \alpha + \beta(x_{ct} - x_{c,t-1}) + \mu_c - \mu_c + v_{ct} - v_{c,t-1} \qquad (11)$$

The First Differenced model is estimated as follows:

$$y_{ct} - y_{c,t-1} = \beta(x_{ct} - x_{c,t-1}) + v_{ct} - v_{c,t-1} \qquad (12)$$

The unobserved cohort fixed effect (μ_c) is removed from the data along with any observed variable that does not change over time (e.g., gender, race, and generation). Calculating the first difference for age characteristic dummy variables resulted in three possible values (i.e., -1, 0 or 1). These age characteristic dummy variables were dropped from the First-Differenced Model because the change in actual mean age for the cohorts between censuses was a constant value of ten. Thus the only variables contained in the First-Differenced Model are the proportional variables representing the young children, marital status, and education characteristics.

Interpretation of the Regression Results

Our analysis of the pseudo panel data set began by estimating the Random Effects and Fixed Effects Models. The results produced for these models are listed in Table 2. Once these models were estimated the following F test was conducted to determine whether the fixed effects were statistically significant:

$$F = \frac{(0.0769 - 0.0548)/27}{(0.0548/52)} = 0.773$$

where: (a) RRSS=.0769, (b) URSS=.0548, (c) N-1=28-1=27, and (d) *NT-N-K* = 88 - 28 – 8 = 52.

The probability value corresponding to the *F* value of .773 (*p* = .76) indicates the fixed effects were not statistically significant. This result indicates the impacts of the gender, race, and generation characteristics were sufficiently consistent across all 28 cohorts such that controlling for all the interactions of these characteristics in the Fixed Effects Model does not significantly improve the fit of the regression. Thus we are able to use the Random Effects Model as the foundation of the analysis due to our finding of no significant fixed cohort effect.

To assist in assessing the relationship of each characteristic to the dependent variable, the Within Transformation, Between Transformation, and First-Differenced Models were also estimated. The results produced for all five models are contained in Table 2.

An Analysis of the Independent Variables

Gender. In the Random Effects Model the coefficient for the female variable (-0.0022) was not significant at the 0.05 level, suggesting no significant difference between the proportion of two-income spouses was found in male cohorts compared to female cohorts. In addition, the coefficient for the female (-.0120) was not significant at the 0.05 level in the Between Transformation Model.

Additional information regarding the relationship between the gender characteristic and the proportion of two-income spouses is not produced by the Fixed Effects, the Within Transformation and the First-Differenced Models. In the Fixed Effects Model the single dummy variable for gender was not estimated because it was interacted with the race and generation variables to produce the 28 cohort dummy variables. Because gender is a time-constant variable, it was eliminated from the Within Transformation and First-Differenced Models. Thus, the impact of gender cannot be estimated in those models.

Race. In the Random Effects Model the coefficient for the African American variable
(-0.0431) was not significant at the 0.05 level, suggesting no significant difference between the proportion of two-income spouses found in the African-American cohorts compared to the Caucasian

cohorts. In addition, the coefficient for the African American variable (-0.0583) was not significant at the 0.05 level in the Between Transformation Model. Additional information regarding the relationship between the race characteristic and the proportion of two-income spouses is not produced by the Fixed Effects, the Within Transformation and the First-Differenced Models. In the Fixed Effects Model the single dummy variable representing the race characteristic is not estimated because it was interacted with the gender and generation variables to produce the 28 cohort dummy variables. Because race is a time-constant variable, it was eliminated from the Within Transformation and First-Differenced Models. Thus, the impact of race cannot be estimated in those models.

Age. The amount of variation in the dependent variables accounted for by the three dummy variables used to represent the age characteristic in the Random Effects Model was significant at the 0.01 level. To understand the life-cycle work pattern we compared and tested coefficients of adjacent age cohorts. The test results of those comparisons revealed that as people aged there was a significant increase in the proportion of two-income spouses until they reached the age category of Mean Age 60. At that point in time the proportion declined to a level that was not significantly different from the proportion estimated for the Mean Age 30 cohort.

Before we drew any conclusions regarding the relationship between the age characteristics variables and the dependent variable, we compared the results of the Fixed Effects Model to the results of the Random Effects Model to assess the robustness of the relationship. The amount of variation in the dependent variables accounted for by the three dummy variables used to represent the age characteristic in the Fixed Effects Model was also significant at the 0.01 level. Comparisons of the adjacent age cohort coefficients verified the life-cycle work pattern estimated by the Random Effects Model. The statistical tests of those adjacent coefficients, however, were not significant except for the decline in the coefficient values from the Mean Age 50 cohort to the Mean Age 60 cohort.

Because the Within Transformation Model is an alternative method of controlling for fixed effects, we anticipated it would produce similar results. As expected, the amount of variation in the dependent variable accounted for by the three dummy variables used to represent the age characteristic in the Within Transformation Model was also significant at the 0.01 level. Comparisons of the adjacent age cohort coefficients verified the life-cycle work pattern and statistical significance estimated by the Fixed Effects Model.

Additional information regarding the relationship between the age characteristic and the proportion of two-income spouses was not produced by the Between Transformation and the First-Differenced Models because any variance in the age characteristic is eliminated by the transformations. Thus, the impact of age cannot be estimated in those models.

Generation. In the Random Effects Model, the amount of variation in the dependent variable accounted for by the six dummy variables used to represent the generation characteristic in the Random Effects Model was significant at the 0.01 level. Once again the adjacent coefficient of these variables were compared and tested. The tests of adjacent coefficients were significant, which suggests that the proportion of two-income spouses increases with each generation.

The Fixed Effects Model contained 27 of the possible 28 three-way interaction variables created from the various levels of the gender, race, and generation characteristics. The variable not included in the series of linearly dependent variables represented the cohort labeled Caucasian, Male, Born 1906-1915. The amount of variation in the dependent variable accounted for by these 27 dummy variables in the Fixed Effects Model was significant at the 0.01 level. To determine whether the generational work pattern revealed by the Random Effects Model also existed for four race-gender subsets (African-American Female, African-American Male, Caucasian Female, and Caucasian Male), we compared and statistically tested the coefficients of adjacent generations within these subsets. These test results indicated each generation had a higher proportion of two-income spouses than its preceding generation. The tests of adjacent coefficients were significant, except for the two youngest Caucasian, female cohorts. Thus the increasing generational work pattern revealed by the Random Effects Model was also found for each of the four race-gender subsets.

Table 2. Regression Results of the Random Effects, Fixed Effects, Within Transformation, Between Transformation, and First-Differenced Models

Type of Model					
Independent variables	Random Effects[a]	Fixed Effects[b]	Within Transformation[c]	Between Transformation[d]	First Differenced[e]
Female	-0.0022 (0.0112)	n/a	n/a	-0.0120 (0.0079)	n/a
African-American	-0.0431 (0.0384)	n/a	n/a	-0.0583 (0.0320)	n/a
Mean age 40	0.0599* (0.0269)	0.0180 (0.0332)	0.0186 (0.0353)	n/a	n/a
Mean age 50	0.1076** (0.0423)	0.0527 (0.0502)	0.0326 (0.0561)	n/a	n/a
Mean age 60	0.0117 (0.0457)	-0.0207 (0.0534)	-0.1010 (0.0581)	n/a	n/a
Born 1916-1925	0.0939** (0.0258)	n/a	n/a	0.0809** (0.0168)	n/a
Born 1926-1935	0.1989** (0.0386)	n/a	n/a	0.1708** (0.0271)	n/a
Born 1936-1945	0.3684** (0.0550)	n/a	n/a	0.3283** (0.0319)	n/a
Born 1946-1955	0.5369** (0.0702)	n/a	n/a	0.4783** (0.0368)	n/a
Years of college ≥ 4	0.1707 (0.3303)	1.4238* (0.6658)	0.2385 (0.3827)	-0.2464 (0.3240)	1.3633 (0.8716)
Caucasian	**Male**				
Born 1916-1925	n/a	0.2459** (0.0560)	n/a	n/a	n/a
Born 1926-1935	n/a	0.3883** (0.0858)	n/a	n/a	n/a
Born 1936-1945	n/a	0.6449** (0.1244)	n/a	n/a	n/a
Born 1946-1955	n/a	0.8108** (0.1494)	n/a	n/a	n/a
Born 1956-1965	n/a	1.0027** (0.1473)	n/a	n/a	n/a
Born 1966-1975	n/a	1.0945** (0.1586)	n/a	n/a	n/a
African	**American**	**Male**			
Born 1906-1915	n/a	-0.2811** (0.1047)	n/a	n/a	n/a
Born 1916-1925	n/a	-0.0679 (0.0720)	n/a	n/a	n/a
Born 1926-1935	n/a	0.2086** (0.0513)	n/a	n/a	n/a
Born 1936-1945	n/a	0.6149** (0.0831)	n/a	n/a	n/a
Born 1946-1955	n/a	0.8923** -0.1223	n/a	n/a	n/a
Born 1956-1965	n/a	1.1036** (0.1445)	n/a	n/a	n/a
Born 1966-1975	n/a	1.2912** (0.1621)	n/a	n/a	n/a

Table 2 (Continued).

Type of Model					
Independent variables	Random Effects[a]	Fixed Effects[b]	Within Transformation[c]	Between Transformation[d]	First Differenced[e]
Coefficient					
Caucasian	**Female**				
Born 1906-1915	n/a	0.0944** (0.0338)	n/a	n/a	n/a
Born 1916-1925	n/a	0.3896** (0.0731)	n/a	n/a	n/a
Born 1926-1935	n/a	0.5656** (0.0977)	n/a	n/a	n/a
Born 1936-1945	n/a	0.7635** (0.1213)	n/a	n/a	n/a
Born 1946-1955	n/a	0.9081** (0.1435)	n/a	n/a	n/a
Born 1956-1965	n/a	1.0138** (0.1486)	n/a	n/a	n/a
Born 1966-1975	n/a	1.0597** (0.1658)	n/a	n/a	n/a
African	**American**	**Female**			
Born 1906-1915	n/a	-0.2707** (0.1051)	n/a	n/a	n/a
Born 1916-1925	n/a	-0.0428 (0.0731)	n/a	n/a	n/a
Born 1926-1935	n/a	0.2465** (0.0597)	n/a	n/a	n/a
Born 1936-1945	n/a	0.6207** (0.0909)	n/a	n/a	n/a
Born 1946-1955	n/a	0.9140** (0.1271)	n/a	n/a	n/a
Born 1956-1965	n/a	1.0842** (0.1436)	n/a	n/a	n/a
Born 1966-1975	n/a	1.2552** (0.1550)	n/a	n/a	n/a
Constant	-0.0177 (0.1076)	-0.9773** (0.2549)	0.2627** (0.0507)	0.0783 (0.0758)	-0.0195 (0.0207)

*significant at the 5% level
**significant at the 1% level
[a] Random Effects Model: N=88; F(16,71)=157.1**; R^2=.97; Root MSE=.033; RSS=.0769
[b] Fixed Effects Model: N=88; R^2=.98; F(35,52)=74.5**; Root MSE=.033; RSS=.0548
[c] Within Transformation Model: N=84; R^2=.8237; F(8,75)=43.8**; Root MSE=.047;RSS=.1665
[d] Between Transformation Model: N=28; R^2=.99; F(13,14)=521.0**; Root MSE=.011; RSS=.0016
[e] First-Differenced Model: N=60; R^2=.7331; F(5,54)=29.7**; Root MSE=.0610; RSS=.2012
[f] The standard errors are enclosed in parentheses

In the Between Transformation Model, the amount of unique variation in the dependent variable accounted for by the six dummy variables used to represent the generation characteristic was significant at the 0.01 level. The tests of adjacent coefficients found the same pattern of a significant increase in the proportion of two-income spouses with each generation that was found in the Random Effects and Fixed Effects Models, further supporting the robustness of the generational pattern.

Since generation is a time-constant variable, it was eliminated from the Within Transformation and First-Differenced Models. Thus, these models could not provide additional information regarding the relationship between the generation characteristic and the proportion of two-income spouses.

Young children. In the Random Effects Model the coefficient for the young child present variable (-0.2873) was significant at the 0.01 level. Because this is a proportional variable, interpreting the coefficient for a one-unit change in the variable is the method of interpretation. A more realistic interpretation is that a 0.10 increase in the proportion of a cohort with at least one young child present was associated with a 0.02873 decrease in the proportion of two-income spouses. The Fixed Effects Model confirmed this same, significant relationship.

In the Within Transformation Model the coefficient for the young child present variable (-0.2778) was also significant. This suggests that over the course of a cohort's life-cycle, the timing of when couples choose to have children significantly impacts changes in the proportion of two-income spouses within that cohort. The coefficient for the young child present variable was not significant in the Between Transformation model at the 0.05 level. This suggests that cohorts which *average* a higher proportion of individuals with young children present do not have a significantly lower proportion of two-income spouses. Finally, in the First-Differenced Model, the coefficient for the young child present variable (-0.6781) was significant at the 0.01 level. This provided further confirmation of the inverse relationship between the proportion of two-income spouses and the proportion of individuals with young children that was found in the Random Effects, Fixed Effects, and Within Transformation Models.

Marital status. To control for changes in a cohort's proportion of married individuals, all five models included a married variable that measured the proportion of a cohort that was married. This variable was not significant at a 0.05 level in any of the models.

Education. In all five models, the amount of unique variation in the dependent variable accounted for by the three dummy variables used to represent the education characteristic was significant at the 0.01 level. The coefficient for the less than HS variable in the Random Effects Model (0.4774) was significant at $\alpha = 0.01$. Apparently, individuals with the least amount of education are more likely to be two-income spouses relative to individuals with a high school education. Assuming education and income are correlated, this finding may be due to greater pressure for both spouses to work if income is low. The coefficient for less than HS variable was also significant and positive in the Fixed Effects and Between Transformation Models, which supports the pattern suggested by the Random Effects Model. The coefficient in the Within Transformation Model was also positive, but it was not significant at $\alpha = 0.05$.

The coefficient for less than HS variable in the First-Differenced Model (3.0199) was significant at the 0.01 level. Because our data set excludes individuals younger than age 25, this indicates that a significant number of individuals completed their high school education after age 25. It should be noted that this coefficient has the largest of any of the proportional variables used to represent the educational characteristics. This coefficient indicated that a 0.10 increase between censuses in the proportion of a cohort with at least a high school education was associated with a 0.30199 decrease in the proportion of two-income spouses. We compared the coefficients for each level of education beyond high school with the coefficient for the adjacent level of education and found no significant difference in the coefficients for any of the five models at the 0.05 level. This suggests that the incentive for both spouses to work to earn the higher salary available with increased education is approximately offset by the decreased need for both spouses to work as higher education allows one spouse to contribute more income to the household.

Summary

In an attempt to identify characteristics that are related to the increase in the proportion of two-income spouses we constructed a pseudo panel data set from the Decennial United States Census collected from 1940 to 2000. A comparison between the Random Effects Model and the Fixed Effects Model revealed that the fixed effects were not statistically significant. Consequently, the Random Effects

Model formed the core of our analysis. Additional insights regarding the relationships of the various characteristics and the proportion of two-income spouses were provided by the results produced by the Within Transformation, Between Transformation, and First-Differenced Models, which used transformed data sets. The Random Effects Model revealed a significant life-cycle pattern of increasing probability of two-income spouses as the cohorts age. The Random Effects, the Fixed Effects, and the Between Transformation Models all showed a significant increase across generation cohorts in the proportion of two-income spouses.

Our findings consistently support the hypothesis that within a cohort, the presence of young children reduces the probability of two-income spouses. Moving from less than a high school education to a high school level of education was significantly associated with a decrease in the proportion of two-income spouses, possibly due to decreased pressure to work as incomes increased with education. The insignificant coefficients for education levels beyond high school appear to reflect offsetting incentives for spouses to work to take advantage of higher potential income and the reduced need to work if household income is higher.

We have sought to provide a reference or guide to researchers who encounter questions that can be addressed with the use of panel data, yet find no true panel data set is available. This paper demonstrated how pseudo panel data can be constructed to address the lack of a true panel data set with special attention given to some of the nuances inherent in its construction. We also described and illustrated how the Random Effects, Fixed Effects, Within Transformation, Between Transformation, and First-Differenced Models were constructed and interpreted when applied to a pseudo panel data set. It is our hope this article will encourage researchers to investigate questions that may have been left unanswered due to a lack of panel data.

References

Baltagi, B. H. (1995). *Econometric analysis of panel data.* New York: John Wiley and Sons, Inc.

Ben-Porath, Y. (1973). Labor force participation rates and the supply of labor. *Journal of Political Economy*, 81, 697-704.

Coleman, M. T. & Pencavel, J. (1993). Changes in work hours of male employees, 1940-1988. *Industrial and Labor Relations Review, 46*(1), 262-283.

Deaton, A. (1985). Panel data from time series of cross-sections. *Journal of Econometrics*, *30*, 109-126.

Fienberg, S. E. & Mason, W. M. (1985). Specification and implementation of age, period, and cohort models. In W. M. Mason & S. E. Fienberg (Eds.), *Cohort analysis in social research: Beyond the identification problem.* New York: Springer-Verlag.

Greene, W. H. (1993). *Econometric analysis.* New Jersey: Prentice-Hall, Inc.

Pencavel, J. (1998). The market work behavior and wages of women. *The Journal of Human Resources*, *33* (4), 771-804.

Rodgers, W. L. (1982). Estimable functions of age, period, and cohort effects. *American Sociological Review*, *47*, 774-787.

Ruggles, S., Sobek, M., Alexander, T., Fitch, C. A., Goeken, R., Hall, P.K., King, M., & Ronnander, C. (2004). *Integrated public use microdata series: Version 3.0* [Machine-readable database]. Minneapolis, MN: Minnesota Population Center [producer and distributor]. http://www.ipums.org

Ryder, N. B. (1965). The cohort as a concept in the study of social change. *American Sociological Review, 30*, 843-861.

Smith, H. L., Mason, W. M. & Fienberg, S. E. (1982). More chimeras of the age-period-cohort accounting framework: Comment on Rodgers. *American Sociological Review, 47*, 787-793.

Send correspondence to: Jeffrey E. Russell, Ph.D.
Ashland University
Email: jrussell@ashland.edu

The Misuse of ANCOVA: The Academic and Political Implications of Type VI Errors In Studies of Achievement and Socioeconomic Status

Susan M. Tracz
California State University, Fresno

Laura L. Nelson
University of Wisconsin, La Crosse

Isadore Newman
University of Akron

Adrian Beltran
California State University, Fresno

This paper examines ANCOVA designs which use SES as the covariate for achievement and Type VI errors. Type VI errors are inconsistencies between the research question and the research methodology, and these errors are discussed in the context of general semantics. The consequences of a Type VI error in studies of achievement differences covariating for SES can be highly misleading. When research with a Type VI error concludes that there are no significant differences in achievement across groups when statistically controlling for SES, the tacit implications are that actual achievement is consistent across groups and that SES can be causally controlled or is somehow not influential. Neither is correct. Authors suggest conducting validity studies of adjusted outcome scores to insure accuracy in interpreting results.

The study of student achievement is a major focus of educational researchers and practitioners. With the recent passage of the No Student Left Behind legislation, the study of the factors related to increasing achievement scores has intensified. One salient variable which is correlated to achievement is SES (Attewell & Battle, 1999; Chapell & Overton, 2002; Gregory, 2000; O'Brien, Martinez-Pons & Kopala, 1999; Verna & Campbell, 1998), and researchers have used it as a covariate for achievement (Dillon & Schemo, 2004; Ferguson, 1981; Kaplan, 2002). The assumed reasoning is that if the variance attributable to SES is removed, the unique variance in achievement can be examined and explained. However, many errors in conceptualization and interpretation are possible with such designs. The purposes of this paper are:

1. To review the way in which ANCOVA is typically explained in textbooks commonly used in university statistics classes,
2. To provide a conceptual framework for discussions of numerical descriptions and the use of language,
3. To explain how Type VI errors, which are inconsistencies between the research question and the research methodology, can lead to inaccurate conclusions and/or interpretations of the data,
4. To provide descriptions of common errors in studies testing for group differences in achievement in which SES is used as the covariate,
5. To discuss the educational and political implications of such errors, and
6. To suggest that the adjusted scores in ANCOVA designs be correlated with other appropriate measures to determine their validity and correct interpretability.

ANCOVA Designs

Isolating and examining the unique variance in a dependent variable in studies of group differences can be undertaken by logical argument, by research design, and by statistical control (McNeil, Newman & Kelly, 1996). While argument is obviously the weakest method, and research design is the strongest, including all confounding variables in a design is not always possible. In such cases, the statistical control provided by analysis of covarience can be a viable alternative. However, there are stringent underlying assumptions which must be met for its appropriate use and interpretation, and these assumptions are frequently violated.

The ANCOVA is a statistical technique used to ascertain group differences on an adjusted dependent variable. This statistical analysis is similar to ANOVA in that it is a vehicle for determining group differences with the exception that instead of examining group means, adjusted group means are studied. In fact, each score is adjusted when the effects of the covariate are statistically removed from the dependent variable.

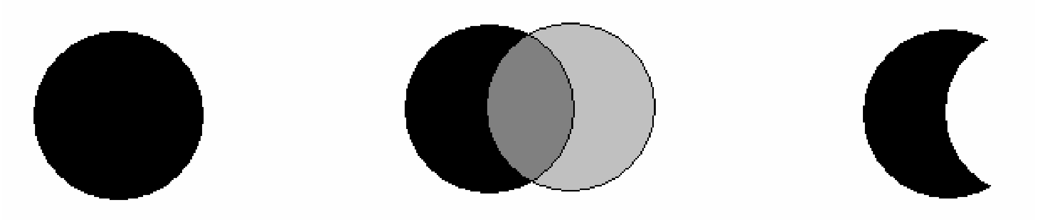
In his classic work, Pedhazur (1982) presents the mathematical logic of covariance with the example of achievement as the dependent variable, intelligence as the covariate, and a treatment as the independent variable. Recalling that when "a variable is residualized, the correlation between the predictor variable and residuals is zero" (p. 496), if intelligence is used to predict achievement, the residuals are then zero correlated with intelligence. These residualized scores are then called the adjusted scores for achievement and are analyzed for group differences. Pedhazur summarized as follows:

$$Y_{ij} = \overline{Y} + T_j + b(X_{ij} - \overline{X}) + e_{ij}$$

where Y_{ij} is the score of subject *i* on treatment *j*, $\overline{Y}$ is the grand mean of the dependent variable, T_j is the effect of the treatment, *b* is the regression coefficient, X_{ij} is the score on the covariate for subject *i* under treatment *j*, $\overline{X}$ is the grand mean for the covariate, and e_{ij} is the error term. This formula also can be

$$Y_{ij} - b(X_{ij} - \overline{X}) = \overline{Y} + T_j + e_{ij},$$

which shows that the adjusted score is equal to the grand mean plus a treatment effect plus error.



Variance in the dependent variable — Overlap in the variance between the dependent variable and the covariate — Variance in the adjusted dependent variable

Figures 1A 1B 1C

Finally, it is important to remember that the outcome or dependent variable in ANCOVA is an adjusted score. To reiterate this point visually, Figure 1A represents the total or 100% of the variance in the dependent variable, for example, achievement. Figure1B represents the overlap between the total variance in the dependent variable and the covariate, for example, achievement and SES. After the effects of the covariate have been statistically controlled or removed from the dependent variable (Figure 1C), the error variance is all that remains. This residualized or adjusted dependent variable is no longer the same as the original dependent variable.

Numbers and the Use of Language

To put the distinctions between the original dependent variable and an adjusted outcome into context, some discussion of general semantics may be useful. Polish mathematician Count Alfred Korzybski (1948) applied the perspective of mathematics to limitations and problems of language in order to help people use language more precisely and effectively, thereby avoiding common problems in communication. His germinal work, *Science and sanity: An introduction to non-Aristotelian systems and general semantics*, gave rise to a highly influential movement in language and communication studies, general semantics. Korzybski recognized a number of highly significant truths about human sign systems that apply to both language and numbers, of which three are relevant here. First, the word or the number is not the phenomenon that is labeled with that word or measured by that number. There is always more to the real world phenomenon than a word or number can capture. Korzybski and his followers summarize this with the aphorism, "The word is not the thing," which might also be paraphrased as "The number is not the thing." A further parallel might be "One's IQ score is not one's intelligence."

Second, the words or numbers used to describe the concrete material world in more complex ways can never represent 100% of what is described, summarized with the aphorism, "The map is not the territory." Anyone who has used a map but still gotten lost has had a practical experience of this truth. Nor can a battery of test scores hope to represent the concrete complexity of a child.

Finally, Korzybski pointed out that while signs in code systems allow communication with one another about the concrete material world in useful ways, abstraction above and beyond that world is inherent in their use. When researchers talk about or measure the immediate environment, they have moved one step away from that environment with those words or numbers. When they discuss the talk or average those measurements, they have moved two steps away. When researchers analyze the discussion about the original talk, or manipulate the averages of the original measurements, they have distanced themselves yet again, and so on. As researchers measure achievement, calculate means, covary for still other measures, the original concept of achievement becomes so abstract as to be something else entirely.

That is, it is difficult to understand what is actually being measured by the adjusted score. It could be a meaningful concept, or it could be something else entirely, possibly error. For example, the original criterion may have been reading achievement, which had validity estimates and made logical sense from a nomological net. The achievement test may have made sense and had adequate validity in terms of its relationship or correlation to other achievement estimates, such as other tests and teacher evaluations. However, when the variance from socioeconomic status is removed from that achievement test, the residual or adjusted score may not have the validity support that the original, unadjusted, reading achievement score had. Therefore, it is possible that the adjusted score, as a criterion, may actually have less validity than the original unadjusted score.

Types of Covariates

To further complicate the situation, the word, "control," can be interpreted in multiple ways. In an attempt to clarify the use of the word, "control," in connection with ANCOVA, Ferguson (1981) proposes that covariates fall into two categories, intrinsic and extrinsic. Intrinsic covariates are attributes which are internal to the subjects such as pretest scores or motivation and may be affected by some aspect in a study, while extrinsic covariates, such as teacher's years of service, are external to both the subjects and the study. SES is, of course, an extrinsic covariate, and as such, is not under the researcher's influence or ability to change. The technical terminology, "controlling for the covariate," however, implies the opposite to those unfamiliar with the language of statistical testing.

Type VI Errors

Given the preceding discussion of semantic meaning, research is still conceptualized as having multiple purposes including predicting; adding to the knowledge base; having a personal, social, institutional, and/or organizational impact; measuring change; understanding complex phenomena; testing new ideas; generating new ideas; informing constituencies; and examining the past (Newman, Ridenour, Newman, & DeMarco, 2003). In the cases of interest here where the effects of SES are removed from achievement, the researcher generally aspires to predicting outcomes or measuring change, both quantitative questions in nature. Such studies have appeal for educational researchers, evaluators and school district personnel, who often speculate whether significant differences exist in achievement between specific programs or ethnic groups but recognize that achievement is confounded with SES.

However, such a research design may be making a Type VI error (Newman, Fraas, Newman & Brown, 2002). Type VI errors occur when the research question does not match the research design. Type VI errors include: "practices that (a) fail to distinguish between statistical analysis and research design issues, (b) do not match the model used in structural equation modeling with the research question, (c) analyze a research question that involves practical significance with an analytical technique that fails to do so, (d) use methods designed to control for inflated Type I error rates that do not match the nature of the research question, and (e) employ multivariate data analysis techniques for research questions that require the application of univariate techniques" (p. 138).

The real question of interest in the ANCOVA design discussed here is whether there are significant differences in achievement, and while researchers may hope such differences exist or not, they really want to know about achievement, the phenomena, and not adjusted achievement, an abstract number. To emphasize this point, predicted scores for the dependent variable of GPA, for example, can be generated in ANOVA and ANCOVA, yet actual GPA and adjusted GPA would look quite different. Adjusted GPA is the residualized variable after the effects of a covariate such as SES has been removed, and it lacks the same meaning as GPA. Adjusted GPA is not in the same metric as GPA, or in other words, it does not have the same mean and standard deviation as GPA. It is inaccurate and misleading to draw conclusions about one variable (GPA) when the analysis has been conducted on a different variable (adjusted GPA).

Common ANCOVA Patterns

SES is so commonly used as a covariate for achievement that such designs have been reported by the popular press (Dillon & Schemo, 2004). In these kinds of designs, the outcome variable is some measure of achievement, the covariate is SES, and the grouping variable may be ethnicity, developmental levels (accelerated, normal, slow), treatment (treatment group, control), school ratings (high, medium, low in achievement), time (pretest, posttest), school type (public, charter, private), or community area (urban, suburban, rural) to name a few of the possibilities.

Typically, these studies begin by accurately describing their variables and analyses, and then slip into some common but unsupportable practices. First, they don't test for homogeneity of regression and so never look for interactions. Second, they fail to understand the implications of the fact that the adjusted scores or residuals will be zero correlated with SES, which overlaps greatly with achievement. Third, their research questions often seem to expect conceptually counterintuitive outcomes, such as students from the poorest performing schools should exhibit academic performances comparable to those students from the highest performing schools if the effects of SES are removed. Fourth, they slip into using achievement and adjusted achievement interchangeably, ignoring the fact that these represent vectors of different scores with different means, different standard deviations, and different patterns of variability. They may even neglect to provide tables of original and adjusted means, or label figures with adjusted means as simply "achievement." Finally, if group means for achievement show one pattern, but means for adjusted achievement appear comparable or even exhibit reversals, misleading conclusions can be made if adjusted means are presented but referred to as original achievement. For example, reading achievement means may show normal-developing students outperforming slow-developing student, but adjusted reading achievement means may show they are comparable. If the results discuss adjusted means but simply refer to them as reading achievement, an erroneous conclusion may be that the two groups are the same, when the truth is that the slow-developing group is still just that, slow-developing and poor in reading.

Determining the Validity of Adjusted Scores

One possible remedy for this problem of interpreting adjusted scores is to treat them as typical variables and to subject them to recognized and acceptable methodological practices. It is a common research practice to assess whether the outcome variable is a valid measure. To ascertain this, the outcome variable is correlated with other recognized, widely used measures of the same construct (Groth-Marnat, 1999; Huck & Cormier, 1996, Nunnally, 1978). Adjusted scores in ANCOVA designs should also undergo the same procedure. If they correlate with other similar measures, their usage is acceptable, and those adjusted scores can be interpreted in a meaningful way. If the adjusted scores do not correlate with other similar measures, such as achievement for example, they are merely residuals or random error, and as such, cannot be interpreted as the original variable, achievement or anything else.

Conclusion

Inconsistencies between the research question and the research methodology, indicative of a Type VI error, can be pervasive and subtle in their semantic expression. Although adjusted scores are used in the analysis cited in this example, conclusions and graphs often tend to refer simply to achievement or to shift in subtle linguist ways that imply that unadjusted achievement is the outcome. Such conclusions are not only inaccurate, but may lead to inappropriate recommendations.

The consequences of a Type VI error in studies of achievement differences covariating for SES can be highly misleading. When research with a Type VI error concludes that there are no significant differences in achievement across groups when statistically controlling for SES, the tacit implications are that actual achievement is consistent across groups and that SES can be causally controlled or somehow is not influential. Neither is correct.

The consequences of a Type VI error in studying achievement when covarying for variables such as SES can be highly misleading. Generally this happens when the researcher asks a question about achievement, conducts the analysis on adjusted achievement scores, but interprets the results in terms of plain, unadjusted achievement. Adjusted achievement is no longer achievement because meaningful, predictable, overlapping variance has been statistically removed. Adjusted achievement is an abstract, unknown construct. That is, what is meant by achievement after the effects of the covariate of SES is statistically removed is unknown. However, achievement scores for children are increasingly being used for high stakes decisions. Researchers need to keep the two straight, and educators need to teach the difference between the two.

The semantic substitution of achievement for adjusted achievement has a far-reaching impact on subsequent discussions or recommendations. What researchers are testing is not what the research question hopes to ascertain, although they sound quite similar. Achievement is a complex interplay of shared variance, some of which can be influenced by teachers, educational systems or other factors and caused to improve, while some of this variance is attributable to factors such as SES which is external and not controllable in the context of educational research. Higher levels of abstraction are extremely valuable as they can reveal truths and trends that would never be perceived without the human ability to abstract through the use of words and numbers. The other edge of the sword is that higher and higher levels of abstraction may also distort those perceptions, misrepresent the concrete material world in significant ways, and mislead thinking as researchers attempt to understand whatever phenomenon is the focus of their attention.

It is also notable that children who have low SES are much more likely to be children of color. When conclusions that no significant differences in achievement between programs or among ethnic groups are reached, the recommendations that follow may find special programs to help students with specific needs or to create parity are unnecessary. Therefore, it is essential that researchers, policy makers and practitioners carefully distinguish between the manner in which achievement is defined and is validated in research questions and research methodologies. These definitions must coincide for conclusions and recommendations to be viable. Furthermore, if professionals lack clarity in differentiating between adjusted and unadjusted means in achievement, there is little reason to expect that the general public will understand this seemingly subtle distinction or will understand that adjusted achievement won't look anything like a child's ability to read.

References

Attewell, P., & Battle, J. (1999). Home computers and school performance. *Information Society, 15*(1), 1-10.

Chapell, M. S., & Overton, W. F. (2002). Development of logical reasoning and the school performance of African American adolescents in relation to socioecomonic status, ethnic identity, and self-esteem. *Journal of Black Psychology, 28*(4), 295-317.

Dillon, S., & Schemo, D. J. (2004, November 23). Charter schools fall short in public school matchup. *New York Times.* Retrieved June 27, 2005, from http://www.nytimes.com

Ferguson, G. A. (1981). *Statistical analysis in psychology and education (5th ed.).* New York: McGraw-Hill.

Groth-Marnat, G. (1999). *Handbook of psychological assessment (3rd ed.).* New York: John Wiley & Sons.

Gregory, S. T. (Ed.). (2000). *The academic achievement of minority students: Perspectives, practices, and prescriptions.* Lanham, MD: University Press of America.

Huck, S. W., & Cormier, W. H. (1996). *Reading statistics and research (2ne ed.).* New York: Harper Collins.

Kaplan, D. (2002). Methodological advances in the analysis of individual growth with relevance to education policy. *Peabody Journal of Education, 77*(4), 189-215.

Korzybski, A. (1948). *Science and sanity: An introduction to non-Aristotelian systems and general semantic*,(3rd ed.). Lakeville, CN: International Non-Aristotelian Library, Institute of General Semantics.

McNeil, K., Newman, I., & Kelly, F. J. (1996). *Testing research hypotheses with the general linear model.* Carbondale, IL: Southern Illinois University Press.

Newman, I., Fraas, J., Newman, C., & Brown. (2002). Research practices that produce Type VI errors. *Journal of Research in Education, 12*(1), 138-145.

Newman, I., Ridenour, C. S., Newman, C., & DeMarco, G. M. P., Jr. (2003). A typology of research purposes and its relationship to mixed methods. In A. Toshakkori & C. Teddie, Eds., *Handbook of Mixed Methods in Social & Behavioral Research* (pp. 167-188), Thousand Oaks, CA: Sage.

Nunnally, J. C. (1978). *Psychometric theory (2nd ed.).* New York: McGraw-Hill.

O'Brien, V., Martinez-Pons, M., & Kopala, M. (1999). Mathematics self-efficacy, ethnic identity, gender, and career interests rerlated in mathematics and science. *Journal of Educational Research, 92*(4), 231-235.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart and Winston.

Verna, M. A., & Campbell, J. R. (1998). *The differential effects of family processes and SES on academic self-concepts and achievement of gifted Asian American and gifted Caucasian high school students.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA, April 13-17. (ED 419 025).

Send correspondence to: Susan M. Tracz
California State University, Fresno
Email: susant@csufresno.edu

Interval Estimates of R^2: An Empirical Investigation of the Influence of Fallible Regressors

Jeffrey D. Kromrey **Melinda R. Hess**
University of South Florida

Recent efforts to improve the analysis of multivariate data have included the use of confidence intervals rather than the more commonplace hypothesis testing. The use of interval estimation in regression analysis not only provides the ability to reject or fail to reject a given hypothesis, it also provides estimates of intervals within which a parameter is expected to reside. This study examines the potential effects of fallible regressors on the precision and accuracy of confidence intervals around R^2 when predictors vary in their reliability. Monte Carlo methods were used to investigate four methods for constructing these intervals around R^2: two percentile approaches based on the asymptotic normality of the distribution of R^2, a Fisher *Z* transformation method, and an interval inversion approach. The factors manipulated in the Monte Carlo study included the population value of R^2, number of regressors, sample size, population distribution shape, regressor intercorrelation, and regressor reliability. Results support the superiority of the interval inversion approach to confidence interval construction. However, as the reliability of the regressors decreased, none of the methods provided accurate intervals.

Recent efforts to improve the analysis of multivariate data have focused on (among other issues) the use of confidence intervals rather than the more commonplace hypothesis testing. In the context of multiple regression, many researchers (e.g. Steiger & Foray, 1992; Algina & Olejnik, 2000; Wilkinson & the APA Task Force, 1999) have provided justifications for the use of confidence intervals contending that they provide more information with better accuracy than the testing of null hypotheses. Of course, when properly applied, a confidence interval approach requires that researchers carefully consider design factors such as adequate sample size and appropriate procedures for sample selection and data collection. The use of interval estimation in regression analysis not only provides the ability to reject or fail to reject a given hypothesis (i.e., if the 1-α confidence band contains the null hypothesized parameter value), it also provides the researcher with estimates of intervals within which a parameter is expected to reside. The recent evolution of using confidence intervals for R^2 has primarily employed the assumption of normal distributions (e.g., Alf & Graf, 1999; Algina, 1999; Algina & Keselman, 1999) although emerging research into the effects of non-normal populations (Kromrey & Hess, 2001) on confidence intervals around R^2 has begun. In all of these investigations, however, there is one more element associated with realistic data that has not yet been addressed, namely the use of regressors that are not perfectly reliable. As such, this study examines the potential effects of fallible regressors on the precision and accuracy of confidence intervals around R^2 when predictors vary in their reliability.

Effects of Random Measurement Error in Regression

Although research on the effects of random measurement errors in regression analysis has a fairly long history (see Pedhazur, 1997, for a brief review) and the effects of measurement errors on the validity of regression analysis can be severe (Cochran, 1968). Jencks et al. (1972) suggested that "The most frequent approach to measurement error is indifference" (p. 330). Despite this apparent indifference in much of the applied research that utilizes regression analysis, the effects of random measurement errors (in either the criterion variable or the regressors) are known to result in a downward bias in the estimation of ρ^2 (Cochran, 1970). In addition, measurement error in the regressors in multiple regression models leads to bias (either positive or negative bias) in the regression coefficients (Cochran, 1968).

Two parameters are of interest when regression is based on predictor variables measured with error. One parameter is the population squared multiple correlation that would have been obtained if the regressors were measured perfectly $\left(\rho^2\right)$. This parameter, which represents a disattenuated multiple correlation, is primarily of interest in explanatory applications of regression, in which researchers are investigating relationships among variables for their theoretical importance. The second parameter of interest is the population squared multiple correlation that would be obtained using the fallible regressors themselves$\left(\rho_*^2\right)$. This parameter is primarily of interest in predictive applications of regression analysis, in which researchers are interested in the predictive power of the regressors as they are measured (i.e., including their measurement error).

Methods of Interval Estimation

Although interval estimates have been infrequently used in drawing inferences about the population squared multiple correlation$\left(\rho^2\right)$, several methods of constructing confidence bands are available. Fisher's (1928) derivation of the density function of R^2 has been implemented by Steiger and Fouladi (1992), using the interval inversion approach. This numerical method evaluates the cumulative distribution function of the sample R^2, given a population value of ρ^2. The method seeks that value of ρ^2 for which the obtained sample R^2 or smaller is expected (for example) 2.5% of the time and 97.5% of the time. These values of ρ^2 provide the endpoints of a 95% confidence band around the sample value of R^2.

Olkin and Finn (1995) provided several methods that were used by Algina (1999) to estimate confidence bands for the squared multiple correlation. The first method uses an estimated variance of R^2, given by

$$\sigma_{R^2}^2 = \frac{4\rho^2\left(1-\rho^2\right)\left(n-k-1\right)^2}{\left(n^2-1\right)\left(n+3\right)}$$

where n = sample size, and
k = number of regressors in the model.

A confidence interval is obtained by substituting the sample R^2 for ρ^2 in the equation (yielding $S_{R^2}^2$) and using $R^2 \pm z_{\alpha/2} S_{R^2}$ to obtain the endpoints of the confidence band.

A second method suggested by Olkin and Finn (1995) provides an estimate of the variance using

$$\dot{\sigma}_{R^2}^2 = \frac{4\rho^2\left(1-\rho^2\right)^2}{n}$$

As with the first method, a sample squared multiple correlation is used to obtain the estimated variance and a confidence band is constructed using the normal distribution.

The third method suggested by Olkin and Finn (1995) uses Fisher's z transformation of the multivariate R to normalize its distribution, resulting in a transformed variable with a variance of 4/n. That is,

$$z^* = \log_e\left(\frac{1+R}{1-R}\right)$$

The endpoints of a confidence band for the z* are given by

$$z^* \pm \frac{2z_{\alpha/2}}{\sqrt{n}}$$

If both endpoints (z_i) of this confidence band are non-negative, the endpoints are transformed to provide endpoints of the confidence band for ρ^2

$$\left[\frac{\exp\left(z_i\right)-1}{\exp\left(z_i\right)+1}\right]^2$$

If the lower endpoint of the confidence band for z* is negative, then the lower endpoint for the band for ρ^2 is set to zero.

The approximations presented by Olkin and Finn (1995) present problems when applied to samples from populations in which the squared multiple correlation is close to zero (Kendall and Stuart, 1977), and an investigation by Lee (1971) suggested that the Fisher transformation method worked poorly unless n was large relative to k. Recent work by Algina (1999) suggests that all of these approximations work poorly in comparison to the inversion method suggested by Steiger and Fouladi.

Previous studies on these estimates have included multivariate normal and non-normal data with perfectly reliable predictors. In reality, predictors in the social sciences have varying degrees of reliability and our intent was to investigate the performance of these confidence bands considering different levels of reliability among predictor variables.

Method

The confidence band estimates were constructed and compared using Monte Carlo methods, in which random samples were generated under known and controlled population conditions. In this Monte Carlo study, samples were generated from multivariate populations and each confidence band estimate was calculated based on each sample.

The Monte Carlo study included six factors in the design. These factors were (a) the true population multiple correlation (with ρ^2 = 0.01, 0.05, 0.10, 0.30, and 0.60), (b) number of regressor variables (with k = 2, 4, and 8), (c) sample sizes (with n = 5*k, 10*k, and 50*k), and (d) population distribution shape (conditions in which each variable evidenced population skewness and kurtosis values, respectively, of 0,0 [i.e., normal distribution]; 1,3; 1.5,5; 2,6; and 0,25), (e) regressor intercorrelation (ρ_{12} = 0.2, 0.4, 0.6, 0.8, 1.0), and (f) regressor reliability (ρ_{xx} = 0.40, 0.60, 0.80, 0.90, and 1.00).

Measurement error was simulated in the data following procedures used by Maxwell, Delaney, and Dill (1984), Jaccard and Wan (1995), and Kromrey and Foster-Johnson (1999). In this method, two normally distributed random variables for each regressor are generated, one of which represents the ‘true score’ on the regressor, the other representing measurement error. Fallible, observed scores on the regessors were calculated as the sum of the true and error components, consistent with classical measurement theory. The reliabilities of the regressors were controlled by adjusting the error variance relative to the true score variance by:

$$\rho_{xx} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2},$$

where σ_T^2 and σ_E^2 are the true and error variances, respectively, and ρ_{xx} is the reliability.

The research was conducted using SAS/IML version 8.1. Conditions for the study were run under Windows 98. Normally distributed random variables were generated using the RANNOR random number generator in SAS. A different seed value for the random number generator was used in each execution of the program. The program code was verified by hand-checking results from benchmark datasets.

For conditions involving nonnormal population distributions, the nonnormal data were produced by transforming the normal random variates obtained from RANNOR using the technique described by Bradley and Fleisher (1994), and operationalized by Ferron, *Yi*, and Kromrey (1997). In this method, a population correlation matrix, R, with a multivariate non-normal shape is constructed by an iterative process in which large simulated samples (*n* = 100,000) are generated from an approximation of R, $\tilde{R}$. The correlation matrix estimated from this large sample $(\hat{R})$ is compared elementwise to R, and the residuals $(R - \hat{R})$ are used to adjust the generating matrix $\tilde{R}$. This sequence of large sample generation, matrix estimation, and adjustment of $\tilde{R}$ continues until the process converges. The resulting matrix, $\tilde{R}$, is used to generate correlated non-normal data for the Monte Carlo study.

For each condition investigated in this study, 10,000 samples were generated. The use of 10,000 estimates provides adequate precision for the investigation of the sampling behavior of these confidence bands. For example, 10,000 samples provides a maximum 95% confidence interval width around an observed proportion that is ± .0098 (Robey & Barcikowski, 1992).

The relative performance of the confidence band estimates was evaluated by a comparison of the confidence band coverage (the proportion of confidence bands that included the population parameter) and the average width of the confidence bands. These indices correspond to statistical bias and estimation precision.

Results

The results were analyzed in terms of the confidence band coverage probabilities for the two parameters of interest, the population squared multiple correlation that would have been obtained with regressors measured without error $\left(\rho^2\right)$, and the squared multiple correlation in the population based upon the fallible regressors actually used $\left(\rho_*^2\right)$. In addition, the widths of the resulting confidence bands were analyzed. With the exception of tables for overall results for different reliability conditions for each of the three parameters of interest, in the interest of space and efficiency, other tables only contain results for the lowest, middle, and highest reliabilities investigated (0.4, 0.8 and 1.0) when three specific factors of the study design (shape, population squared multiple correlation, and sample size) are discussed. Many of the figures provided to illustrate results reflect conditions with the other reliabilities to maximize information. Specific results for reliabilities of 0.6 and 0.9 may be obtained from the authors if desired. Results for the different regressor intercorrelations did not show appreciable differences and are therefore not included in the detailed discussion.

Confidence Band Coverage of ρ^2

Figure 1 presents the distributions of coverage probabilities for the population squared multiple correlation based upon perfectly reliable regressors (i.e., ρ^2) across all conditions in this research. The band coverage for this parameter is rather poor under most conditions using a 95% confidence interval with extremely low coverage for specific conditions. The Steiger and Fouladi method provided the best band coverage overall, and the Olkin and Finn 3 method had notably poorer performance when compared to Olkin and Finn methods 1 and 2. Additional analyses of these coverage probabilities were addressed by considering the average coverage probabilities for differences in reliability, shape, population squared multiple correlation, and sample size relative to the number of regressors.

Reliability. Table 1 presents the distributions of coverage probabilities across all five reliabilities reflected in the conditions in this research. These values indicate the likelihood that ρ^2 will fall within the confidence interval when the fallibility of the regressors is not taken into account. Regardless of the method employed, all improve as reliability improves and the Steiger and Fouladi method consistently outperforms the other three methods, in spite of the number of regressors.

Distribution Shape. In Table 2, the impact of distribution shape on the ability of the different techniques to provide adequate coverage is explored based on the number of regressors as well as the

Figure 1. Distribution of confidence band coverage for ρ^2.

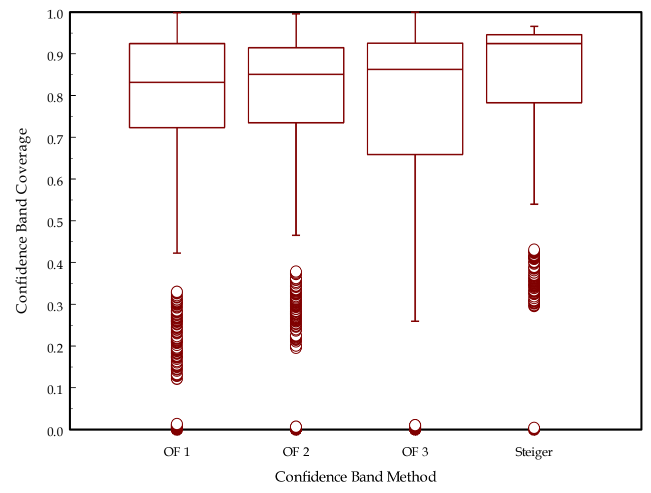


Table 1. Confidence Band Coverage for ρ^2 by Number of Regressors and Measurement Reliability

		Reliability				
	Method	0.4	0.6	0.8	0.9	1.0
2 Regressors	**O&F 1**	0.64	0.71	0.81	0.86	0.87
	O&F 2	0.67	0.74	0.82	0.86	0.86
	O& F 3	0.84	0.83	0.86	0.89	0.88
	S&F	0.71	0.77	0.86	0.91	0.92
4 Regressors	**O&F 1**	0.63	0.70	0.82	0.87	0.88
	O&F 2	0.66	0.72	0.82	0.86	0.86
	O& F 3	0.62	0.68	0.76	0.80	0.80
	S&F	0.63	0.70	0.81	0.88	0.92
8 Regressors	**O&F 1**	0.56	0.67	0.76	0.78	0.81
	O&F 2	0.60	0.68	0.77	0.79	0.80
	O& F 3	0.54	0.62	0.69	0.70	0.72
	S&F	0.55	0.70	0.78	0.88	0.92

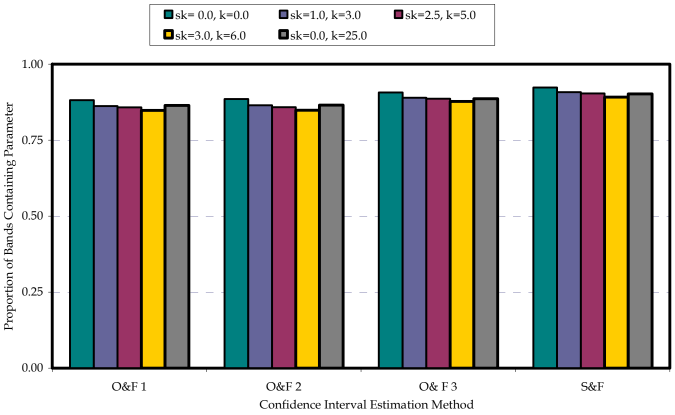


Figure 2. Proportion of confidence bands containing ρ^2 by distribution shape. Two regressors, reliability = 0.9

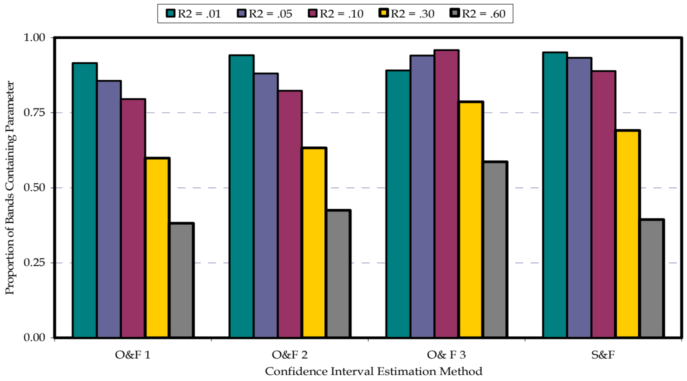


Figure 3. Proportion of confidence bands containing ρ^2 by ρ^2 . Two regressors, reliability = 0.6.

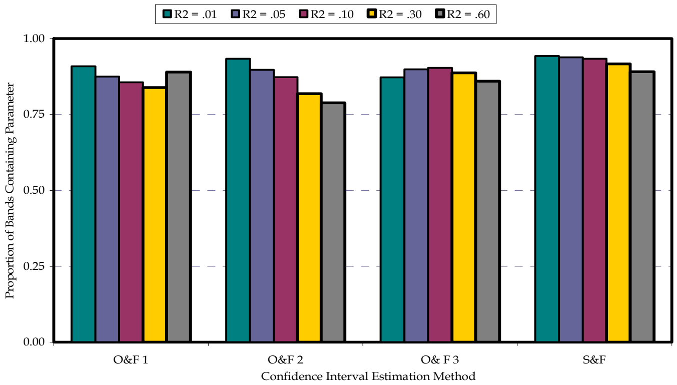


Figure 4. Proportion of confidence bands containing ρ^2 by ρ^2 . Two regressors, reliability = 1.0

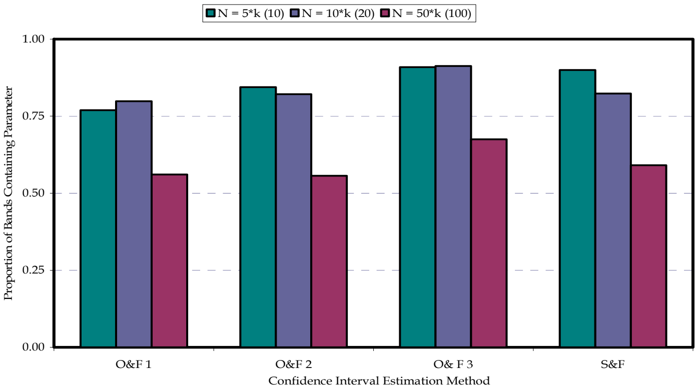


Figure 5. Proportion of confidence bands containing ρ^2 by sample size . Two regressors, reliability = 0.6.

reliability. An examination of these data as well as representation of coverage for 2 regressors with a reliability of 0.9 in Figure 2 indicates that the distribution shape has very little effect on the confidence band coverage within each method and that none of the methods provide very good coverage when ρ^2 is estimated.

Coverage is especially poor when the number of regressors is high (k=8) and reliability low (r = 0.4) with the Steiger and Fouladi method providing a consistent coverage probability of 0.55 at the low end and Olkin and Finn method 2 providing an almost consistent rate at approximately 0.60 across shapes. No method was consistently better across shapes with ρ^2, although the Steiger and Fouladi method did provide notably better coverage in a few isolated instances, i.e., when k = 8, r = 1.0 Steiger and Fouladi averaged a 0.91 coverage rate compared to the next best method, Olkin & Finn 1 which had a 0.80 coverage rate. Such instances of such clear superior performance were few.

Population Squared Multiple Correlation. When coverage was examined as a function of the population squared multiple correlation, there was a notable difference in how well the different methods performed. Table 3 provides estimated coverage for all conditions with reliabilities of 0.4, 0.8, and 1.0. Figures 3 and 4 illustrate these results for two specific conditions with 2 regressors (r = 0.6 and r = 1.0, respectively). When reliability is 0.6, Olkin and Finn method 3 tends to perform slightly better for higher values of ρ^2, however, as reliability increases, this slight superiority diminishes.

Table 2. Confidence Band Coverage for ρ^2 by Number of Regressors, Measurement Reliability and Distribution Shape.

			Distribution Shape				
	Reliability	Method	Sk = 0.0 Kurt = 0.0	Sk = 1.0 Kurt = 3.0	Sk = 1.5 Kurt = 5.0	Sk = 2.0 Kurt = 6.0	Sk = 0.0 Kurt = 25.0
2 Regressors	0.4	O&F 1	0.64	0.64	0.63	0.64	0.63
		O&F 2	0.68	0.68	0.67	0.68	0.67
		O& F 3	0.83	0.84	0.84	0.84	0.84
		S&F	0.71	0.71	0.70	0.71	0.70
	0.8	O&F 1	0.83	0.82	0.81	0.80	0.81
		O&F 2	0.84	0.83	0.82	0.81	0.82
		O& F 3	0.87	0.87	0.86	0.86	0.86
		S&F	0.87	0.86	0.85	0.85	0.85
	1.0	O&F 1	0.90	0.88	0.86	0.85	0.88
		O&F 2	0.89	0.87	0.85	0.83	0.87
		O& F 3	0.91	0.89	0.87	0.86	0.89
		S&F	0.95	0.93	0.92	0.90	0.93
4 Regressors	0.4	O&F 1	0.63	0.61	0.63	0.63	0.66
		O&F 2	0.66	0.64	0.66	0.66	0.69
		O& F 3	0.62	0.62	0.62	0.63	0.63
		S&F	0.62	0.60	0.63	0.63	0.66
	0.8	O&F 1	0.83	0.82	0.81	0.80	0.82
		O&F 2	0.83	0.81	0.81	0.80	0.82
		O& F 3	0.77	0.77	0.76	0.74	0.74
		S&F	0.82	0.81	0.81	0.80	0.82
	1.0	O&F 1	0.91	0.89	0.88	0.86	0.88
		O&F 2	0.89	0.86	0.85	0.83	0.87
		O& F 3	0.84	0.83	0.80	0.78	0.78
		S&F	0.95	0.93	0.91	0.90	0.92
8 Regressors	0.4	O&F 1	0.56	0.56	0.56	0.56	0.56
		O&F 2	0.60	0.60	0.60	0.60	0.59
		O& F 3	0.54	0.54	0.54	0.54	0.54
		S&F	0.55	0.55	0.55	0.55	0.55
	0.8	O&F 1	0.77	0.76	0.75	0.75	0.75
		O&F 2	0.78	0.77	0.76	0.76	0.76
		O& F 3	0.70	0.69	0.68	0.68	0.67
		S&F	0.79	0.78	0.78	0.77	0.77
	1.0	O&F 1	0.84	0.82	0.81	0.80	0.80
		O&F 2	0.83	0.81	0.79	0.78	0.79
		O& F 3	0.75	0.73	0.72	0.71	0.71
		S&F	0.95	0.93	0.92	0.90	0.91

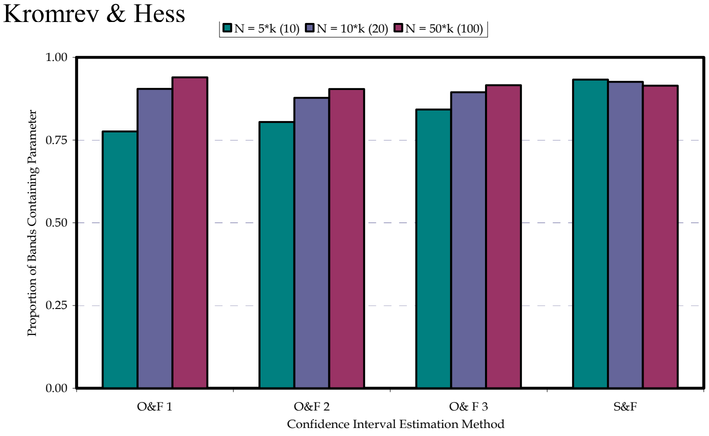
Again, the poorest results were evident when low reliability and large numbers of regressors were involved. Effectiveness of the different methods varied. The Steiger and Fouladi method tended to do better when larger numbers of regressors were considered and reliability was 0.8 or 1.0 regardless of the population squared multiple correlation coefficient. Olkin & Finn 3, performed better than the others when reliability was low. When considering a ρ^2 of 0.6, O&F3 had a coverage of 0.60 when reliability was 0.4 with two regressors, compared with 0.31 by O&F 2, 0.27 by S & F, and 0.25 by O&F 1.

Table 3. Confidence Band Coverage for ρ^2 by Number of Regressors, Measurement Reliability and ρ^2.

			Population Squared Multiple Correlation				
	Reliability	Method	$\rho^2 = 0.01$	$\rho^2 = 0.05$	$\rho^2 = 0.1$	$\rho^2 = 0.3$	$\rho^2 = 0.6$
2 Regressors	0.4	O&F 1	0.92	0.83	0.73	0.46	0.25
		O&F 2	0.94	0.86	0.76	0.51	0.31
		O& F 3	0.90	0.95	0.97	0.77	0.60
		S&F	0.95	0.91	0.83	0.57	0.27
	0.8	O&F 1	0.91	0.87	0.84	0.77	0.66
		O&F 2	0.94	0.90	0.87	0.78	0.63
		O& F 3	0.88	0.92	0.94	0.88	0.69
		S&F	0.95	0.94	0.93	0.85	0.61
	1.0	O&F 1	0.91	0.88	0.86	0.84	0.89
		O&F 2	0.93	0.90	0.87	0.82	0.79
		O& F 3	0.87	0.90	0.90	0.89	0.86
		S&F	0.94	0.94	0.93	0.92	0.89
4 Regressors	0.4	O&F 1	0.89	0.85	0.74	0.48	0.19
		O&F 2	0.93	0.87	0.75	0.51	0.22
		O& F 3	0.66	0.88	0.81	0.58	0.20
		S&F	0.95	0.87	0.74	0.45	0.12
	0.8	O&F 1	0.89	0.88	0.87	0.80	0.64
		O&F 2	0.92	0.91	0.88	0.79	0.59
		O& F 3	0.64	0.86	0.89	0.83	0.57
		S&F	0.95	0.93	0.92	0.79	0.50
	1.0	O&F 1	0.88	0.87	0.87	0.87	0.92
		O&F 2	0.92	0.89	0.87	0.83	0.80
		O& F 3	0.63	0.83	0.86	0.86	0.84
		S&F	0.94	0.94	0.93	0.91	0.89
8 Regressors	0.4	O&F 1	0.76	0.79	0.69	0.46	0.12
		O&F 2	0.84	0.83	0.70	0.49	0.13
		O& F 3	0.56	0.80	0.72	0.53	0.11
		S&F	0.94	0.80	0.64	0.32	0.04
	0.8	O&F 1	0.73	0.81	0.84	0.80	0.60
		O&F 2	0.82	0.85	0.86	0.77	0.54
		O& F 3	0.52	0.77	0.84	0.79	0.52
		S&F	0.95	0.93	0.90	0.72	0.40
	1.0	O&F 1	0.77	0.77	0.80	0.86	0.92
		O&F 2	0.81	0.81	0.81	0.81	0.78
		O& F 3	0.71	0.70	0.77	0.82	0.82
		S&F	0.94	0.94	0.93	0.91	0.88

Sample Size. Table 4 provides coverage results by sample size. Coverage varied again across reliabilities and regressors, although there was little difference in coverage for conditions with 2 or 4 regressors under similar conditions; however, 8 regressor conditions typically showed a decrease.

At low reliabilities, O&F3 again did slightly better than the others, although it quickly lost ground as reliability increased. All methods performed very poorly when sample sizes were large and reliabilities were less than 0.9. For the reliability of 0.6 with two regressors, O&F 3 outperformed the others, as illustrated in Figure 5. However, this method was typically outperformed by the other Olkin and Finn methods as well as the Steiger and Fouladi methods under higher reliabilities as evident by Figure 6 which illustrates results for two regressors under perfect reliability conditions. At a sample size of 10*k for k = 4 and a reliability of 0.8, the O&F 3 method had a coverage rate of 0.81 while the others had approximately a 90% coverage rate.



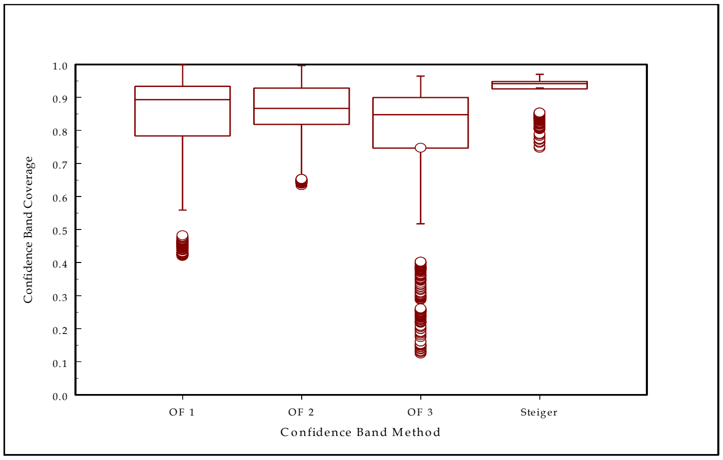
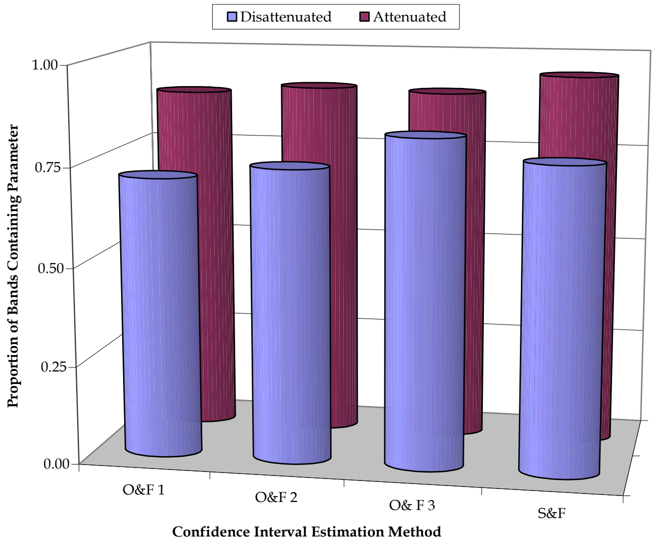


Figure 6. Proportion of confidence bands containing ρ_*^2 by sample size. Two regressors, reliability = 1.0

Figure 7. Distribution of confidence band coverage for ρ_*^2



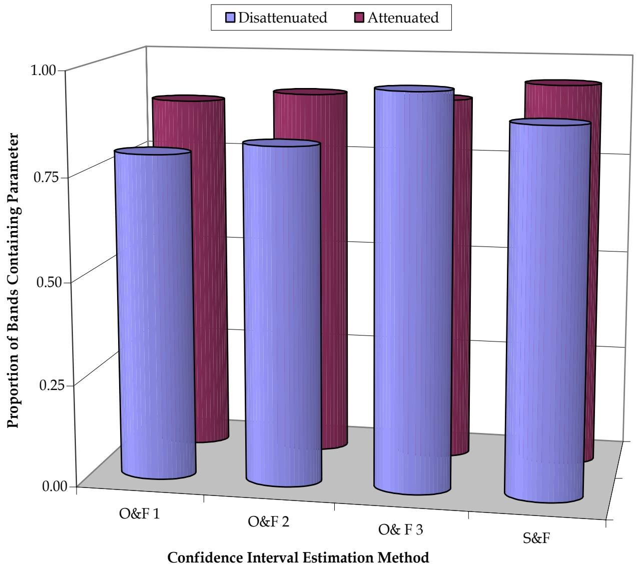


Figure 8. Comparison of the Proportion of Confidence Bands Containing ρ_*^2 , k=2, r = 0.6, sk = 0.0 & kurt = 0.0.

Figure 9. Comparison of the Proportion of ConfidenceBands Containing ρ_*^2 , k=2, r = 0.6, and ρ_*^2 = 0.1.

Confidence Band Coverage of ρ_*^2

Figure 7 presents the distributions of coverage probabilities for the population squared multiple correlation when the fallibility of the regressors was taken into account (i.e., ρ_*^2). The use of 95% confidence bands based upon the Steiger and Fouladi method continued to provide the best band coverage overall. The Olkin and Finn method 2 had the next best results, although Figure 7 clearly shows distinctly poorer performance than the Steiger and Fouladi method. Olkin and Finn methods 1 and 3 present many conditions with poor band coverage, with method 3 providing extremely low coverage for some conditions. Additional analyses of these coverage probabilities were addressed by considering the average coverage probabilities for each factor in the Monte Carlo study design

Reliability. Table 5 presents the distributions of coverage probabilities for the parameter ρ_*^2 by number of regressors and reliability. When one compares and contrasts the values in Table 1 with the values in Table 5, it is readily apparent how coverage is improved when the fallibility of the regressors is taken into account. Once again, the Steiger and Fouladi method outperforms the other three under all conditions, with coverage ranging from 92% to 94% throughout the conditions. The other three

Table 4. Confidence Band Coverage for ρ ^{2} by Number of Regressors, Measurement Reliability and Sample Size.

			Sample Size		
	Reliability	Method	N=5*k	N=10*k	N=50*k
2 Regressors	0.4	O&F 1	0.74	0.73	0.44
		O&F 2	0.83	0.75	0.44
		O& F 3	0.92	0.92	0.67
		S&F	0.87	0.76	0.49
	0.8	O&F 1	0.79	0.88	0.77
		O&F 2	0.85	0.89	0.74
		O& F 3	0.89	0.92	0.78
		S&F	0.93	0.90	0.74
	1.0	O&F 1	0.78	0.90	0.94
		O&F 2	0.80	0.88	0.90
		O& F 3	0.84	0.89	0.92
		S&F	0.93	0.93	0.91
4 Regressors	0.4	O&F 1	0.74	0.72	0.43
		O&F 2	0.82	0.73	0.44
		O& F 3	0.71	0.67	0.49
		S&F	0.80	0.68	0.41
	0.8	O&F 1	0.81	0.90	0.74
		O&F 2	0.85	0.89	0.70
		O& F 3	0.76	0.81	0.71
		S&F	0.81	0.90	0.74
	1.0	O&F 1	0.78	0.91	0.95
		O&F 2	0.79	0.88	0.91
		O& F 3	0.71	0.80	0.89
		S&F	0.93	0.92	0.91
8 Regressors	0.4	O&F 1	0.64	0.65	0.40
		O&F 2	0.72	0.67	0.40
		O& F 3	0.60	0.59	0.43
		S&F	0.72	0.60	0.32
	0.8	O&F 1	0.71	0.85	0.71
		O&F 2	0.78	0.84	0.68
		O& F 3	0.66	0.73	0.66
		S&F	0.89	0.82	0.63
	1.0	O&F 1	0.65	0.85	0.94
		O&F 2	0.69	0.82	0.90
		O& F 3	0.60	0.72	0.85
		S&F	0.93	0.92	0.91

performed similarly when conditions only called for two regressors, however, the Olkin and Finn 3 quickly fell behind in effectiveness as the number of regressors increased, i.e., when k = 4 and r = 0.4, the confidence band coverage probabilities for the methods, from best to worst were: (1) 0.94 for S&F, (2) 0.90 for O&F 2, (3) 0.88 for O&F 1, and (4) 0.69 for O&F 3. When the number of regressors increased to 8, the Steiger and Fouladi method maintained its performance in the mid 90's, however, all of the other three fell even more in their coverage rates, with O&F3 continuing to have much poorer performance than the other two Olkin and Finn methods which consistently had similar performance (Table 5).

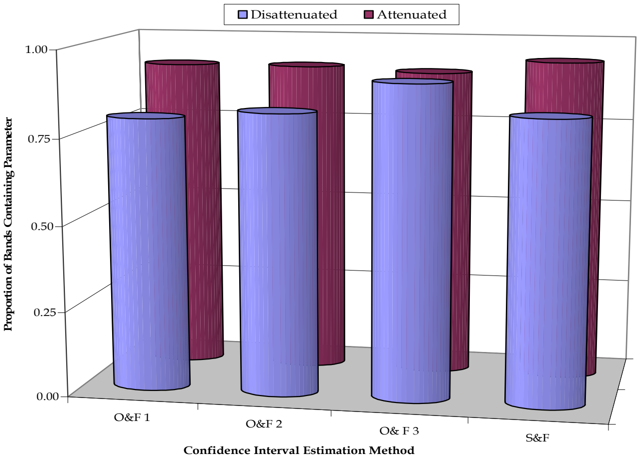


Figure 10. Comparison of the Proportion of Confidence Bands Containing ρ^2 and ρ_*^2 k=2, r = 0.6, N = 10*k (20).

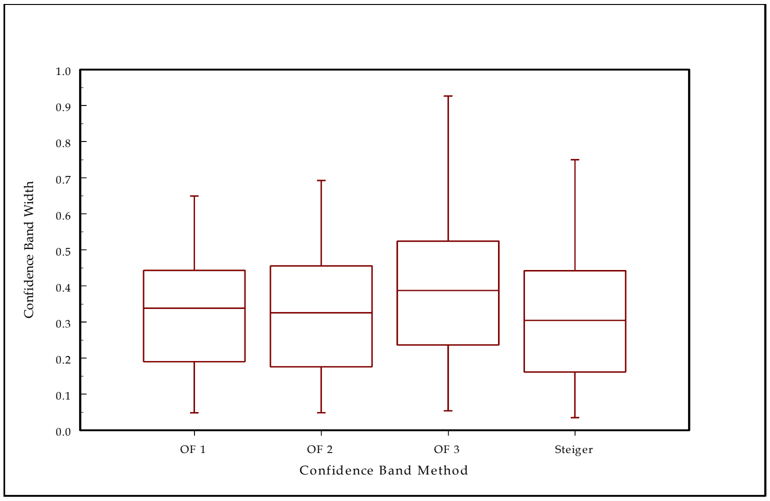


Figure 11. Distribution of Confidence Band Widths

Distribution Shape. Confidence band coverage probabilities for ρ_*^2 as a function of distribution shape is shown in Table 6. As with the confidence coverage of ρ^2, shape does not seem to have an appreciable effect on coverage of this parameter. However, once again, the Steiger and Fouladi method is superior to the other two under all conditions. The Olkin and Finn 3 method consistently provides the poorest performance. Figure 8 illustrates the improved coverage by all methods for one condition which is fairly representative of all conditions. When error in the regressors is accounted for, all methods have coverage's much closer to 95% although Steiger and Fouladi is still superior. Olkin and Finn 3 shows the least amount of improvement although it had the better coverage when regressor fallibility was not considered.

Population Squared Multiple Correlation. Table 7 and Figure 9 present coverage probabilities when considering differences in the population squared multiple correlation. Once again the Steiger and Fouladi method has consistently superior results when measurement error is accounted for. Mean coverages for this method range from 0.91 to 0.95. The three Olkin and Finn methods have similar performance when reliabilities are high; However, all three show dramatic drops in coverage when reliability falls to 0.4 and the number of regressors is greater than 2, with the Olkin and Finn method 3 performing very poorly even for high reliabilities under conditions containing eight regressors. When reliability is 0.8 and the number of regressors is 8, O&F 3 shows a coverage of only 45%, compared to 71% for O&F 1, 80% for O&F 2, and 95% for S&F. Once again, coverage is improved dramatically when results from ρ_*^2 are compared with ρ^2 as illustrated in Figure 9, at least for three of the four methods. Olkin and Finn 3, the least consistent of the methods actually shows comparable coverage whether ρ_*^2 or ρ^2 is considered under these conditions.

Sample Size. As sample size increases, so does the coverage for ρ_*^2 as clearly shown in Table 8. Steiger and Fouladi remains fairly consistent across sample size, regardless of the number of regressors or the reliability. However, Olkin and Finn 3, while somewhat satisfactory for conditions with large samples and a small number of regressors (0.93 when n= 50*k, r = 0.8, and k = 2), drops quite a bit with smaller sample sizes and a large number of regressors (0.57 when n = 5*k. r = 0.8, and k = 8). Olkin and Finn 1 and 2 also show relatively poor performance for small samples sizes (0.77 and 0.81, respectively when n = 5*k, r = 0.8, and k = 4). Once again, coverage is clearly improved across sample sizes when ρ_*^2 is estimated, rather than ρ^2. Figure 10 shows how all methods are close to 0.95 coverage when rho^2 is attenuated, with marked improvement for the Olkin and Finn 1 and 2 methods as well as the Steiger and Fouladi method. Interestingly, there was very little improvement for the Olkin and Finn 3 method.

Table 5.*Confidence Band Coverage for ρ_*^2 by Measurement Reliability.*

		Reliability				
	Method	0.4	0.6	0.8	0.9	1.0
2 Regressors	O&F 1	0.89	0.88	0.87	0.88	0.87
	O&F 2	0.92	0.90	0.88	0.87	0.86
	O& F 3	0.88	0.89	0.89	0.89	0.88
	S&F	0.94	0.94	0.93	0.93	0.92
4 Regressors	O&F 1	0.88	0.88	0.88	0.88	0.88
	O&F 2	0.90	0.89	0.88	0.87	0.86
	O& F 3	0.69	0.77	0.79	0.80	0.80
	S&F	0.94	0.94	0.93	0.93	0.92
8 Regressors	O&F 1	0.76	0.79	0.80	0.81	0.81
	O&F 2	0.81	0.81	0.81	0.80	0.80
	O& F 3	0.60	0.67	0.70	0.71	0.72
	S&F	0.94	0.94	0.93	0.93	0.92

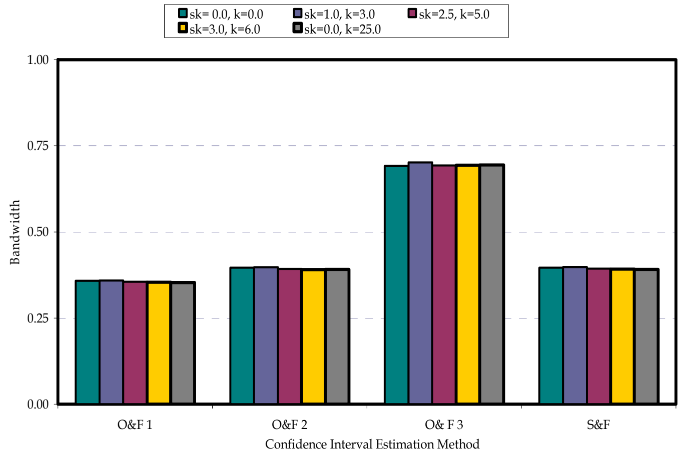


Figure 12. Width of confidence band by distribution shape. Two regressors, reliability = 0.6.

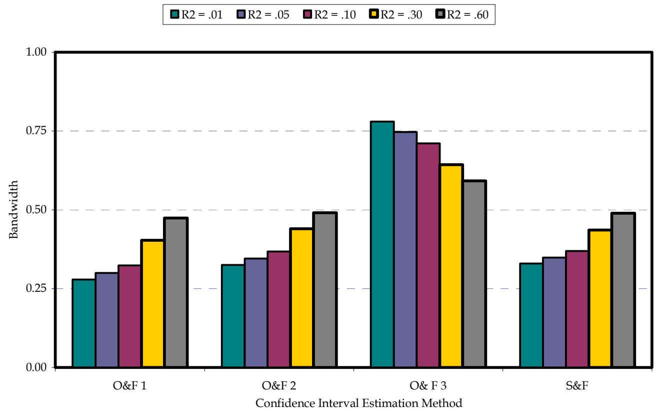


Figure 13. Width of confidence band by ρ^2. Two regressors, reliability = 0.6.

Confidence Band Width

Figure 11 presents the distributions of confidence band widths across all conditions in this research. This figure suggests that the Olkin and Finn method 3 produced slightly larger bands than the other methods. Additional analyses of these coverage probabilities were addressed by considering the average band widths for each factor in the Monte Carlo study design.

Reliability. Table 9 contains the bandwidths under different reliability conditions. In most cases, bandwidth increased as reliability did with the exception of the Olkin and Finn 3 method. This method showed either decreasing or approximately consistent bandwidths as reliability increased. When conditions contained 2 regressors, the bandwidth decreased by 0.11 when reliability changed from 0.4 to 1.0. Under conditions containing 4 and 8 regressors, bandwidths only varied 0.02 between different reliabilities. However, these isolated cases of width constriction and consistency are not of great consequence as they are either much larger than those provided by the other three methods (when k=2) or about the same as the other three methods (when k = 4 or k = 8). Steiger and Fouladi outperformed the other methods for conditions with a large number of regressors (width = 0.20 when k = 8, r = 0.4 compared to 0.24 or 0.25 for the others) and has widths that are either smaller than, or similar to, those constructed by the other methods for conditions with fewer regressors.

Distribution Shape. The confidence band widths were not appreciably related to the population distribution shape (Table 10). Across all distribution shapes, the average width of confidence bands constructed by Olkin and Finn method 3 were larger than those of the other methods, and the Olkin and Finn method 1 and Steiger and Fouladi approaches produced the smallest bands. The tendency for bandwidths constructed by Olkin and Finn 3 to be larger than the others is especially evident for conditions with only two regressors (Figure 12). As the number of regressors increased, the widths resulting from this method were very similar to those obtained by the other two Olkin and Finn methods, and usually just slightly larger than those constructed by the Steiger and Fouladi method.

Population Squared Multiple Correlation. The relationship between average band width and ρ^2 is presented in Table 11. All methods typically showed progressively larger bands as the value of the parameter increased, although there were some exceptions. As the number of regressors increased, the bandwidths tended to tighten a bit. For low values of ρ^2, the Steiger method was typically the best performer. When k=8, r = 0.4 and ρ^2 = 0.01, the Steiger method had a mean bandwidth of 0.16 while Olkin and Finn methods 1, 2, and 3, and widths of 0.20, 0.22, and 0.23 respectively. However, as ρ^2 increased (specifically 0.3 and 0.6), Steiger tended to provide bandwidths slightly larger than those produced by Olkin and Finn methods 2 and 3.

For conditions with a small number of regressors (k = 2), bands tended to get smaller when the Olkin and Finn 3 method was employed (Figure 13 is an example of this phenomena). However, when one takes into account the large bandwidth for smaller values of ρ^2 under these k = 2 conditions, this constriction does not provide any notable benefit compared to the other methods.

Sample Size. As anticipated, all of the methods produced smaller confidence intervals with larger samples (Table 12). The Olkin and Finn 1 and Steiger and Fouladi methods produced the smallest bands with small samples, but with samples of 10 or 50 times the number of regressors (10*k or 50*k) that occurred in conditions with either 4 or 8 regressors (k = 4 or k = 8), the difference in the confidence interval widths was negligible. However, when the number of regressors was small, k = 2, the bandwidths resulting from the Olkin and Finn method 3 were wider (Figure 14).

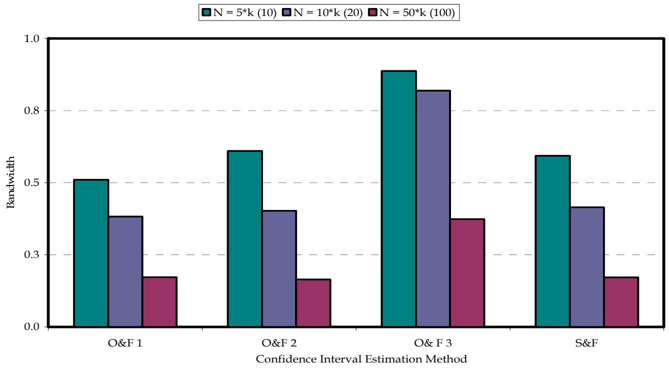


Figure 14. Width of confidence band by sample size. Two regressors, reliability = 0.6.

Conclusions

The general superiority of the Steiger and Fouladi approach to confidence band construction is evident in these results. This approach provided consistently more accurate confidence intervals across most of the conditions examined. Further, these more accurate bands were obtained without substantially increasing the confidence band width. Such results are consistent with a previous comparison of these methods with normal and non-normal data measured without error (Kromrey & Hess, 2001).

However, notable differences were obtained between the accuracy of the confidence bands for the two parameters of interest in this study. The population squared multiple correlation obtained with fallible regressors $\left(\rho_*^2\right)$was accurately estimated with the Steiger and Fouladi bands, but the parameter that would be realized if the regressors were measured without error $\left(\rho^2\right)$ was poorly estimated in most conditions. The difference in these parameters and the differential success in estimating them, suggests that it is incumbent upon researchers to remain cognizant of the distinction. Concern for the psychometric characteristics and the consequences of poor reliability needs to become a part of the variable selection process in regression applications. We concur with others that unreliable regressors can and should be avoided (cf. Cohen & Cohen, 1983).

Obtaining accurate estimates of the parameter ρ^2 in the context of fallible regressors may require a disattenuation of the sample R^2 as a part of the confidence band construction. For example, Fuller and Hidirouglou (1978) described a correction for attenuation as a sample covariance matrix that is modified using reliabilities or error estimates obtained from a source independent of the sample covariance matrix.

However, the sampling distribution of a disattenuated ρ^2 will differ from that of the sample R^2, and the interval inversion approach of Steiger and Fouladi will need to be adjusted for these sampling characteristics. Further research in this direction is certainly in order.

It is becoming increasingly critical in today's high stakes educational environment to ensure that measurement methods and statistical analyses are as accurate and informative as possible. A continued reliance solely on the testing of null hypotheses in conjunction with unrealistic assumptions (i.e., normality; perfectly reliable predictors) limits the amount of information that we may glean from available information on students, teachers, methods and the myriad of other elements that compose educational systems. Although hypothesis testing provides probabilistic information about the accuracy of rejecting a null hypothesis, the proper development and use of confidence intervals for correlation applications under less than perfect conditions will allow us to estimate the bounds within the parameter is expected to reside with increased accuracy. As a result of this expected increase in precision, practitioners will be better equipped to make critical decisions with greater confidence.

Table 6. Confidence Band Coverage for ρ_*^2 by Number of Regressors, Reliability & Distribution Shape.

			Distribution Shape				
	Reliability	Method	Sk = 0.0 Kurt = 0.0	Sk = 1.0 Kurt = 3.0	Sk = 1.5 Kurt = 5.0	Sk = 2.0 Kurt = 6.0	Sk = 0.0 Kurt = 25.0
2 Regressors	0.4	O&F 1	0.90	0.89	0.90	0.89	0.89
		O&F 2	0.92	0.92	0.92	0.91	0.91
		O& F 3	0.89	0.89	0.89	0.88	0.88
		S&F	0.95	0.95	0.94	0.94	0.94
	0.8	O&F 1	0.89	0.88	0.87	0.86	0.88
		O&F 2	0.90	0.88	0.87	0.87	0.88
		O& F 3	0.91	0.90	0.89	0.88	0.89
		S&F	0.95	0.94	0.93	0.92	0.93
	1.0	O&F 1	0.90	0.88	0.86	0.85	0.88
		O&F 2	0.89	0.87	0.85	0.83	0.87
		O& F 3	0.91	0.89	0.87	0.86	0.89
		S&F	0.95	0.93	0.92	0.90	0.93
4 Regressors	0.4	O&F 1	0.88	0.88	0.88	0.88	0.87
		O&F 2	0.91	0.90	0.90	0.90	0.90
		O& F 3	0.70	0.72	0.70	0.69	0.66
		S&F	0.95	0.95	0.94	0.94	0.93
	0.8	O&F 1	0.90	0.88	0.87	0.86	0.87
		O&F 2	0.90	0.88	0.87	0.86	0.88
		O& F 3	0.81	0.81	0.79	0.78	0.77
		S&F	0.95	0.94	0.92	0.92	0.92
	1.0	O&F 1	0.91	0.89	0.88	0.86	0.88
		O&F 2	0.89	0.86	0.85	0.83	0.87
		O& F 3	0.84	0.83	0.80	0.78	0.78
		S&F	0.95	0.93	0.91	0.90	0.92
8 Regressors	0.4	O&F 1	0.77	0.76	0.76	0.76	0.75
		O&F 2	0.82	0.81	0.81	0.81	0.80
		O& F 3	0.61	0.60	0.60	0.60	0.60
		S&F	0.95	0.94	0.94	0.94	0.93
	0.8	O&F 1	0.82	0.80	0.80	0.79	0.79
		O&F 2	0.83	0.81	0.80	0.80	0.80
		O& F 3	0.72	0.71	0.70	0.69	0.69
		S&F	0.95	0.93	0.92	0.91	0.91
	1.0	O&F 1	0.84	0.82	0.81	0.80	0.80
		O&F 2	0.83	0.81	0.79	0.78	0.79
		O& F 3	0.75	0.73	0.72	0.71	0.71
		S&F	0.95	0.93	0.92	0.90	0.91

Table 7. Confidence Band Coverage for ρ_*^2 by Number of Regressors, Measurement Reliability & ρ^2.

			Population Squared Multiple Correlation				
	Reliability	Method	$\rho^2 = 0.01$	$\rho^2 = 0.05$	$\rho^2 = 0.1$	$\rho^2 = 0.3$	$\rho^2 = 0.6$
2 Regressors	0.4	O&F 1	0.92	0.91	0.90	0.88	0.86
		O&F 2	0.95	0.94	0.93	0.90	0.87
		O& F 3	0.85	0.87	0.89	0.91	0.91
		S&F	0.95	0.95	0.95	0.94	0.94
	0.8	O&F 1	0.91	0.89	0.87	0.84	0.85
		O&F 2	0.94	0.91	0.89	0.84	0.82
		O& F 3	0.87	0.90	0.91	0.90	0.89
		S&F	0.95	0.94	0.94	0.93	0.91
	1.0	O&F 1	0.91	0.87	0.86	0.84	0.89
		O&F 2	0.93	0.90	0.87	0.82	0.79
		O& F 3	0.87	0.90	0.90	0.89	0.86
		S&F	0.94	0.94	0.93	0.92	0.89
4 Regressors	0.4	O&F 1	0.88	0.88	0.88	0.87	0.87
		O&F 2	0.92	0.92	0.91	0.89	0.87
		O& F 3	0.38	0.64	0.75	0.85	0.87
		S&F	0.95	0.95	0.95	0.94	0.93
	0.8	O&F 1	0.88	0.87	0.87	0.87	0.89
		O&F 2	0.92	0.90	0.89	0.85	0.83
		O& F 3	0.56	0.80	0.85	0.87	0.87
		S&F	0.95	0.94	0.94	0.92	0.90
	1.0	O&F 1	0.88	0.87	0.87	0.87	0.92
		O&F 2	0.92	0.89	0.88	0.83	0.80
		O& F 3	0.62	0.82	0.86	0.86	0.85
		S&F	0.94	0.94	0.93	0.91	0.89
8 Regressors	0.4	O&F 1	0.70	0.72	0.75	0.79	0.83
		O&F 2	0.80	0.80	0.81	0.82	0.82
		O& F 3	0.31	0.52	0.62	0.75	0.81
		S&F	0.95	0.95	0.94	0.93	0.92
	0.8	O&F 1	0.71	0.76	0.79	0.84	0.89
		O&F 2	0.80	0.81	0.82	0.82	0.80
		O& F 3	0.45	0.67	0.75	0.82	0.83
		S&F	0.95	0.94	0.94	0.92	0.89
	1.0	O&F 1	0.77	0.77	0.80	0.86	0.92
		O&F 2	0.81	0.81	0.81	0.81	0.78
		O& F 3	0.71	0.70	0.77	0.82	0.82
		S&F	0.94	0.94	0.93	0.91	0.88

Table 8. Confidence Band Coverage for ρ_*^2 by Number of Regressors, Measurement Reliability and Sample Size.

			Sample Size		
	Reliability	Method	N=5*k	N=10*k	N=50*k
2 Regressors	0.4	O&F 1	0.79	0.94	0.95
		O&F 2	0.86	0.94	0.95
		O& F 3	0.82	0.89	0.94
		S&F	0.95	0.95	0.94
	0.8	O&F 1	0.78	0.91	0.94
		O&F 2	0.83	0.90	0.92
		O& F 3	0.85	0.91	0.93
		S&F	0.94	0.94	0.92
	1.0	O&F 1	0.78	0.90	0.94
		O&F 2	0.80	0.88	0.90
		O& F 3	0.84	0.89	0.92
		S&F	0.93	0.93	0.91
4 Regressors	0.4	O&F 1	0.75	0.92	0.96
		O&F 2	0.83	0.92	0.95
		O& F 3	0.55	0.68	0.85
		S&F	0.95	0.94	0.94
	0.8	O&F 1	0.77	0.91	0.95
		O&F 2	0.81	0.90	0.92
		O& F 3	0.69	0.79	0.89
		S&F	0.94	0.93	0.92
	1.0	O&F 1	0.78	0.91	0.95
		O&F 2	0.79	0.88	0.91
		O& F 3	0.71	0.80	0.89
		S&F	0.93	0.92	0.91
8 Regressors	0.4	O&F 1	0.54	0.81	0.93
		O&F 2	0.68	0.83	0.92
		O& F 3	0.43	0.58	0.79
		S&F	0.94	0.94	0.93
	0.8	O&F 1	0.63	0.84	0.93
		O&F 2	0.70	0.83	0.90
		O& F 3	0.57	0.70	0.85
		S&F	0.93	0.93	0.92
	1.0	O&F 1	0.65	0.85	0.94
		O&F 2	0.69	0.82	0.90
		O& F 3	0.60	0.72	0.85
		S&F	0.93	0.92	0.91

Table 9. Width of Confidence Band by Measurement Reliability.

		Reliability				
	Method	0.4	0.6	0.8	0.9	1.0
2 Regressors	O&F 1	0.32	0.36	0.39	0.40	0.41
	O&F 2	0.36	0.39	0.41	0.42	0.41
	O& F 3	0.73	0.69	0.66	0.64	0.62
	S&F	0.36	0.39	0.42	0.43	0.43
4 Regressors	O&F 1	0.29	0.32	0.34	0.35	0.35
	O&F 2	0.31	0.33	0.34	0.34	0.36
	O& F 3	0.32	0.33	0.34	0.34	0.33
	S&F	0.27	0.30	0.32	0.33	0.33
8 Regressors	O&F 1	0.24	0.25	0.27	0.26	0.27
	O&F 2	0.25	0.25	0.27	0.27	0.26
	O& F 3	0.25	0.26	0.26	0.27	0.26
	S&F	0.20	0.23	0.24	0.25	0.24

References

Algina, J. (1999). A comparison of methods for constructing confidence intervals for the squared multiple correlation coefficient. *Multivariate Behavioral Research*, *34*, 493-504.

Bradley, D. R. & Fleisher, C. L. (1994). Generating multivariate data from nonnormal distributions: Mihal and Barrett revisited. *Behavior Research Methods, Instruments, and Computers*, *26*, 156-166.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics*, *10*, 637 – 666.

Cochran, W. G. (1970). Some effects of errors of measurement on multiple correlation. *Journal of the American Statistical Association*, *65*, 22 – 34.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ferron, J., Yi, Q. & Kromrey, J.D. (1997). NNCORR: A SAS/IML program for generating nonnormal correlated data. *Applied Psychological Measurement*, *21*, 64.

Fisher, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society*, *121*, 654-673.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521-532.

Fuller, W. A. & Hidirouglou, M.A. (1978). Regression estimation after correcting for attenuation. *Journal of the American Statistical Association*, *73*, 99-104

Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum.

Jaccard, J & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin*, *117*, 348-357.

Jencks, C., and others (1972). *Inequality: A reassessment of the effect of family and schooling in America*. New York: Basic Books.

Kendall, M. G. & Stuart, A. (1977). *The advanced theory of statistics* (4th Ed., Vol. 2). London: Griffin.

Table 10. Width of Confidence Band by Number of Regressors, Reliability and Distribution Shape.

			Distribution Shape				
	Reliability	Method	Sk = 0.0 Kurt = 0.0	Sk = 1.0 Kurt = 3.0	Sk = 1.5 Kurt = 5.0	Sk = 2.0 Kurt = 6.0	Sk = 0.0 Kurt = 25.0
2 Regressors	0.4	O&F 1	0.32	0.32	0.32	0.32	0.32
		O&F 2	0.36	0.36	0.36	0.36	0.36
		O& F 3	0.72	0.73	0.72	0.73	0.73
		S&F	0.37	0.36	0.36	0.36	0.36
	0.8	O&F 1	0.39	0.39	0.39	0.38	0.39
		O&F 2	0.42	0.42	0.41	0.41	0.42
		O& F 3	0.66	0.66	0.66	0.66	0.66
		S&F	0.42	0.42	0.42	0.41	0.42
	1.0	O&F 1	0.41	0.41	0.41	0.40	0.41
		O&F 2	0.42	0.42	0.41	0.40	0.42
		O& F 3	0.62	0.62	0.62	0.62	0.62
		S&F	0.43	0.43	0.43	0.42	0.43
4 Regressors	0.4	O&F 1	0.29	0.29	0.29	0.29	0.28
		O&F 2	0.31	0.31	0.31	0.31	0.30
		O& F 3	0.32	0.32	0.32	0.32	0.31
		S&F	0.27	0.28	0.27	0.27	0.26
	0.8	O&F 1	0.34	0.34	0.34	0.34	0.33
		O&F 2	0.35	0.35	0.34	0.35	0.33
		O& F 3	0.35	0.35	0.34	0.34	0.33
		S&F	0.33	0.33	0.33	0.33	0.31
	1.0	O&F 1	0.35	0.35	0.35	0.35	0.34
		O&F 2	0.34	0.34	0.33	0.33	0.33
		O& F 3	0.34	0.33	0.33	0.33	0.33
		S&F	0.33	0.33	0.33	0.33	0.32
8 Regressors	0.4	O&F 1	0.24	0.24	0.24	0.24	0.24
		O&F 2	0.25	0.25	0.25	0.25	0.25
		O& F 3	0.25	0.25	0.25	0.25	0.25
		S&F	0.20	0.20	0.20	0.20	0.20
	0.8	O&F 1	0.27	0.27	0.27	0.27	0.27
		O&F 2	0.27	0.27	0.27	0.27	0.27
		O& F 3	0.26	0.26	0.26	0.26	0.26
		S&F	0.24	0.24	0.24	0.24	0.24
	1.0	O&F 1	0.27	0.27	0.27	0.27	0.27
		O&F 2	0.26	0.26	0.26	0.26	0.26
		O& F 3	0.26	0.26	0.26	0.25	0.26
		S&F	0.25	0.24	0.24	0.24	0.24

Kromrey, J. & Foster-Johnson, L. (1999, April). *Bias, Type I Error Control and Statistical Power in Multiple Regression: An Empirical Comparison of OLS and Errors-in-Variables Regression Algorithms.* Paper presented at the annual conference of the American Educational Research Association, New Orleans, LA.

Kromrey, J. & Hess, M. (2001, April). *Interval estimates of R^2: An empirical comparison of accuracy and precision under violations of the normality assumption.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Lee, Y. (1971). Some results on the sampling distribution of the multiple correlation coefficient. *Journal of the Royal Statistical Society*, *B*, *33*, 117-130.

Table 11. Width of Confidence Band by Number of Regressors, Measurement Reliability and ρ 2.

			Population Squared Multiple Correlation				
	Reliability	Method	$\rho^2 = 0.01$	$\rho^2 = 0.05$	$\rho^2 = 0.1$	$\rho^2 = 0.3$	$\rho^2 = 0.6$
2 Regressors	0.4	O&F 1	0.27	0.29	0.30	0.34	0.39
		O&F 2	0.32	0.33	0.34	0.38	0.43
		O& F 3	0.78	0.77	0.75	0.69	0.64
		S&F	0.32	0.34	0.35	0.38	0.43
	0.8	O&F 1	0.28	0.32	0.35	0.46	0.53
		O&F 2	0.33	0.36	0.39	0.48	0.51
		O& F 3	0.77	0.72	0.67	0.60	0.53
		S&F	0.33	0.36	0.39	0.48	0.53
	1.0	O&F 1	0.29	0.33	0.39	0.51	0.52
		O&F 2	0.33	0.38	0.42	0.51	0.43
		O& F 3	0.77	0.69	0.64	0.56	0.44
		S&F	0.34	0.38	0.42	0.51	0.49
4 Regressors	0.4	O&F 1	0.24	0.25	0.27	0.31	0.35
		O&F 2	0.27	0.28	0.29	0.33	0.37
		O& F 3	0.29	0.29	0.30	0.33	0.36
		S&F	0.23	0.24	0.25	0.30	0.34
	0.8	O&F 1	0.25	0.28	0.32	0.40	0.44
		O&F 2	0.27	0.31	0.34	0.40	0.39
		O& F 3	0.29	0.31	0.34	0.38	0.38
		S&F	0.23	0.27	0.31	0.39	0.42
	1.0	O&F 1	0.25	0.30	0.33	0.43	0.43
		O&F 2	0.27	0.32	0.35	0.41	0.32
		O& F 3	0.29	0.32	0.35	0.39	0.32
		S&F	0.23	0.28	0.32	0.42	0.38
8 Regressors	0.4	O&F 1	0.20	0.21	0.22	0.26	0.29
		O&F 2	0.22	0.23	0.24	0.27	0.30
		O& F 3	0.23	0.23	0.24	0.26	0.29
		S&F	0.16	0.17	0.18	0.23	0.28
	0.8	O&F 1	0.21	0.23	0.26	0.31	0.34
		O&F 2	0.23	0.25	0.27	0.31	0.28
		O& F 3	0.23	0.25	0.26	0.29	0.28
		S&F	0.16	0.20	0.23	0.31	0.32
	1.0	O&F 1	0.24	0.24	0.27	0.33	0.32
		O&F 2	0.26	0.26	0.28	0.30	0.23
		O& F 3	0.25	0.25	0.27	0.30	0.23
		S&F	0.21	0.21	0.25	0.32	0.27

Maxwell, S. E., Delaney, H. D., & Dill, C. A. (1984). Another look at ANCOVA versus blocking. *Psychological Bulletin*, *95*, 136-147.

Olkin, I. & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin*, *188,* 155-164.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd Ed.). Fort Worth: Harcourt Brace.

Robey, R. R. & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, *45*, 283-288.

Steiger, J. H. & Fouladi, R. T. (1992). R2: A computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation. *Behavior Research, Methods, Instruments, and Computers*, *4*, 581-582.

Table 12. Width of Confidence Band by Number of Regressors, Reliability and Sample Size.

			Sample Size		
	Reliability	Method	N=5*k	N=10*k	N=50*k
2 Regressors	0.4	O&F 1	0.49	0.34	0.13
		O&F 2	0.59	0.37	0.13
		O& F 3	0.91	0.86	0.41
		S&F	0.57	0.38	0.14
	0.8	O&F 1	0.53	0.43	0.21
		O&F 2	0.62	0.43	0.19
		O& F 3	0.86	0.77	0.35
		S&F	0.62	0.44	0.20
	1.0	O&F 1	0.53	0.45	0.23
		O&F 2	0.61	0.44	0.20
		O& F 3	0.81	0.73	0.32
		S&F	0.63	0.46	0.20
4 Regressors	0.4	O&F 1	0.45	0.30	0.11
		O&F 2	0.51	0.31	0.11
		O& F 3	0.50	0.33	0.11
		S&F	0.43	0.28	0.10
	0.8	O&F 1	0.49	0.37	0.16
		O&F 2	0.53	0.36	0.15
		O& F 3	0.51	0.36	0.15
		S&F	0.49	0.34	0.14
	1.0	O&F 1	0.49	0.38	0.18
		O&F 2	0.51	0.35	0.15
		O& F 3	0.50	0.36	0.15
		S&F	0.50	0.34	0.15
8 Regressors	0.4	O&F 1	0.37	0.25	0.09
		O&F 2	0.42	0.26	0.09
		O& F 3	0.40	0.26	0.09
		S&F	0.33	0.21	0.08
	0.8	O&F 1	0.40	0.29	0.12
		O&F 2	0.42	0.28	0.11
		O& F 3	0.40	0.27	0.11
		S&F	0.37	0.25	0.10
	1.0	O&F 1	0.40	0.30	0.13
		O&F 2	0.40	0.27	0.11
		O& F 3	0.39	0.27	0.11
		S&F	0.38	0.25	0.10

Steiger, J. H. & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum, p. 221-257.

Wilkinson & APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, p. 594-604.

Send correspondence to: Jeffrey D Kromrey, Ph.D.
University of South Florida
Email: kromrey@tempest.coedu.usf.edu

Effect Size and Confidence Intervals in General Linear Models for Categorical Data Analysis

Randall E. Schumacker
University of North Texas, Health Science Center

In general linear models for categorical data analysis, goodness-of-fit statistics only provide a broad significance test of whether the model fits the sample data. Hypothesis testing has traditionally reported the chi-square or G^2 likelihood ratio (deviance) statistic and associated p-value when testing the significance of a model or comparing alternative models. The effect size (log odds ratio) and confidence interval (ASE) need to receive more attention when interpreting categorical response data using the logistic regression model. This trend is supported by recent efforts in general linear models for continuous data (t-test, analysis of variance, least squares regression) that have criticized the sole use of statistical significance testing and the $p < .05$ criteria for a Type I error rate..

The American Psychological Association has recently advocated that hypothesis testing go beyond statistical significance testing at $p < .05$ for Type I error rate (Wilkinson, L., & APA Task Force on Statistical Inference, 1999). The traditional statistical significance testing has placed an emphasis upon the probability of a statistical value occurring beyond a chance level given the sampling distribution of the statistic (Harlow, Mulaik, and Steiger, 1997). Recently, more emphasis has been placed on the practical interpretation of results that include effect size, confidence interval, and confidence intervals around an effect size; however, the discussion centered on statistical applications that use continuous data (Kirk, 1996). The present study highlights a typical application that uses the general linear model for categorical data analysis (DeMaris, 1992; Fox, 1997). The logistic regression goodness-of-fit criteria for categorical data analysis will be presented (Klienbaum, 1994). The results go beyond the statistical test of significance and highlight the important role that effect size (odds ratio, log odds ratio, relative risk or probability ratio) and confidence interval (asymptotic standard error; ASE) have in the general linear model for categorical data analysis.

Categorical data analysis techniques are used when subject responses are binary and mutually exclusive. The typical method of analyzing relationships amongst categorical variables is to use the chi-square statistic or phi correlation coefficient (Upton, 1978). The general linear model for categorical response variables however has become more widely used in the behavioral sciences because many research questions involve a categorical dependent variable and one or more categorical independent variables.

Logistic regression is a special case of log-linear regression where both the dependent and independent variables are categorical in nature (Hosmer & Lemeshow, 1989; Klienbaum, 1994). It offers distinct advantages over the chi-square method for analysis of categorical variables. In logit models, natural log odds of the frequencies are computed that allow different models and different model parameters to be compared given the additive nature of the G^2 component for each model. If a non-significant likelihood-ratio chi-square (G^2) value is computed, then a given model fits the observed data.

Goodness-of-fit Criteria

A theoretical logit regression model is generally postulated (null model or base model). A common practice is then to create alternative models where each new model contains parameters of the previous model, plus a hypothesized new parameter. The theoretical model can be tested beginning with a null model and adding parameters, or with a saturated model deleting parameters. Several logit regression models may fit equally well based on various goodness-of-fit criteria that are used to determine whether the model fits the data in the logit regression model. The goodness-of-fit criteria typically reported are:

1. Pearson chi-square
2. Likelihood-ratio chi-square (G^2)
3. Predictive efficacy (R-squared type measure)
4. Deviance ($-2\,[L_M - L_S]$)

Pearson chi-square is calculated as: $\chi^2 = \Sigma\ (O - E)^2\ /\ E$. The chi-square distribution is defined by degrees of freedom, df. The mean of the chi-square distribution is equal to *df* with the standard deviation equal to $\sqrt{2df}$. As the degrees of freedom, df, increases the chi-square sampling distribution goes from a right skewed distribution to a normal distribution.

The likelihood-ratio chi-square (G^2) is based on the ratio of maximum likelihood values, Λ, and expressed in logarithm form as $-2 \log (\Lambda)$. The G^2 statistic can also be expressed as: $G^2 = 2\ \Sigma\ Oij \log (Oij\ /\ Eij)$ where *O* is the observed cell frequency, E is the expected cell frequency, and the *i* and *j* subscripts represent the individual cells in the cross-tabulated table. The log transformation yields an approximate chi-squared sampling distribution with a minimum value of zero and larger values suggesting rejection of the null hypothesis. The *p*-value simply indicates the strength of evidence against the null hypothesis.

Predictive efficacy refers to whether a model generates accurate predictions of group membership on the dependent variable. It is possible to have an excellent fit between the logit model and the data without having predictive efficacy. Recall, if $G^2 = 0$, a saturated model exists which perfectly fits the data, yet predictive efficacy can be far from perfect. The R^2 type measure for logistic regression is not meant as a variance accounted for interpretation, as traditionally noted in least squares regression, because it under estimates the proportion of variance explained in the categorical variables. Instead, the R^2 type measure is an approximation for assessing predictive efficacy ranging from zero (0) [independence model] to one (1) [saturated model].

The deviance value provides a way to examine differences in nested logistic regression models. The G^2 from one model is simply subtracted from the G^2 of the second model. This is similar to testing a full versus restricted model in multiple regression. The deviance value is $-2[L_m - L_s]$ where L represents the respective log-likelihood function of each model with the degrees of freedom equal to the difference in the degrees of freedom of the two models. The deviance is the likelihood ratio statistic (G^2) for comparing model *M* to the saturated model *S*. Since the saturated model has $G^2 = 0$, this reduces to the G^2 statistic for the hypothesized logistic regression model. If G^2 is non-significant, then additional independent categorical predictors in the model are not needed. This type of test is only appropriate for the likelihood-ratio chi-square and not the Pearson chi-square because adding additional independent categorical predictor variables will never result in a poorer fit of the model to the data.

Effect Size and Confidence Interval Criteria

Effect size measures and the asymptotic standard error (ASE) play a major role in interpreting the practical significance of estimated parameters in general linear models for categorical variables. The parameter estimates in logistic regression are calculated using maximum likelihood estimation and possess asymptotic properties. As sample size increases, the parameter estimates become unbiased and consistent with population parameters. The sampling distribution also approaches normality with variance lower than other unbiased estimation procedures.

The effect size measures typically used in categorical data analysis are:

1. *z* test
2. odds ratio
3. log odds ratio
4. relative risk or probability ratio

The *z* test, given larger samples, can be used to test a parameter's significance and compute a confidence interval. The formula for z is: $z = B\ /\ \text{ASE}$. The confidence interval is computed as: z +/- 1.96*σ; where $\sigma = [p(1\text{-}p)/n]^{1/2}$. The significance test simply indicates whether an estimated parameter is reasonable whereas the confidence interval yields a range of possible values for the parameter, given sampling error.

Odds ratios are computed as: Odds = $p\ /\ 1 - p$. If the probability of success is .8, the probability of failure is .2, and the odds ratio is .8 / .2 = 4. This indicates 4 successes for every one failure. Unfortunately, odds ratios in small to moderate samples have skewed sampling distributions and therefore are not widely used.

The log odds ratio or natural logarithm of the odds ratio, log (θ), is preferred for interpreting an effect size. Independence of categorical variables is equivalent to log (θ) = 0, i.e. odds ratio = 1 is equal to log odds ratio = 0. The sampling distribution of the log odds ratio approximates a normal distribution as sample size increases with a mean of log (θ) and standard deviation ASE. Parameter estimates in logit models can be readily interpreted as a log-odds ratio. This is calculated as e^{β} for a single parameter, or $e^{\beta_1 - \beta_2}$ for differences between two parameters. This is useful when examining contrasts between levels of two independent categorical predictor variables.

The relative risk or probability ratio should be interpreted separately from the odds ratio (Cohen, 2000). The relative risk (RR) indicates a probability and is computed as probability p_1 divided by probability p_2 [RR = p_1 / p_2]. In contrast the odds ratio (OR) is (p_1 / 1 – p_1) divided by (p_2 / 1 – p_2). The odds ratio is therefore related, but different from relative risk (OR = [PR–p_1]/1–p_1] or RR x [1–p_2/1–p_1]). For logistic model interpretation, a gender coefficient (male = 0 and female = 1) of $e^{1.67}$ would indicate the odds of females over males participating, whereas the statement females were two-thirds more likely than males to participate is a relative risk or probability statement.

The asymptotic standard error (ASE) or standard deviation of the log transform sampling distribution is computed as: ASE (log π) = $[1/n_1 + \ldots + 1/n_k]^{1/2}$. A 95% confidence interval around the log odds ratio is then computed as log(π) +/- 1.96 ASE[log(π)]. The confidence interval should contain the value 1.0 otherwise the true odds will be different for the two groups being compared. The confidence interval also provides valuable information about the range of minimum and maximum log odd ratios.

Method

The logistic regression model ($\log(\pi) = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k$) was applied to a set of categorical data (Stokes, Davis, & Koch, 1995). The goodness-of-fit criteria, effect size, confidence interval, and confidence interval around the effect size are reported. The importance of effect size and confidence interval reporting above and beyond significance testing is then discussed.

Data Analysis

An example data set relating myocardial infarction and aspirin use is provided as follows (Agresti, 1996):

Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11.037

The proportion, p_1, or placebo odds ratio is 189 / 11,034 = .0171 and indicates that .0171 percent suffered myocardial infarction while taking a placebo. In contrast, proportion, p_2, or aspirin odds ratio is 104 / 11,037 and indicates that .0094 suffered myocardial infarction while taking aspirin. The percent difference is .0077 with standard error of .0015. z = .0077/.0015 = 5.133, which is statistically significant. The 95% confidence interval for this true difference is .0077 + /- 1.96(.0015) or (.005, .011), so taking aspirin appears to diminish the risk of myocardial infarction. The relative risk is .0171 divided by .0094 or 1.82. Using relative risk, the proportion of MI cases was 82% higher for the group taking the placebo. The 95% confidence interval is (1.43, 2.30), thus we can be 95% confident that the proportion of MI cases for the group taking the placebo was at least 43% higher than the group taking the aspirin. The relative risk indicates that the difference isn't trivial and may have important health implications.

The natural log odds ratio is log (1.82) = .599. The ASE (log π) is computed as $[1/189 + 1/10{,}845 + 1/104 + 1/10{,}933]^{1/2}$ = .123. The 95% confidence interval for log (π) is (.358, .840). The corresponding confidence interval for π is (1.43, 2.30). Since it does not contain 1.0, the true odds of myocardial infarction appear to be different for the two groups.

Results and Interpretation

The categorical data example indicates a statistically significant z-test of the difference between the proportion of myocardial infarction cases for the placebo and aspirin usage groups. The effect size (odds ratio, log odds ratio, and relative risk or probability ratio) provides a more practical interpretation of the efficacy of using aspirin to thwart myocardial infarction in patients. Moreover, the confidence interval and especially the confidence interval around the effect size (log odds ratio) provided important additional information to our interpretation of results.

Statistical significance testing has come under attack by scholars in recent years because it is influenced by a researcher's choice of sample size, power, and Type I error rate. The reported research literature however has focused on continuous data analysis techniques and not fully included categorical data analysis methods. The American Psychological Association and Editors of several popular journals are now requiring educational researchers to report effect size and confidence intervals. The use and interpretation of effect size and confidence interval in categorical data analysis is therefore also important to understand and report.

References

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. NY: John Wiley & Sons, Inc.

Cohen, M.P. (2000). Note on the Odds Ratio and the Probability Ratio. *Journal of Educational and Behavioral Statistics*, *25*(2), 249-252.

DeMaris, A. (1992). *Logit modeling: Practical Applications*. Sage University Paper series on Quantitative Applications in the Social Sciences, no. 07-086. Newbury Park, CA: Sage.

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. Newbury Park, CA: Sage.

Harlow, L.L., Mulaik, S.A., & Steiger, J.H. (editors.) (1997). *What if there were no significance tests?* NJ: Lawrence Erlbaum Associates, Inc.

Hosmer, D.W. & S. Lemeshow (1989). *Applied Logistic Regression*. NY: John Wiley & Sons, Inc.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759.

Kleinbaum, D.G. (1994). *Logistic Regression*. NY: Springer-Verlag.

Stokes, M.E., C.S. Davis, & G.G. Koch (1995). *Categorical Data Analysis Using the SAS System*. Cary, NC: SAS Institute, Inc.

Send correspondence to: Randall E. Schumacker, Ph.D.
University of North Texas
Health Science Center
Email: rschumacker@unt.edu

Multiple Linear Regression Viewpoints

Information for Contributors

Multiple Linear Regression Viewpoints (*MLRV*) is a publication sponsored by the American Educational Research Association's Special Interest Group on Multiple Linear Regression: The General Linear Model (MLR/GLM SIG). It is published twice a year to facilitate communication among professionals who focus their research on the theory, application, or teaching of multiple linear regression models and/or the general linear model. Manuscripts submitted to *MLRV* should conform to the language, style, and format of the Publication Manual of the American Psychological Association (5th ed., 2001). Three copies (two blind) of a doubled spaced manuscript (including equations, footnotes, quotes, and references) of approximately 25 pages in length, a 100 word abstract, and an IBM formatted diskette with the manuscript formatted in WordPerfect or Word should be submitted to one of the editors listed below.

Mathematical and Greek symbols should be clear and concise. All figures and diagrams must be photocopy-ready for publication. Manuscripts will be anonymously peer reviewed by two editorial board members. Author identifying information should appear on the title page of only one submitted manuscript. The review process will take approximately 3 months. A letter acknowledging receipt of the manuscript will be sent to the first author, and upon review, a letter indicating the peer review decision will be sent to the first author. Potential authors are encouraged to contact the editors to discuss ideas for contributions or determine if their manuscript is suitable for publication in *MLRV*.

EDITORS

T. Mark Beasley, Editor *MLRV*
Department of Biostatistics
School of Public Health
343C Ryals Public Health Bldg.
University of Alabama at Birmingham
Birmingham, AL 35294
(205) 975-4957 (voice)
(205) 975-2540 (fax)
mbeasley@uab.edu

Robin K. Henson, Associate Editor
Department of Technology and Cognition
College of Education,
P.O. Box 311337
University of North Texas
Denton, Texas 76203-1337
(940) 369-8385 (voice)
(940) 565-2185 (fax)
rhenson@tac.coe.unt.edu

ORDER INFORMATION
Jeffrey B. Hecht, MLR/GLM SIG Executive Secretary
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL   60115-2854
jbhecht@niu.edu

Check out our website at: http://www.coe.unt.edu/mlrv/

POSTMASTER: Send address changes to:
Jeffrey B. Hecht, MLR/GLM SIG Executive Secretary
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL   60115-2854

Multiple Linear Regression Viewpoints (ISSN 0195-7171) is published by the
AERA Special Interest Group on Multiple Linear Regression: General Linear Model
through the **University of Alabama at Birmingham** and the **Dallas Independent School District**.