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 37 • Number 2 • Fall 2011

Table of Contents

Multiple Linear Regression Viewpoints

David A. Walker, Editor
Northern Illinois University

T. Mark Beasley, Associate Editor
University of Alabama-Birmingham

Isadore Newman, Editor Emeritus
Florida International University

Randall E. Schumacker, Editor Emeritus
University of Alabama-Tuscaloosa

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 **Northern Illinois University** and the **University of Alabama-Birmingham**.

Subscription and SIG membership information can be obtained from:
Cynthia Campbell, Managing Editor
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL 60115-2854
ccampbell@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*.

Multiple Linear Regression Viewpoints

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 (SIG/MLR: GLM). 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* (6th ed., 2010). Manuscripts should be prepared in Word, be doubled-spaced, use 12 font, contain a 100 word abstract, have author(s) identifying information appear on the title page only, and consist of no more than 30 pages in length (including equations, footnotes, quotes, and references). Mathematical and Greek symbols should be clear and concise. Tables, figures, and diagrams must be photo copy ready for publication. All manuscripts should be submitted electronically to the editor.

Once received by the editor, manuscripts will be anonymously peer-reviewed by two editorial board members. The review process will take approximately 2 to 3 months. A letter acknowledging receipt of the manuscript will be sent to the first author, and upon review completion, a letter indicating the peer-review decision will be sent to the first author. Potential authors are encouraged to contact the editor to discuss ideas for contributions or determine if their manuscript is suitable for publication in *MLRV*.

EDITOR

David Walker
Northern Illinois University
Department of Educational Technology, Research and Assessment
204A Gabel Hall
DeKalb, IL 60115
Phone: (815) 753-7886
Fax: (815) 753-9388
Email: dawalker@niu.edu

ASSOCIATE EDITOR

T. Mark Beasley
University of Alabama-Birmingham
Department of Biostatistics
School of Public Health
Ryals Public Health Bldg.
Birmingham, AL 35294
Phone: (205) 975-4957
Fax: (205) 975-2540
Email: MBeasley@ms.soph.uab.edu

ORDER INFORMATION

Cynthia Campbell, Managing Editor
Northern Illinois University
Department of Educational Technology, Research and Assessment
DeKalb, IL 60115
Phone: (815) 753-8471
Fax: (815) 753-9388
Email: ccampbell@niu.edu

Model Selection with Information Complexity in Multiple Linear Regression Modeling

Hongwei Yang
University of Kentucky

Hamparsum Bozdogan
University of Tennessee

This paper aims to introduce to applied researchers a new family of information model selection criteria in multiple linear regression models. These criteria are known as information complexity (*ICOMP*) criteria. The paper provides supportive evidence under the R language to show the effectiveness of *ICOMP* and its tendency to outperform some other traditional criteria: *AIC*, *SBC*, etc. This paper also creates a framework on which to base future work in applying *ICOMP* to more general regression modeling problems in R.

The selection of an appropriate model from a potentially large class of candidate models is an issue that is central to regression, time series modeling, and generalized linear models (McQuarrie & Tsai, 1998). In multiple linear regression, statistical model evaluation and selection involves evaluating a pool of subsets of predictors and selecting the best subset that predicts the response with sufficient accuracy from predictor variables that can be measured cheaply (Miller, 2002). Given a large number of predictor variables, the hope is to identify a small subset of them that gives adequate prediction accuracy for a reasonable cost of measurement. On the other hand, it is well known that, for multiple linear regression models fitted using least squares, the variance of the predicted response values increases monotonically with the number of predictor variables used in the prediction equation, and this increased prediction variability is traded off against reduced prediction bias. The question of how this trade-off should be handled is a critical problem in this field of subset selection in multiple linear regression modeling.

The problem of selecting the best regression subset is not trivial particularly when there are a large number of potential predictors. This is so because, usually without a precise knowledge of the relationship between the response and the predictors, researchers have to find a way of developing, validating, evaluating and selecting regression models and the increase in the number of predictors complicates the process. In addition to theoretical considerations, researchers also rely on data-adaptive approaches to regression model selection. Hypothesis-test-based stepwise regression is one of many data-adaptive model selection techniques that are commonly used today, which adds and/or removes predictors based on partial *F* or *t* statistics with arbitrarily set probabilities of entry and removal after controlling the contributions of other predictors, if any, already in the model. However, hypothesis-test-based stepwise regression has known problems. First, there is no guarantee that the final model from stepwise regression is optimal in any specified sense (Tamhane & Dunlop, 1999). Stepwise procedures can sometimes err by identifying a suboptimal regression model as "best" (Kutner, Nachtsheim, & Neter, 2004). Second, the probabilities for entry and removal of predictors are arbitrarily set, so plenty of subjectivity exists in the model search process.

As an alternative to model selection via hypothesis testing, information model selection criteria are recommended for comparing and evaluating competing regression and other statistical models (Burnham & Anderson, 2002). As is compared with the usual methods of hypothesis testing, the use of information criteria in model selection has had a much shorter exposure in statistics. Information criteria belong to the group of relative fit criteria which select the best model from a pool of models that we have specified. Relying on information criteria, we can identify the model that appears to be the best among its competitors (Skrondal & Rabe-Hesketh, 2004), and the model is the best in the sense of optimizing information criteria. So, a critical task for users of information criteria is to set up more appropriate competing models by making use of knowledge regarding the object (Konishi & Kitagawa, 2010). Information criteria can be used with many data-adaptive automatic model selection algorithms including stepwise regression, all-possible-subset regression, and genetic algorithms (Bozdogan, 2004).

There are two approaches to information model selection criteria: 1) Information- theoretic approach, and 2) Bayesian approach (Ando, 2010; Konishi & Kitagawa, 2010). The former approach includes Akaike's Information Criterion or *AIC* (Akaike, 1973; 1987), Consistent Akaike's Information Criteria or *CAIC* (Bozdogan, 1987), etc. The latter approach includes Schwartz Bayesian Criterion or *SBC* (Schwartz, 1978), etc. The *AIC*-type criteria and their variants are constructed as estimators of the Kullback-Leibler

(K-L) information (Kullback & Leibler, 1951) between a statistic's model and the true distribution generating the data. In contrast, the Bayes approach for selecting a model is to choose the model with the largest posterior probability among a set of competing models. Information criteria usually assess how badly a model fits the data while adjusting for the level of complexity of a model (i.e., the number of free parameters, interdependency of parameter estimates, etc.) (Bozdogan, 2004), so the best approximating model is selected as the one that minimizes the criterion. Due to the availability of multiple criteria, matching appropriate selection criteria to a given problem or data set has received much attention in the literature (McQuarrie & Tsai, 1998).

Many information criteria appear similar in form to *AIC* because they all take the form of 1) a penalized log likelihood: a badness/lack of fit term, or a negative log likelihood term, plus 2) a penalty term (Sclove, 1987). For example, the formula for *AIC* is (-2) times the maximized log likelihood function plus 2 times the number of free parameters, with the former term describing lack of fit and the latter penalizing the number of free parameters in the model. In *AIC*, a measure of model complexity is comprised of the number of free parameters (Bozdogan, 2004). Like *AIC*, many other information criteria also contain two terms that serve similar purposes. They usually use the same lack of fit term as *AIC*, but differ in how to penalize model complexity.

Bozdogan's Information Complexity Criterion or *ICOMP* is a relatively new family of model selection criterion (Bozdogan, 2004). Like *AIC* and other criteria, *ICOMP* uses (-2) times the maximized log likelihood to measure the lack of fit of the model. On the other hand, the complexity of the model is measured based on a generalization of the covariance complexity index introduced by Van Emden (1971). Unlike *AIC*, which defines model complexity as number of free parameters, *ICOMP* measures this concept with both the number of free model parameters and the interdependency of parameter estimates. According to Bozdogan (2004), Konishi and Kitagawa (2010), and Mulaik (2009), a generic formula of *ICOMP* is:

$$ICOMP = -2logL(\hat{\boldsymbol{\theta}}) + 2C(\hat{\Sigma}_M),$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of the parameter vector under the model whose covariance matrix is denoted by $\hat{\Sigma}_M = Est.Cov(\hat{\boldsymbol{\theta}})$, and where C represents a real-valued complexity measure of $\hat{\Sigma}_M$. Usually two types of C measures exist denoted by $C_1(*)$ and $C_{1F}(*)$, respectively. Both of them are designed to *transform* a covariance matrix into a scalar value, which is then used to measure model complexity. The covariance matrix inside the parenthesis of the two complexity measures is called the inverse Fisher Information Matrix (*IFIM*). Bozdogan (2004) developed several *IFIM*s to handle different modeling conditions (e.g., mis-specification resistant vs. otherwise). Loosely speaking, when applying a complexity measure (either by $C_1(*)$ or $C_{1F}(*)$) to *IFIM*, the model complexity part of *ICOMP* is created, which is combined with the lack of fit part to construct an *ICOMP* criterion.

Although the use of *AIC*, *CAIC*, and *SBC* in regression analysis is well documented in the literature (Burnham & Anderson, 2002; Claeskens & Hjort, 2008; McQuarrie & Tsai, 1998; Miller 2002) partially because they have been made readily available by major statistics programs, the research on applying *ICOMP* to regression modeling is very limited. Bozdogan and Haughton (1998) examined the performance of six *ICOMP* criteria using only the $C_1(*)$ measure of complexity in its early stage of development. Since then, more *ICOMP* criteria have been created that have extended the way model complexity is measured. So, this paper revisits the topic of *ICOMP*-based regression model selection using more recent *ICOMP* criteria that approach model complexity from beyond the $C_1(*)$ perspective to include the $C_{1F}(*)$ measure. Also, prior implementations of *ICOMP* have used MATLAB®, a program preferred mainly by engineers/mathematicians. Coding *ICOMP* in R is desired because R is more readily available and is better accepted in non-engineering/non-math fields

In sum, this study aims to achieve the following: 1) familiarizing applied researchers using regression with *ICOMP*, 2) comparing the performance of *ICOMP* in regression with that of other criteria, and 3) creating *ICOMP* routines in R (available upon request from the authors) to present the criteria in a better accepted environment.

Before continuing, some key general issues in model selection are briefly discussed:

Best approximating model: This is the model in the pool of candidate models that is "closet" to the true model (Bozdogan & Haughton, 1998). The objective of modeling is to obtain a "good" model, rather than the true model (Konishi & Kitagawa, 2010). This true model, which in the background generated the data, might be very complex and almost always unknown. For working with the data, it may be more practical to work instead with a simpler, but almost-as-good model, and, hence, the best approximating model. A true model can be defined explicitly only in some special situations such as in computer simulations. In this paper, the *good* model and the *best* model are both used to refer to the best approximating model.

Consistency: A model selection criterion is considered to be consistent if the probability of selecting the best approximating model converges to one as the sample size goes to infinity. Because an infinitely large sample is impossible to obtain, the paper focuses on the behavior of *ICOMP* criteria as the sample size is finite and keeps increasing. If the performance of *ICOMP* improves as sample size increases, it provides supportive evidence of *ICOMP* being consistent.

Overfitting and underfitting: Statistical modeling has to balance simplicity (i.e., fewer parameters in a model, lower variability in the predicted response, but with more modeling bias) against complexity (i.e., more parameters in a model, higher variability in the predicted response, but with smaller modeling bias). Statistical model selection criteria have to seek a proper balance between overfitting (i.e., a model with too many parameters, more than actually needed) and underfitting (i.e., a model with too few parameters, not capturing the right signal) (Claeskens & Hjort, 2008*)*. A criterion underfits/overfits a model when it selects a model that contains fewer/more parameters than does the best approximating model (Bozdogan & Haughton, 1998).

Theoretical Framework

A multiple linear regression model under normality is defined by:

$$\underset{(n\text{x}1)}{\mathbf{y}} = \underset{(n\text{x}q)(q\text{x}1)}{\mathbf{X}\,\boldsymbol{\beta}} + \underset{(n\text{x}1)}{\boldsymbol{\varepsilon}} \tag{1}$$

where $\mathbf{y}$ is an (nx1) vector of observed values of the response variable, $\mathbf{X}$ is an (nxq) full rank matrix representing n observations with each one measured on k variables and $q = k + 1$, $\boldsymbol{\beta}$ is a (qx1) matrix of unknown regression coefficients, and $\boldsymbol{\varepsilon}$ is an (nx1) vector of i.i.d. random errors. Further, suppose $\mathbf{y} \sim \boldsymbol{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$ with σ^2 being the unknown variance of random errors.

To evaluate how well an estimated regression model under Equation (1) fits the observed data, *ICOMP* criteria are presented below. *ICOMP* criteria share the same badness/lack of fit term as *AIC*, *CAIC*, etc., which equals (-2) times the maximized log likelihood function, but *ICOMP* criteria measure model complexity differently.

Badness/Lack of Fit Term of ICOMP

Given the multiple regression model in Equation (1) the maximum likelihood estimates or MLE's of $\boldsymbol{\beta}$ and σ^2 are given by:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}\,, \tag{2}$$

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'\,(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \tag{3}$$

Hence, the maximized log likelihood function is

$$logL(\widehat{\boldsymbol{\beta}}, \hat{\sigma}^2) = -\frac{1}{2}\,n\,log(2\pi) - \frac{n}{2}\,log(\hat{\sigma}^2) - \frac{1}{2}\,n \tag{4}$$

The badness/lack of fit part of *ICOMP* is thus:

$$-2logL(\widehat{\boldsymbol{\beta}}, \hat{\sigma}^2) = n\,log(2\pi) + n\,log(\hat{\sigma}^2) + n \tag{5}$$

Model Complexity Term of ICOMP

The model complexity term of *ICOMP* takes various forms, so various versions of *ICOMP* can be defined. Basically, this term is defined as the complexity of inverse the Fisher Information Matrix or *IFIM* (Bozdogan, 2004). There are two ways to measure the complexity of a matrix, namely $C_1(*)$ and

$C_{1F}(*)$. There are also two different forms of *IFIM*, namely *IFIM* and mis-specified *IFIM*. Presented next are three approaches to model complexity in *ICOMP* with different combinations of 1) complexity measure ($C_1(*)$ vs. $C_{1F}(*)$) and 2) *IFIM* (*IFIM* vs. mis-specified *IFIM*).

The first approach to *ICOMP* complexity takes the $C_1(*)$ complexity of $\mathbf{F}^{-1}$, denoted by $C_1(\mathbf{F}^{-1})$, where $\mathbf{F}^{-1}$ is the estimated inverse Fisher Information Matrix of the regression model given by

$$\mathbf{F}^{-1} = Est.Cov(\widehat{\boldsymbol{\beta}},\sigma^2) = \begin{bmatrix} \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0}' & \frac{2\hat{\sigma}^4}{n} \end{bmatrix}$$

Now invoking the complexity measure the $C_1$to $\mathbf{F}^{-1}$we have the scalar value of its complexity given by:

$$C_1(\mathbf{F}^{-1}) = \frac{s}{2} log\left[\frac{tr(\mathbf{F}^{-1})}{s}\right] - \frac{1}{2} log\left|\mathbf{F}^{-1}\right|, \quad (6)$$

where $$s = dim(\mathbf{F}^{-1}) = rank(\mathbf{F}^{-1}) \quad (7)$$

For the regression model in Equation (1), $s = dim(\mathbf{F}^{-1}) = rank(\mathbf{F}^{-1}) = q$. Further suppose the eigenvalues of $Est.Cov(\widehat{\boldsymbol{\beta}},\sigma^2)$ are $\lambda_1, \lambda_2, \ldots, \lambda_q$. Therefore,

$$\begin{aligned} C_1(\mathbf{F}^{-1}) &= \frac{q}{2} log\left[\frac{tr(\mathbf{F}^{-1})}{q}\right] - \frac{1}{2} log\left|\mathbf{F}^{-1}\right| \\ &= \frac{q}{2} log\left[\frac{\sum_{j=1}^{q}\lambda_j}{q}\right] - \frac{1}{2} log\left|\prod_{j=1}^{q}\lambda_j\right| \\ &= \frac{q}{2} log\left[\frac{\bar{\lambda}_a}{\bar{\lambda}_g}\right] \end{aligned}$$

where $\bar{\lambda}_a = \frac{1}{q}\sum_{j=1}^{q}\lambda_j$ is the arithmetic mean of the eigenvalues of $\mathbf{F}^{-1}$ and $\bar{\lambda}_g = \left[\prod_{j=1}^{q}\lambda_j\right]^{\frac{1}{q}}$ is the corresponding geometric mean.

The second approach to *ICOMP* complexity takes the $C_{1F}(*)$ complexity of $\mathbf{F}^{-1}$ denoted by $C_{1F}(\mathbf{F}^{-1})$. This second complexity measure is used to avoid the problematic situation where $C_1(\mathbf{F}^{-1})$ becomes zero; it measures the relative variation in the eigenvalues and is given by:

$$C_{1F}(\mathbf{F}^{-1}) = \frac{1}{4\bar{\lambda}_a^2}\sum_{j=1}^{q}(\lambda_j - \bar{\lambda}_a)^2 \ . \quad (8)$$

The third approach to *ICOMP* complexity uses both $\mathbf{F}^{-1}$ and its outer product form $\mathbf{R}$. For the regression model in Equation (1), the estimated outer product form of the Fisher Information Matrix is given by:

$$\mathbf{R} = \begin{bmatrix} \frac{1}{n\hat{\sigma}^4}\mathbf{X}'\mathbf{D}^2\mathbf{X} & \mathbf{X}'\mathbf{1}\frac{Sk}{2\hat{\sigma}^3} \\ (\mathbf{X}'\mathbf{1}\frac{Sk}{2\hat{\sigma}^3})' & \frac{(Kt-1)}{4\hat{\sigma}^4} \end{bmatrix}, \quad (9)$$

where $\mathbf{D}^2 = \text{diag}[\hat{\epsilon}_1^2, \hat{\epsilon}_2^2, , \ldots, \hat{\epsilon}_n^2]$ with , $i = 1,2, \ldots, n$, being squared residuals from the fitted regression model, *Sk* is the estimated residual skewness, *Kt* the estimated residual kurtosis, and **1** is an (nx1) vector of ones. Formulas for *Sk* and *Kt* are respectively given by:

$$Sk = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^3}{\hat{\sigma}^3} \quad , \quad \text{and} \quad Kt = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^4}{\hat{\sigma}^4}$$

With $\mathbf{F}^{-1}$ and $\mathbf{R}$, the mis-specified version of the estimated *IFIM* can be defined:

$$Est.Cov(\widehat{\boldsymbol{\beta}},\sigma^2)_{Mis} = \mathbf{F}^{-1}\mathbf{R}\mathbf{F}^{-1}$$

Therefore, the third approach to *ICOMP* complexity takes the $C_1(*)$ complexity of $\mathbf{F}^{-1}\mathbf{R}\mathbf{F}^{-1}$ denoted by $C_1(\mathbf{F}^{-1}\mathbf{R}\mathbf{F}^{-1})$. This version of *ICOMP* provides a protection against model mis-specification (Bozdogan, 2004).

ICOMP and Non-ICOMP Criteria

Based on the information presented previously, formulas for several *ICOMP* criteria are given below, along with formulas for several non-*ICOMP* criteria.

$$AIC = nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2(k+1) \quad (10)$$

$$AIC_C = nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2\left[\frac{n(k+1)}{n-k-2}\right] \quad (11)$$

$$CAIC = nlog(2\pi) + nlog(\hat{\sigma}^2) + n + [log(n)+1]k \quad (12)$$

$$SBC = nlog(2\pi) + nlog(\hat{\sigma}^2) + n + [log(n)]k \quad (13)$$

$$\begin{aligned} ICOMP_{C1} &= nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2C_1(\mathbf{F}^{-1}) \\ &= nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2\left[\frac{q}{2}log\left(\frac{\bar{\lambda}_a}{\bar{\lambda}_g}\right)\right] \end{aligned} \quad (14)$$

$$\begin{aligned} ICOMP_{C1F} &= nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2C_{1F}(\mathbf{F}^{-1}) \\ &= nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2\left[\frac{1}{4\bar{\lambda}_a^2}\sum_{j=1}^{q}(\lambda_j - \bar{\lambda}_a)^2\right] \end{aligned} \quad (15)$$

Finally, according to the mis-specified *IFIM* or $Est.Cov(\hat{\boldsymbol{\beta}},\sigma^2)_{Mis}$, the mis-specified *ICOMP* can be defined by:

$$\begin{aligned} ICOMP_{Mis} &= nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2C_1\left[Est.Cov(\hat{\boldsymbol{\beta}},\hat{\sigma}^2)_{Mis}\right] \\ &= nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2C_1[\mathbf{F}^{-1}\mathbf{R}\mathbf{F}^{-1}] \end{aligned} \quad (16)$$

Further analyses are based on the seven criteria presented above. Data sources and the simulation protocol are detailed in the next section.

Monte Carlo Simulation Examples

Simulation Protocol

Determining the effectiveness of an information criterion involves evaluating cumulative model selection results from repeated random sampling: running the simulation repeatedly and finding the number of times that the best approximating model is identified by each criterion. Data sets used in the study are generated using Monte Carlo methods (Bozdogan & Haughton, 1998). The study simulates data sets where the true regression model has five predictors, namely $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$, and $\mathbf{x}_5$. And the analysis is performed respectively for three sample sizes, namely n = 50, 100, and 1000.

Suppose $z_i \sim N(0,1)$, i = 1, 2, . . ., 6. The following simulation protocol is used:

$$\mathbf{x}_i = \sqrt{1-\alpha_1^2}z_i + \alpha_1 z_6 \text{ when } i = 1, 2, 3$$
$$\mathbf{x}_i = \sqrt{1-\alpha_2^2}z_i + \alpha_2 z_6 \text{ when } i = 4, 5.$$

α_1 and α_2 are parameters controlling the degree of multicollinearity, and $\alpha_1^2 = 0.3$ and $\alpha_2^2 = 0.5$ to yield a reasonable covariance structure for $\mathbf{X} = \{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4,\mathbf{x}_5\}$. Given $\mathbf{X}$ already generated using the above protocol, the focus is now on obtaining $\boldsymbol{\beta}$. Here, $\boldsymbol{\beta}$ is generated from the eigenvectors of $(\mathbf{X}'\mathbf{X})$. Three $\boldsymbol{\beta}$ vectors are obtained from $(\mathbf{X}'\mathbf{X})$ and used to produce three sets of $(\mathbf{X}\boldsymbol{\beta})$ values having different degrees of variability, namely $\boldsymbol{\beta}_{max}$, $\boldsymbol{\beta}_{min}$, and $\boldsymbol{\beta}_{int}$. The eigenvector corresponding to the largest eigenvalue of $(\mathbf{X}'\mathbf{X})$ is denoted as $\boldsymbol{\beta}_{max}$, that corresponding to the smallest eigenvalue as $\boldsymbol{\beta}_{min}$, and that equal to $\frac{1}{2}(\boldsymbol{\beta}_{max}+\boldsymbol{\beta}_{min})$ as $\boldsymbol{\beta}_{int}$. So, according to Johnson and Wichern (1992), $(\mathbf{X}\boldsymbol{\beta}_{max})$ possesses the largest variability, $(\mathbf{X}\boldsymbol{\beta}_{min})$ the smallest variability, and $(\mathbf{X}\boldsymbol{\beta}_{int})$ the intermediate variability. Given $\mathbf{X}$ and $\boldsymbol{\beta}$, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Here, $\boldsymbol{\varepsilon}$ is simulated from a normal distribution with a mean of 0 and a user-specified variance, σ^2.

Two Modeling Conditions

Given $\mathbf{X}$ and $\mathbf{y}$, the performance of information criteria is examined under two conditions. One condition has the true model included in the pool of candidate models, whereas the other one does not. The good

model is to be identified in both conditions. When the true model is in the pool, the good model is just the true model. Otherwise, the good model is the one that is "closest" to the true model.

When the True Model is Included

This part of the analysis assesses the number of times that *ICOMP* criteria successfully identify the true model, which *ICOMP* criteria overfit a model, and that *ICOMP* criteria underfit a model. To add more competing models to the pool, two additional variables $\mathbf{x}_6$ and $\mathbf{x}_7$ are added to $\mathbf{X}$ with both of them generated from an exponential distribution Exp (0.1). A total of seven models are evaluated and compared using information criteria, namely $\{\mathbf{x}_1\}$, $\{\mathbf{x}_1,\mathbf{x}_2\}$, …, and $\{\mathbf{x}_1,\mathbf{x}_2, ..., \mathbf{x}_K\}$, $K = 3, 4, \ldots, 7$. The true model is the one with five predictors: $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4,\mathbf{x}_5\}$.

When the True Model is Not Included

This part of the analysis assesses the number of times that *ICOMP* criteria select the good model minimizing the K-L distance between the true model and each estimated model. Here, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, and $\mathbf{x}_4$ are used to create the pool of candidate models. A total of four models are created, evaluated, and compared using information criteria, namely $\{\mathbf{x}_1\}$, $\{\mathbf{x}_1,\mathbf{x}_2\}$, $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3\}$, and $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4\}$. The true model is still the one with five predictors: $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4,\mathbf{x}_5\}$, although it is not in the pool of competing models. The model in the pool that minimizes the K-L distance from the true model is the one with four predictors: $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4\}$. Hereafter, Models 1 through 7 refer to the regression models with 1 through 7 predictor variables, respectively. For example, Model 3 is the regression model that contains just three predictors $\mathbf{x}_1$ through $\mathbf{x}_3$, or $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3\}$.

Simulation Results

With the True Model Included

Tables 1, 2, and 3 present the model selection results from the case when the true model is included, with Table 1 corresponding to $\boldsymbol{\beta}_{max}$, Table 2 to $\boldsymbol{\beta}_{int}$, and Table 3 to $\boldsymbol{\beta}_{min}$. In each table, seven model selection criteria are scored to evaluate seven regression models: Models 1 to 7 described above under three sample sizes (i.e., small, medium, and large): $n_{min} = 50$, $n_{int} = 100$, and $n_{max} = 1000$. Since it is Model 5 that simulates the data, the goal of using model selection criteria is to identify this model as the best model.

Under each $\boldsymbol{\beta}$ by n combination, two sets of simulations are run. In the first set of simulations, a total of 100 runs are performed, whereas in the second set, as many as 10,000 runs are performed. So, cells in each table contain two integers separated by a forward slash sign which are frequencies of each competing model being selected under the two sets of simulations (100 runs/10,000 runs), respectively. Model selection results from the two sets of simulations are compared with each other in a few aspects: frequency and/or percentage of identifying the best approximating model, etc. Conclusions are drawn from the patterns found from both sets of simulations. Given any inconsistency in results between the two sets of simulations, those from the second set with a larger number of simulations prevail, because they explore a larger model space.

In addition to model selection frequencies in each of the tables, Figures 1 and 2 present the average percentage of the true model (Model 5) selection as a function of sample size and variability in ($\mathbf{X}\boldsymbol{\beta}$), respectively. Finally, Figure 3 compares all seven criteria in terms of the range of percentages of each of Models 1 through 7 being selected.

The model selection results are examined in the following three aspects:

(1) The increase in sample size tends to improve the performance of all seven criteria in identifying the true model, or Model 5, and this supports the consistency property of all seven criteria. This trend is indicated relatively clearly in all seven line graphs in Figure 1, particularly when the number of runs is larger. In that figure, when the number of runs is 10,000, with an increase in sample size (from 50 to 100, again to 1,000), each line graph keeps showing an upward trend, which indicates that the average percentage of successfully identifying the true model is increasing. When the number of runs is only 100, five of the seven information criteria present an upward trend with an increase in sample size. Two of them, AIC_C and $ICOMP_{C1F}$,

Table 1. Frequency of Model Selection Given Maximum Variability with True Model (100/10,000 runs)

Criterion	*n*	1	2	3	4	5*	6	7
AIC	50	0/0	0/0	0/6	2/143	72/7179	15/1506	11/1166
	100	0/0	0/0	0/0	0/0	78/7582	12/1467	10/951
	1000	0/0	0/0	0/0	0/0	73/7822	15/1337	12/841
*AIC*c	50	0/0	0/0	0/11	2/210	79/8112	12/1085	7/582
	100	0/0	0/0	0/0	0/1	84/8052	9/1251	7/696
	1000	0/0	0/0	0/0	0/0	73/7874	15/1313	12/813
CAIC	50	0/0	0/1	0/52	6/526	89/8940	4/381	1/100
	100	0/0	0/0	0/0	0/15	99/9698	1/252	0/35
	1000	0/0	0/0	0/0	0/0	99/9947	1/49	0/4
SBC	50	0/0	0/1	0/30	3/371	87/8751	7/613	3/234
	100	0/0	0/0	0/0	0/9	97/9490	3/414	0/87
	1000	0/0	0/0	0/0	0/0	99/9890	1/100	0/10
$ICOMP_{C1}$	50	0/0	0/0	0/0	0/41	95/9437	4/407	1/115
	100	0/0	0/0	0/0	0/0	97/9532	3/387	0/81
	1000	0/0	0/0	0/0	0/0	96/9615	4/331	0/54
$ICOMP_{C1F}$	50	0/0	0/0	0/0	0/22	50/5152	31/2849	19/1977
	100	0/0	0/0	0/0	0/0	53/5041	26/2969	21/1990
	1000	0/0	0/0	0/0	0/0	37/5007	39/3028	24/1965
$ICOMP_{Mis}$	50	0/0	0/0	0/0	0/100	93/9236	6/532	1/132
	100	0/0	0/0	0/0	0/2	97/9423	3/482	0/93
	1000	0/0	0/0	0/0	0/0	94/9578	6/362	0/60

Table 2. Frequency of Model Selection Given Intermediate Variability with True Model (100/10,000 runs)

Criterion	*n*	1	2	3	4	5*	6	7
AIC	50	0/6	0/142	10/644	14/1696	54/5190	13/1282	9/1040
	100	0/1	1/53	2/305	14/1204	61/6221	12/1334	10/882
	1000	0/0	0/0	0/0	0/6	73/7818	15/1337	12/839
*AIC*c	50	0/9	0/198	12/854	17/2033	57/5541	8/877	6/488
	100	0/2	1/62	2/343	15/1333	66/6510	9/1123	7/627
	1000	0/0	0/0	0/0	0/6	73/7870	15/1313	12/811
CAIC	50	1/111	2/602	21/1658	20/2459	51/4864	4/240	1/66
	100	0/16	4/292	9/917	26/2144	60/6428	1/181	0/22
	1000	0/0	0/0	0/0	0/25	99/9922	1/49	0/4
SBC	50	1/51	1/412	18/1316	20/2318	54/5292	4/434	2/177
	100	0/11	4/212	9/751	21/1954	63/6681	3/329	0/62
	1000	0/0	0/0	0/0	0/23	99/9867	1/100	0/10
$ICOMP_{C1}$	50	0/0	0/5	0/99	10/831	85/8548	4/403	1/114
	100	0/0	0/5	0/31	7/447	90/9051	3/386	0/80
	1000	0/0	0/0	0/0	0/1	96/9614	4/331	0/54
$ICOMP_{C1F}$	50	0/0	0/1	0/42	5/411	45/4761	31/2822	19/1963
	100	0/0	0/1	0/11	5/172	49/4863	25/2966	21/1987
	1000	0/0	0/0	0/0	0/0	37/5007	39/3028	24/1965
$ICOMP_{Mis}$	50	0/0	0/31	2/195	10/1232	81/7924	6/492	1/126
	100	0/0	0/9	1/78	6/639	90/8719	3/467	0/88
	1000	0/0	0/0	0/0	0/2	94/9576	6/362	0/60

* The true model

Table 3. Frequency of Model Selection Given Minimum Variability with True Model (100/10,000 runs)

Criterion	*n*	1	2	3	4	5*	6	7
AIC	50	0/0	3/133	4/755	13/1599	58/5205	14/1271	8/1037
	100	0/0	1/67	6/647	15/1567	61/5608	8/1262	9/849
	1000	0/0	0/19	10/504	13/1500	56/6059	13/1162	8/756
*AIC*c	50	0/0	3/184	5/969	15/1886	60/5616	12/867	5/478
	100	0/0	1/77	6/735	17/1693	63/5832	6/1050	7/613
	1000	0/0	0/19	11/509	12/1508	56/6098	13/1139	8/727
CAIC	50	1/9	5/576	13/1854	21/2244	56/4977	3/263	1/77
	100	0/0	4/331	19/1799	20/2252	56/5430	1/161	0/27
	1000	0/0	1/106	20/2060	31/2344	47/5462	1/27	0/1
SBC	50	0/2	4/405	7/1494	20/2149	61/5337	6/450	2/163
	100	0/0	2/228	14/1507	21/2173	61/5741	2/283	0/68
	1000	0/0	1/87	19/1832	23/2287	56/5734	1/56	0/4
$ICOMP_{C1}$	50	0/0	0/2	2/80	6/680	87/8719	4/404	1/115
	100	0/0	0/2	0/50	6/564	91/8917	3/386	0/81
	1000	0/0	0/0	1/30	2/458	93/9130	4/328	0/54
$ICOMP_{C1F}$	50	0/0	0/0	0/24	1/265	49/4894	31/2843	19/1974
	100	0/0	0/0	0/6	3/183	50/4856	26/2967	21/1988
	1000	0/0	0/0	0/0	1/83	36/4924	39/3028	24/1965
$ICOMP_{Mis}$	50	0/0	0/16	2/194	10/1091	83/8080	4/493	1/126
	100	0/0	0/6	0/123	8/889	89/8428	3/463	0/91
	1000	0/0	0/0	2/52	4/543	88/8989	6/356	0/60

* The true model

have a turning point when the sample size is medium, indicating that they perform the best when the sample is neither largest nor smallest. This observation under only 100 simulations is not consistent with that when the number of runs is 10,000, thus we consider it to be untrustworthy due to the small number of simulations. Finally, the performance of $ICOMP_{C1F}$ does not seem to be very consistent with that of the rest. Its performance under 10,000 runs of simulations increases only slightly when the sample size jumps from 50 to as large as 1,000, whereas all other criteria show a marked increase in the average percentage of identifying the true model when increasing the sample size.

(2) The increase in the variability of (**Xβ**) tends to improve the performance of all seven criteria. This trend is clearly indicated in Figure 2 for both sets of simulations for six of the seven criteria (excluding $ICOMP_{C1F}$); and, the two trend lines representing 100 and 10,000 simulations in each of the six graphs almost completely overlap, so that they are almost indistinguishable from each other. When sample size increases from 50 to 1,000, a marked increase in the average percentage of identifying the true model is observed for *AIC* (approximately from 60% to 78%), AIC_C (approximately from 60% to 80%), *SBC* (approximately from 60% to 96%), and *CAIC* (approximately from 58% to 98%). A relative moderate increase is observed for $ICOMP_{C1}$ (approximately from 90% to 99%) and $ICOMP_{Mis}$ (approximately from 90% to 98%). These two *ICOMP* criteria are already successful at as high as 90% of the time when (**Xβ**) assumes the minimum variability, so there is not much room for improvement for the two of them given more variability in (**Xβ**). Finally, $ICOMP_{C1F}$ fails to meet our expectations again this time. When the other criteria are becoming more and more capable of identifying the true model with increasing variability in (**Xβ**), the increase in the performance of $ICOMP_{C1F}$ is negligible under the larger set of simulations.

(3) An overall comparison of all seven criteria is found in Figures 1, 2, and 3. In Figures 1 and 2, it can be seen that on average both $ICOMP_{C1}$ and $ICOMP_{Mis}$ tend to outperform non-*ICOMP*

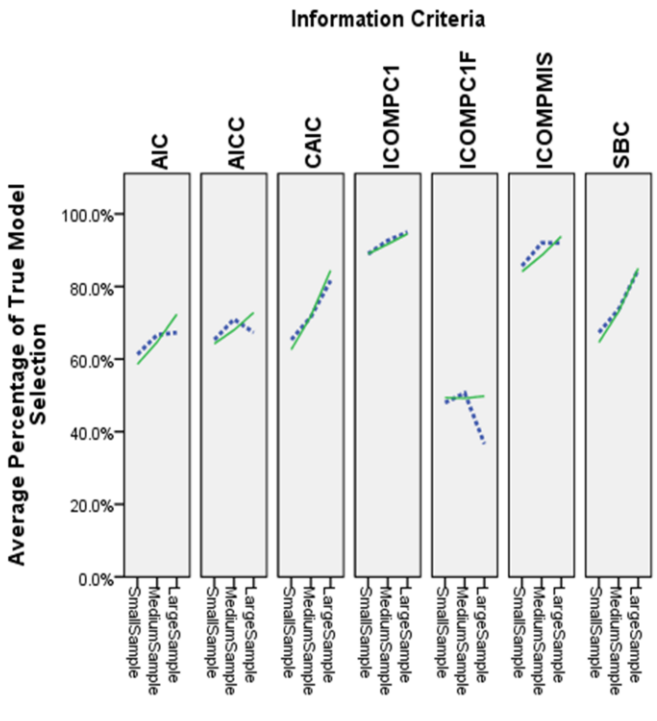

Figure 1. Comparison of average percentage of true model selection (Model 5) as a function of sample s under 100 and 10000 runs of simulations.

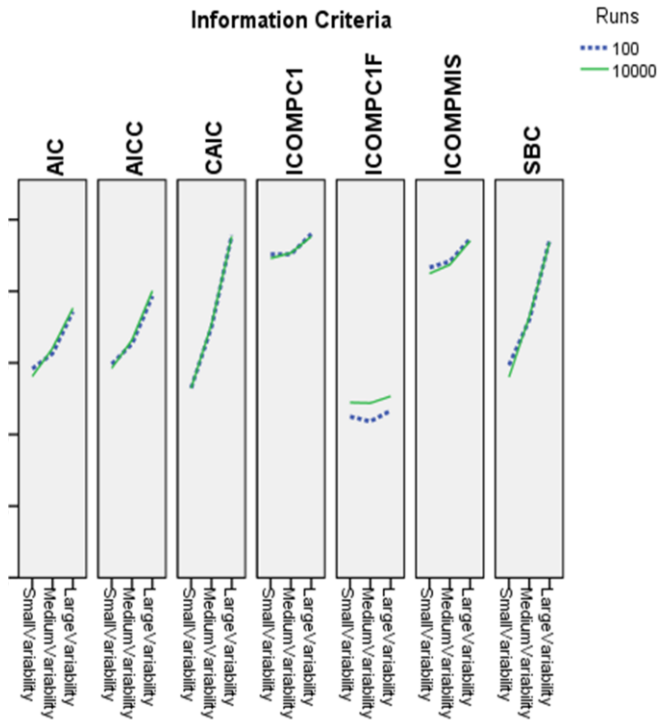

Figure 2. Comparison of average percentage of true model selection (Model 5) as a function of variability under 100 and 10000 runs of simulations.

criteria: *AIC*, AIC_C, *SBC*, and *CAIC*, and, in Figure 3, the range of percentages of successfully identifying the true model from each simulation condition tends to be higher for the two *ICOMP* criteria than for all other criteria. However, $ICOMP_{C1F}$ does not seem to perform as well as the other two *ICOMP* criteria, and is probably the worst of all seven criteria in terms of the likelihood of identifying the true model. The bad performance of this criterion is due to its tendency to select more complex models, either Model 6 or Model 7. In Figure 3, such an overfitting tendency of $ICOMP_{C1F}$ is clearly observed. This criterion is much more likely to select either Model 6 or Model 7 than all other criteria, thus causing it to be less successful in identifying the true model, or Model 5.

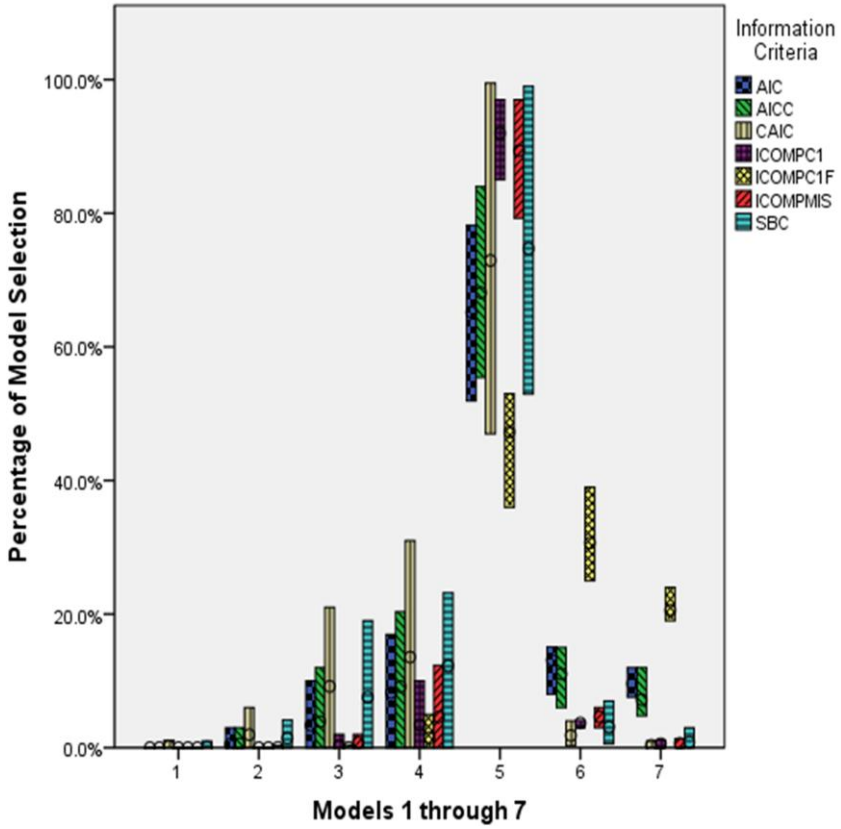

Figure 3. Comparison of range of percentages of model selection by all seven information criteria with true model.

With the True Model Excluded

Tables 4, 5, and 6 present the model selection results from the case when the true model is excluded, with Table 4 corresponding to $\boldsymbol{\beta}_{max}$, Table 5 to $\boldsymbol{\beta}_{int}$, and Table 6 to $\boldsymbol{\beta}_{min}$. In each figure, seven model selection criteria are scored to evaluate four regression models: Models 1 to 4 described above, with Model 4 being the best approximating model of the true model: Model 5. Three different sample sizes (i.e., small, medium, and large) are used, namely n_{min} = 50, n_{int} = 100, and n_{max} = 1000.

Similar to the previous case with the true model included, under each $\boldsymbol{\beta}$ by n combination, two sets of simulations are performed for the purpose of cross-validating model selection results. The first set contains 100 runs of simulations whereas the second set 10,000 runs. So, cells in each of Tables 4, 5, and 6 also contain two integers separated by a forward slash sign which represent frequencies of each competing model being selected under the two sets of simulations (100 runs/10,000 runs), respectively.

Besides, Figures 4 and 5 present the average percentage of the best approximating model (Model 4) selection as a function of sample size and variability in ($\mathbf{X\beta}$), respectively. Finally, Figure 6 compares all seven criteria using the range of percentages of each of Models 1 through 4 being selected under each simulation condition.

Under the second case, where Model 4 is the best, similar patterns of criterion performance are found. In Figure 4, the two lines of 100 and 10,000 simulations both show a continuing upward trend with an increase in sample size for all seven criteria (i.e., the $ICOMP_{C1}$ line for the smaller number of simulations

Table 4. Frequency of Model Selection Given Maximum Variability Without True Model (100/10,000 runs)

Criterion	*n*	1	2	3	4*
AIC	50	0/0	1/1	2/278	97/9721
	100	0/0	0/0	0/9	100/9991
	1000	0/0	0/0	0/0	100/10000
*AIC*c	50	0/0	1/2	5/346	94/9652
	100	0/0	0/0	0/14	100/9986
	1000	0/0	0/0	0/0	100/10000
CAIC	50	0/0	1/25	9/828	90/9147
	100	0/0	0/0	1/76	99/9924
	1000	0/0	0/0	0/0	100/10000
SBC	50	0/0	1/12	7/611	92/9377
	100	0/0	0/0	1/51	99/9949
	1000	0/0	0/0	0/0	100/10000
$ICOMP_{C1}$	50	0/0	0/0	0/40	100/9960
	100	0/0	0/0	0/0	100/10000
	1000	0/0	0/0	0/0	100/10000
$ICOMP_{C1F}$	50	0/0	0/0	0/31	100/9969
	100	0/0	0/0	0/0	100/10000
	1000	0/0	0/0	0/0	100/10000
$ICOMP_{Mis}$	50	0/0	0/0	1/153	99/9847
	100	0/0	0/0	0/3	100/9997
	1000	0/0	0/0	0/0	100/10000

Table 5. Frequency of Model Selection Given Intermediate Variability Without True Model (100/10,000 runs)

Criterion	*n*	1	2	3	4*
AIC	50	0/38	4/392	25/2023	71/7547
	100	0/1	2/218	12/1226	86/8555
	1000	0/0	0/0	0/3	100/9997
*AIC*c	50	0/45	5/481	27/2244	68/7230
	100	0/2	3/234	12/1285	85/8479
	1000	0/0	0/0	0/3	100/9997
CAIC	50	1/238	7/1068	39/3041	53/5653
	100	0/35	7/634	21/2133	72/7198
	1000	0/0	0/2	0/10	100/9988
SBC	50	1/145	6/828	33/2762	60/6265
	100	0/23	5/513	20/1919	75/7545
	1000	0/0	0/1	0/8	100/9991
$ICOMP_{C1}$	50	0/5	2/83	8/747	90/9165
	100	0/0	0/39	4/380	96/9581
	1000	0/0	0/0	0/0	100/10000
$ICOMP_{C1F}$	50	0/5	2/58	4/575	94/9362
	100	0/0	0/25	3/266	97/9709
	1000	0/0	0/0	0/0	100/10000
$ICOMP_{Mis}$	50	0/6	1/159	13/1092	86/8743
	100	0/0	0/57	7/579	93/9364
	1000	0/0	0/0	0/0	100/10000

* The best approximating model

may deviate a little bit, though), thus supporting their property of consistency. In Figure 5, such a continuing upward trend is also observed for all seven criteria when the variability in (**Xβ**) increases. Finally, the performance of *ICOMP* criteria is generally better than that of non-*ICOMP* criteria. This is true of all three *ICOMP* criteria. In Figures 4 and 5, the average performance of each *ICOMP* criterion under smallest sample size or smallest (**Xβ**) variability is generally the same as or even better than that of each non-*ICOMP* criterion under largest sample size or largest (**Xβ**) variability. In Figure 6, the range of percentages of successfully identifying the best approximating model under each simulation condition tends to be higher for the three *ICOMP* criteria than for the four non-*ICOMP* criteria. Although $ICOMP_{C1F}$ performs less satisfactorily in the previous case that includes the true model, it performs as well as the other two *ICOMP* criteria in this second case. Such an increase in performance is probably

Table 6. Frequency of Model Selection Given Minimum Variability Without True Model (100/10,000 runs)

Criterion	*n*	1	2	3	4*
AIC	50	3/127	5/572	20/2720	72/6581
	100	1/10	2/296	29/2533	68/7161
	1000	0/0	0/58	22/2094	78/7848
AICc	50	3/177	8/680	20/2921	69/6222
	100	1/10	3/315	29/2665	67/7010
	1000	0/0	0/58	23/2099	77/7843
CAIC	50	8/604	14/1269	29/3513	49/4614
	100	1/48	9/767	43/3765	47/5420
	1000	0/0	1/196	39/4023	60/5781
SBC	50	8/401	10/1042	26/3322	56/5235
	100	1/31	8/618	39/3549	52/5802
	1000	0/0	1/167	37/3821	62/6012
$ICOMP_{C1}$	50	1/3	1/76	6/883	92/9038
	100	0/1	0/34	2/677	98/9288
	1000	0/0	0/4	10/523	90/9473
$ICOMP_{C1F}$	50	1/3	0/32	6/582	93/9383
	100	0/1	0/9	1/388	99/9602
	1000	0/0	0/1	2/184	98/9815
$ICOMP_{Mis}$	50	1/12	3/159	6/1509	90/8320
	100	0/2	0/54	10/1077	90/8867
	1000	0/0	0/8	10/616	90/9376

* The best approximating model

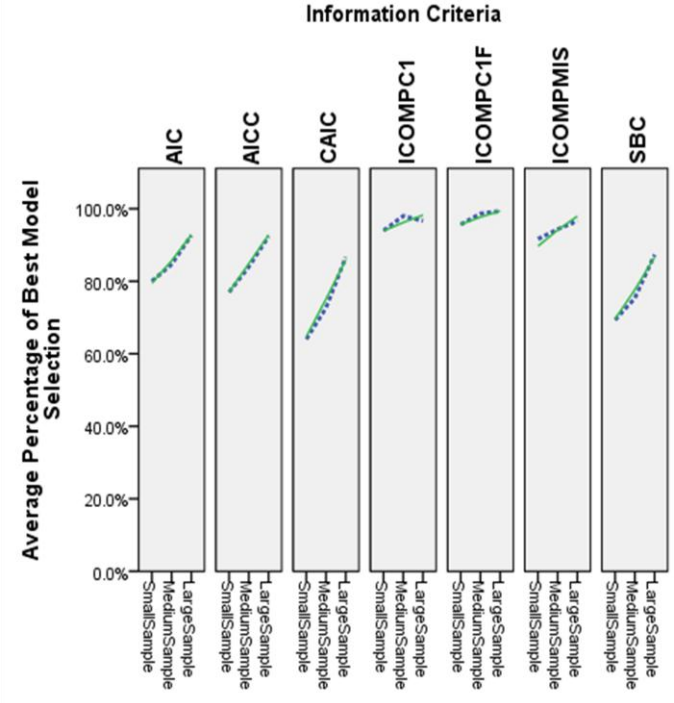

Figure 4. Comparison of average percentage of best approximating model selection (Model 4) as a function of sample size under 100 and 10000 runs of simulations.

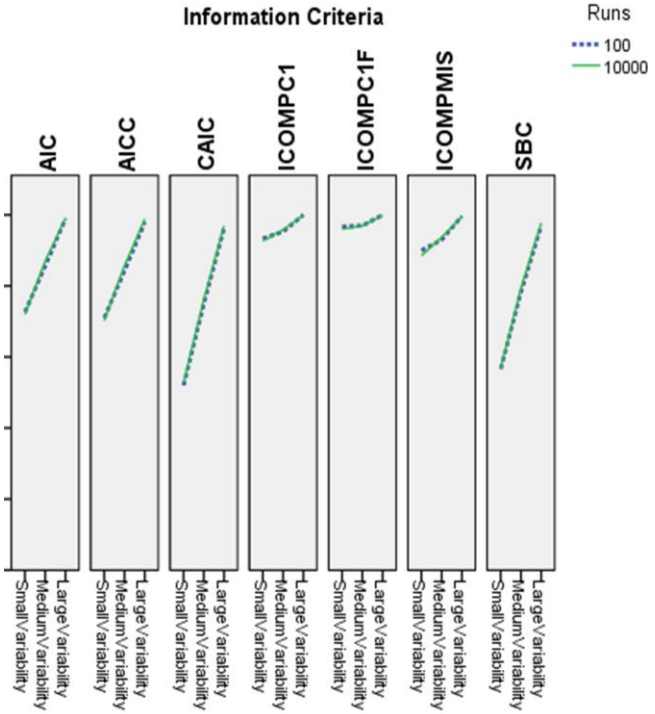

Figure 5. Comparison of average percentage of best approximating model selection (Model 4) as a function of variability under 100 and 10000 runs of simulations.

because this criterion tends to overfit a model and the best approximating model in the second case is already the most complex model. In other words, in both cases, $ICOMP_{C1F}$ tends to select a more complex model and, in the second case only, the most complex model happens to be the best approximating model.

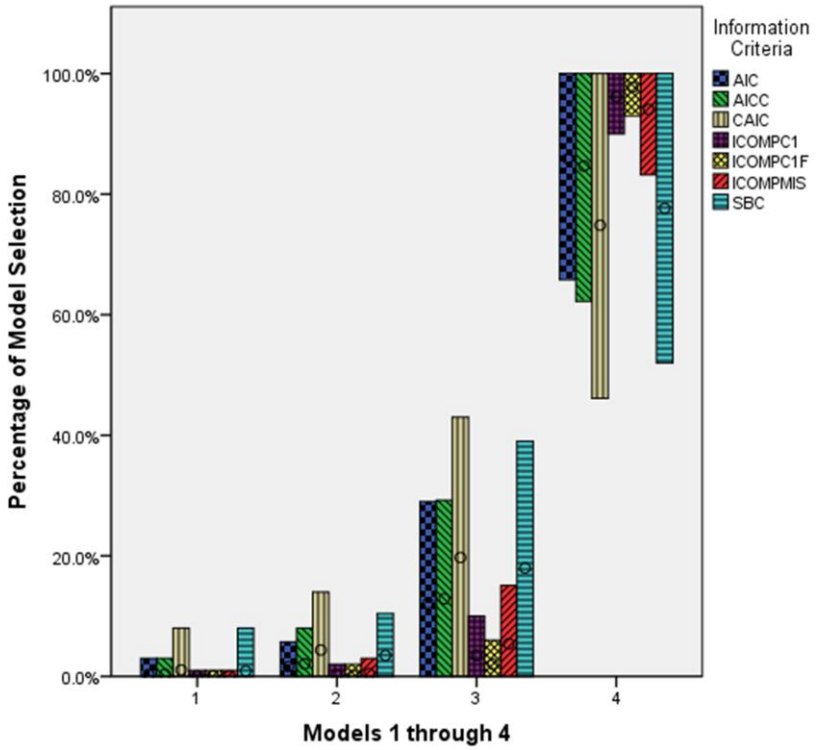

Figure 6. Comparison of range of percentages of model selection by all seven information criteria without true model.

Conclusion

The paper provides support for the use of two *ICOMP* criteria in multiple linear regression to supplement existing information criteria commonly found in major statistics programs: *AIC*, *CAIC*, *SBC*, etc. The two recommended *ICOMP* criteria are $ICOMP_{C1}$ and $ICOMP_{Mis}$. However, this paper has some reservations for the third *ICOMP* criterion, or $ICOMP_{C1F}$, because it is usually prone to overfitting.

The two recommended *ICOMP* criteria are usually more capable of successfully identifying the best approximating model than other criteria under the simulations of multiple linear regression modeling in this study. And their effectiveness can generally be improved by either increasing sample size or increasing the variability in (**Xβ**).

Future research on *ICOMP* could focus on its application to linear and nonlinear mixed models, which are extensions of the type of linear models covered in this paper. Mixed models consist of both fixed and random components and are capable of analyzing grouped, nested, or hierarchical data structures that are more commonly seen in many fields of study. *ICOMP* would be used to select fixed and/or random components in mixed models. Special *ICOMP* formulas should be developed for mixed models that correspond to formulas for marginal and conditional *AIC* (Vaida & Blanchard, 2005).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267.281). Budapest, Hungary: Akademiai Kiado.

Akaike, H. (1987). Factor analysis and *AIC*. *Psychometrika, 52*, 317-332.

Ando, T. (2010). *Bayesian model selection and statistical modeling*. Boca Raton, FL: Chapman and Hall.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (*AIC*): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

Bozdogan, H. (2004). Intelligent statistical data mining with information complexity and genetic algorithms. In H. Bozdogan (Ed.), *Statistical data mining and knowledge discovery* (pp. 15-56). Boca Raton, FL: Chapman and Hall.

Bozdogan, H., & Haughton, H. (1998). Information complexity criteria for regression models. *Computational Statistics and Data Analysis, 28*, 51-76.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge, UK: Cambridge University Press.

Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Konishi, S., & Kitagawa, G. (2010). *Information criteria and statistical modeling*. New York: Springer.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22,* 79-86.

Kutner, M., Nachtsheim, C., & Neter, J. (2004). *Applied linear regression models* (4th ed.). New York: McGraw-Hill.

McQuarrie, A. D. R., & Tsai, C. L. (1998). *Regression and time series model selection*. River Edge, NJ: World Scientific Publishing.

Miller, A. (2002). *Subset selection in regression* (2nd ed.). Boca Raton, FL: Chapman and Hall.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Boca Raton, FL: Chapman and Hall.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52,* 333-343.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall.

Tamhane, A. C., & Dunlop, D. (1999). *Statistics and data analysis: From elementary to intermediate.* Upper Saddle River, NJ: Prentice Hall.

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika, 92,* 351-370.

Van Emden, M. H. (1971). An analysis of complexity. In *Mathematical Centre Tracts* (Vol. 35). Amsterdam, Netherlands: Mathematisch Centrum.

Send correspondence to:
Hongwei Yang
University of Kentucky
Email: hya222@uky.edu

Multiple Linear Regression: A Return to Basics in Educational Research

Winona Burt Vesey
University of Houston, Clear Lake

Jermaine T. Vesey
University of Texas, San Antonio

Antionette D. Stroter
Liberty University

Kyndra V. Middleton
Howard University

A content analysis of the *American Educational Research Journal* and the *Educational Evaluation and Policy Analysis* journal for the use of multiple linear regression (MLR) was conducted. Two hundred articles were randomly sampled and coded to determine if basic reporting guidelines were followed. Results showed that standard reporting methods for MLR were not followed and the use of stepwise MLR was on the rise. Manuscripts using MLR did not apply necessary corrections for inflated Type I error rates. The majority of the sampled articles did not include key summary statistics, which violated the American Educational Research Association's principle of transparency. The lack of consistency in reporting hindered critique of work, meta-analysis, and theory development.

Multiple linear regression (MLR) is a common statistical technique used in educational research (Elmore & Woehlke, 1996). Used in experimental and non-experimental research designs alike, MLR involves the use of one or more predictor (i.e., independent) variables predicting some criterion (i.e., dependent) variable. Research conclusions based on MLR are used to influence education policy decisions (Clements & Sarama, 2008; Ingersoll, 2001; Stern, Dayton, Paik, & Weisberg, 1989), inform school reform efforts (Desimone, Smith, Baker, & Uano, 2005), determine variables considered in college admission (e.g., Zwick & Sklar, 2005), and identify relationships between school climate and student achievement (e.g., McInerney, Roche, McInerney, & Marsh, 1997; Pianta, Belsky, Vandergrift, Houts, & Morrison, 2008). Given the pervasive use of this technique in educational research, the improper interpretation of MLR results are far reaching. Hierarchical MLR (HMLR) and stepwise MLR (SMLR) are particularly susceptible to misreporting. Regrettably, as is detailed in this study, many educational researchers fail to avoid common misinterpretations of HMLR and SMLR. Research and statistical analysis classes typically focus on the scientific method and its fundamental guidance to conducting good research. Specifically, theory is developed and tested to determine if it can be supported. When theories endure multiple tests they are regarded as robust and acceptable – statistical analysis is a critical component in this process. In this respect, replication and statistical validity are critical to the development of sound theory. If researchers are unable to replicate research due to the lack of standardized reporting, or if researchers are concluding that hypotheses are supported when in fact they are not; proper development and acceptance of theory is jeopardized.

To facilitate replication of studies, transparency in reporting is essential. Transparency in research findings refers to the practice of revealing the key methodological components necessary for scrutiny and replication of research findings (American Educational Research Association (AERA), 2006). Transparency allows researchers to critique the statistical validity of a study, conduct replication studies, and ultimately ensure the accuracy of research claims. Despite the known importance of standard reporting practices and statistical validity, many researchers have failed to provide adequate transparency in their analysis. The purpose of this paper is to call attention to the frequency of inaccurate statistical claims (i.e., specifically as related to regression), call for statistical reporting standards, and provide recommendations to aid in the development of said standards.

.

Importance of Study

The use of MLR is common among education researchers (Elmore & Woehlke, 1996). In Elmore and Woehlke's review of articles published in the *American Educational Research Journal* (AERJ), *Educational Researcher* (ER), and the *Review of Educational Research* (RER) from 1988 to 1995, MLR/correlation emerged as the third most used statistical method following analysis of variance/analysis of covariance and descriptive methods (1996). MLR continues to be utilized in contemporary research. In our random sample of 200 articles, 35% of our articles were published from 1968 to 1995. The remaining 65% of the articles were published between 1996 and 2008, supporting the pervasiveness of the method in

recent years. Given the popularity of MLR (Elmore & Woehlke), it is important to establish standardized reporting conventions and ensure that MLR results are reported accurately. In a recent article, Zientek and Thompson (2009) called for at a minimum, the reporting of matrix summaries when using continuous data (e.g., correlation matrix and standard deviations, or the variance-covariance matrices) as matrices support and encourage meta-analytic thinking. In an earlier paper, Thompson (2007) also called for basic reporting and research standards. The following quote from Thompson reflects his belief in the value of creating statistical reporting standards that reflect transparency and enable replication: "Vital aspects of scholarship include exposing one's conclusions and their warrants to public scrutiny and disseminating one's findings" (Thompson, 2007, p. 18).

Standards for Reporting Empirical Research

Having standards in reporting statistical findings is important for a variety of reasons, namely it makes it easier to generalize findings across fields, ensures accuracy, enables understanding, provides information necessary for replication, and allows researchers to conduct meta-analysis. Our call for reporting standards is shared by other researchers and professional organizations. For example, both AERA and the American Psychological Association (APA) provide guidelines that encourage reporting standards and transparency. Specifically, AERA (2006) outlines two overarching principles for reporting empirical research: sufficiency and transparency. In short, adequate evidence should be provided to support results and conclusions and reports should be transparent. "Reporting that takes these principles into account permits scholars to understand one another's work, prepares that work for public scrutiny, and enables others to use that work" (AERA, 2006, p. 33). AERA provides further detail for the area of analysis and interpretation reporting. Specifically, for quantitative methods, AERA calls for a statement of statistical analyses and why they were appropriate; descriptive and inferential statistics; discussion of considerations that arose during data collection and processing such as, missing data or attrition; considerations identified as a result of data analysis (e.g., violations of assumptions); and inclusion of a measure of effect size for each statistical result; standard error or confidence interval; test statistics and its significance level for hypothesis testing; and a qualitative description of the index of the effect.

The APA (2010) publication manual also provides guidelines for standard reporting practices. Specifically, they call for summary descriptive data, variance-covariance or correlation matrices, and results of inferential statistics (e.g., observed values, degrees of freedom, *p* values, standard errors, and effect sizes).

Conventions for reporting analysis of variance, *t*-tests, and correlation results have been well established for many years (Daniel, 2001; Schafer, 1991). Nonetheless, similar standards for reporting MLR results have not been established (Courville & Thompson, 2001; Schafer). In his 1991 editorial, Schafer offers recommendations for reporting hierarchical regression results. Schafer proposes that authors report both descriptive and inferential statistics (e.g., correlation matrix for the predictors, *df* column, R^2 change column, and *p* values for each *F* ratio), and predictors listed in the order of their inclusion in the analysis. Schafer emphasizes reporting results with sufficient information in an interpretable way without losing the ability for replication, thereby highlighting the importance of transparency in research. Schafer further notes that if exact *p* values are reported, then it is not as important to indicate which are below the alpha level set by the researcher, but simply indicate the *a priori* alpha level in the text. Near the end of the article, Schafer states "Whether this or some other format becomes popular remains to be seen, but it seems clear that some conventional way to report multiple regression outcomes is needed" (Schafer, p. 3). As evidenced in the content review below, conventional reporting for MLR has yet to take hold. Moreover, a standard for identifying the type of MLR and proper adjustments for experimentwise error rates were absent in most cases, resulting in inflated Type I error rates and inaccurate statistical claims.

Multiple Regression Typology

Multiple regression is the process of predicting a dependent, outcome variable from a set of independent, predictor variables. The dependent and independent variables can take several forms: continuous, dichotomous, or polytomous, to name a few. The level of measurement of the dependent variable (DV) is typically used to define the multiple regression technique. For example, in logistic regression the dependent variable is dichotomous; whereas in MLR the dependent variable is continuous.

Regression techniques are further classified by the procedures used to enter the independent variables (IV) to obtain the final equation. Simultaneous MLR is when all independent variables are entered into the regression equation at one time, which results in one hypothesis being tested. Hierarchical multiple linear regression (HMLR) is when the entry order of the independent variables is predetermined by the researcher and there are two or more stages of variable entry into the regression equation. For example, a researcher might first enter demographic variables as control variables in stage one and then enter a second set of variables in stage two, and focus on a change in R^2 (i.e., the amount of variance in the DVs accounted for by the IVs) at each stage. As a second example, the researcher might enter a single IV, such as self-efficacy, followed by another single variable, such as school climate, at stage two. This process would continue for all variables the researcher chooses to enter. Lastly, stepwise regression (also known as empirical multiple regression) is when statistical software determines the entry order of IVs based on which variables contribute most to prediction at a given step in the regression equation (Hoyt, Leierer, & Millington, 2006).

For the purposes of this study, we have categorized multiple regression techniques using the following terms: simultaneous multiple regression, hierarchical multiple regression, and stepwise multiple regression (Hoyt et al., 2006). Methodologists in the psychological community have recommended that SMLR be used rarely, or not at all, in academic research (Cohen, Cohen, West, & Aiken, 2003; Thompson, 1995). The primary reason was that stepwise procedures yield data-dependent results that are unlikely to generalize to future samples (Hoyt et al.). These authors take a similar stance and recommend that researchers never use the stepwise method in education or any other discipline. A computer program is not sufficient to determine the importance of a variable; instead the literature and theory should guide decisions. Additionally, statistical conclusion validity, as discussed in the next section, is often violated when conducting stepwise regression. Furthermore, as demonstrated in our content analysis, neither hierarchical nor stepwise regressions are properly identified in education research and results are often misreported.

Compromising Statistical Conclusion Validity

Statistical conclusion validity refers to the accuracy of a conclusion regarding the relationship between variables (Shaddish, Cook, & Campbell, 2002). In MLR, statistical conclusion validity is often violated when researchers fail to properly adjust alpha levels to compensate for multiple hypothesis testing. When researchers use HMLR or SMLR to test multiple combinations of variables, they are in fact testing multiple hypotheses. Each variable or set of variables entered into and removed from the regression equation represents a separate hypothesis test. When using stepwise regression procedures (e.g., forward and backward regression), the software executes the adding and removing of the variables and provides the model with the best R^2 to the researcher. This is particularly dangerous if the researcher is unaware of the exact number and order of steps used by the software to derive the final model and further supports the authors' position to never use SMLR.

Using hierarchical and stepwise regression and not adjusting the alpha level is the same as testing multiple hypotheses while holding the alpha level constant. The reason for the adjustment of the alpha level is the same reason that researchers conduct an ANOVA when there are three or more groups being compared. If three separate *t* tests are conducted, the result of this practice is that the researcher is testing the hypothesis at an inflated Type I error rate, which could result in variables being identified as "significant" predictors when they are not. Additionally, running simultaneous regression and not adjusting the alpha level produces a similar problem as hierarchical and stepwise regression in that multiple tests are performed on the same sample data in an attempt to find the best predictors. Not adjusting for the multiple tests can inflate Type I error if several different predictors are tested. In the following section, we provide a more thorough explanation of the experimentwise error rate problem.

Error Rates

To begin, we offer a brief reminder of the difference between alpha levels, *p* values, and error rates. The alpha level is the standard set by the researcher before statistical tests are conducted. Alpha levels are commonly set at .05, .01, and .001 in education research and determine the probability of obtaining a sample mean in the critical region when the null hypothesis is true. In other words, the alpha level controls the risk of making a mistake or a Type I error (Gravetter & Wallnau, 2007). Ultimately then, the

risk of a Type I error is in the control of the researcher. The *p* value, or probability value, is related to the test statistic and defines the probability of observing the sample results actually obtained, given that the null hypothesis is true. When the *p* value is equal to or less than the alpha level, the null hypothesis is rejected. Lastly, the error rate, or Type I error (synonymous with the alpha level), is also defined as the probability of rejecting a null hypothesis when in fact it is true (Salkind, 2007).

When conducting multiple hypothesis tests in an experiment, you have both a testwise error rate and an experimentwise error rate. Testwise error rate is the probability of making a Type I error in a single hypothesis test and should be set by researchers *a priori*. Experimentwise error rate, also known as familywise error rate, is the probability of having made a Type I error within a set of hypothesis tests (Thompson, 1995). The experimentwise error rate is inflated for every hypothesis tested on a single set of data in a given experiment (Altman, 2000). Each hypothesis test conducted can be considered as a separate experiment.

Experimentwise Type I error rate is affected by the number of tests (hypotheses) ran using a single sample (Thompson, 1995). When conducting multiple hypothesis tests, the inflated experimentwise error rate (α_{ew}) can be calculated using the Bonferroni inequality (Love, 1988):

$$\alpha_{ew} \leq 1 - (1 - \alpha_{Tw})^k \quad (1)$$

where k is the number of perfectly uncorrelated hypotheses being tested and α_{Tw} is the testwise alpha level (Altman, 2000). As an example, if you have three different models with variables being entered separately, an alpha initially set at .05 becomes an alpha of .14 using the Bonferroni inequality.

Mundfrom, Perrett, Schaffer, Piccone, and Roozeboom (2006) further propose that when unadjusted *t* tests are used for individual variable selection in simultaneous linear regression, Type I error is even further affected. Researchers using unadjusted alpha levels exponentially inflate the Type I error rate depending on the number of independent variables in the model and the number of independent variables that are correlated with the dependent variable (Mundfrom et al.). As a result, Type I errors are committed, which means variables are identified as "significant" predictors when in fact they may not be. Mundfrom et al. suggest that when conducting multiple hypothesis tests, researchers should control the testwise error rate by using the Bonferroni correction (Altman, 2000):

$$\alpha_{Tw}^{*} = \alpha_{Tw/k} \quad (2)$$

where k is the number of hypothesis tests being conducted and α_{Tw} is the testwise error rate. Roozeboom, Mundfrom, and Perrett (2008) later developed a modified Bonferroni correction in an effort to maintain greater statistical power

$$\alpha_{Tw}^{*} = \alpha_{Tw/k(1-q)} \quad (3)$$

where the numerator remains the same nominal alpha value as in equation 2, but the denominator becomes the number of tests performed (k), multiplied by one minus the proportion of nonzero relationships between the dependent and independent variables.

In general, the Bonferroni correction (also known as the Dunn test) adjusts the inflated experiment-wise alpha level by dividing the original testwise error rate (α_{Tw}) by the number of hypotheses being tested (k) yielding a new testwise error rate (α_{Tw}^{*}). Consequently each hypothesis (or *post hoc*) test uses the new testwise error rate to keep the experimentwise error rate at the appropriate level. For example, if there are three comparisons made with an overall alpha level of .05, each comparison would be held to an alpha level of .02 (i.e., .05/3 = .02), thereby maintaining the experiment wise error rate of .05 (Gravetter & Wallnau, 2007). Below we mention additional alternatives to the traditional Bonferroni correction that purport to have greater power than Bonferroni's correction yet maintain its flexibility for use with tests such as MLR and correlations.

Sidak-Bonferroni

Sidak (1967) suggested a modification of the Bonferroni formula that would have less impact on statistical power than the Bonferroni method and retain much of its flexibility (Keppel & Wickens, 2004). Instead of dividing by the number of comparisons, there is a slightly more complicated formula:

$$\alpha_{S\text{-}B} = 1 - (1 - \alpha_{FWE})^{1/c} \quad (4)$$

where $\alpha_{S\text{-}B}$ is the Sidak-Bonferroni alpha level used to determine statistical significance (a value less than .05), α_{FWE} is the computed testwise error according to Formula 1, and *c* is the number of comparisons or

statistical tests conducted in the study. The *p* values obtained from the results of the analysis must be smaller than $\alpha_{S\text{-}B}$ to be considered significant (Olejnik, Li, Supattathum, & Huberty, 1997).

Methodology

Data Sources and Procedures

We used a qualitative research design to assess the current multiple regression reporting standards. Specifically, we conducted a content analysis of two educational research journals published by the AERA: the AERJ and the *Educational Evaluation and Policy Analysis* (EEPA) journal. The Sage search engine was used to conduct a query using the key word regression, within each of the two journals. The search created a sampling frame of 590 articles in AERJ and 289 articles in EEPA. One hundred articles were then randomly selected from the sampling frames of each journal, resulting in a final sample of 200 articles. The articles were then analyzed using a qualitative analysis approach; document/content analysis (Creswell, 2003). Articles that did not contain actual multiple regression techniques were replaced. For example, some of the randomly selected articles were book reviews or made a one-line reference to the term regression but did not conduct an analysis using regression. Articles reviewed in AERJ covered 40 years ranging from 1968 to 2008 and the articles in EEPA ranged from 1979 to 2008. The difference in years occurred because the search was not restricted by years, but instead simply by the use of multiple regression and the random sampling.

Each article was then coded by two coders using the following categorizations: (a) simultaneous multiple regression, or stepwise/hierarchical regression depending on the method used to enter the independent variables, (b) whether or not corrections of Type I error rates were made for HMLR/SMLR, (c) whether authors properly identified HMLR/SMLR when used, (d) inclusion of correlation matrices, basic descriptive statistics data (e.g., mean and standard deviation), effect size statistics (e.g., R^2), standard errors, and lastly, whether *F* statistics and *t* statistics were provided.

Results

About 30% of the articles from AERJ and EEPA used HMLR/SMLR methods. Ninety percent of the articles that used HMLR/SMLR in AERJ and 95% of the articles in EEPA failed to adjust their testwise error rate. In short, of the articles that used HMLR/SMLR, only 10% of the articles in AERJ and 5% of the articles in EEPA used a procedure to ensure a reduction in Type I error rate. The remaining articles found significance when in fact, if the researchers had adjusted their testwise error rate, the results may have been different. Additionally, in light of the recommendations against the use of stepwise regression, it was noteworthy that of the 60 articles using stepwise or hierarchical regression, more than 50% of the articles were published within the last 10 years. This finding heightens the urgency of this study.

The content analysis also confirmed inconsistencies in the reporting of basic regression summary statistics, which hinders transparency and replicability. In particular, it is recommended that researchers report the following statistics when conducting regression analysis: means, standard deviations, bivariate correlations, overall *F* value, regression coefficients, R^2, and changes in R^2 (for stepwise regression methods). In Table 1, a summarization is provided listing the number of articles that included these basic data in their manuscripts.

Table 1. Percentage of Manuscripts that Included Summary Statistics

	Journal	
Descriptive Statistic	AERJ	EEPA
Means	83%	90%
Standard Deviation	77%	82%
Correlation Matrix	48%	55%
Overall *F* value	95%	95%
t statistics	80%	85%
Regression coefficients	95%	95%
R^2	95%	95%
*Change in R^2	90%	95%
Standard Errors	65%	70%

Note. *Change in R^2 is only reported for those articles that used stepwise regression

Table 2[a] Probability of Entry into a Bachelor's Program: Fall 1990 First-time Freshman (Survey Respondents) Logistic Regression

	Model 1			Model 2			Model 3		
Variable	B	SE	%	B	SE	%	B	SE	%
Immigrant origin									
Foreign born, U.S. HS	0.13	0.06	3.1*	0.30	0.09	6.8***	0.46	0.13	10.3***
Foreign born, foreign HS	-0.25	0.11	-5.8*	-0.16	0.15	-3.4	-0.07	0.16	-1.4
Race/ethnicity	--	--	--						
Black	--	--	--	-0.43	0.09	-8.5***	-0.38	0.09	-7.1***
Hispanic	--	--	--	-0.04	0.09	-0.9	-0.01	0.10	-0.2
Asian	--	--	--	0.00	0.13	0.1	-0.01	0.13	-0.2
GED	--	--	--	-0.69	0.11	-12.7***	-0.70	0.11	-12.1***
Aspirations	--	--	--	1.02	0.10	24.4***	1.02	0.10	24.2***
Gender (F = 1)	--	--	--	0.36	0.07	8.2***	0.47	0.08	10.5***
Age (minus 18)	--	--	--	-0.03	0.01	-0.6***	-0.03	0.01	-0.6
Enrolled part-time, F90	--	--	--	-0.12	0.11	-2.5	-0.10	0.11	-2.1
Supporting Children	--	--	--	-0.74	0.14	-13.4***	-0.73	0.14	-12.5***
Employment, F90	--	--	--						
Part-time	--	--	--	-0.16	0.08	-3.4**	-0.17	0.08	-3.4**
Full-time	--	--	--	-0.33	0.12	-6.6***	-0.33	0.12	-6.3***
Household income	--	--	--						
16K to 31K	--	--	--	-0.16	0.10	-3.4*	-0.15	0.10	-3.0
31K+	--	--	--	-0.15	0.10	-3.2	-0.14	0.10	-2.8
Missing income	--	--	--	-0.45	0.10	-8.8***	-0.44	0.10	-8.1***
Parent's education	--	--	--						
High school degree	--	--	--	-0.14	0.09	-3.0	-0.15	0.09	-3.0*
Some college	--	--	--	-0.11	0.10	-2.4	-0.13	0.11	-2.6
College degree	--	--	--	-0.17	0.11	-3.5	-0.19	0.11	-3.8*
Graduate/professional	--	--	--	0.18	0.13	3.9	0.16	0.13	3.4
Hybrid college	--	--	--	-1.22	0.09	-19.4***	-1.22	0.09	-18.2***
CDSEEK	--	--	--	1.28	0.10	30.7***	1.29	0.10	30.8***
Assessment tests	--	--	--						
Math	--	--	--	0.73	0.04	17.4***	0.72	0.04	16.6***
Reading	--	--	--	0.44	0.04	10.2***	0.58	0.06	13.1***
Interactions									
FB, U.S. HS*reading	--	--	--	--			-0.22	0.09	-4.2**
FB, FRGN HS*reading	--	--	--	--			-0.53	0.13	-9.5***
FB, U.S. HS*female	--	--	--	--			-0.24	0.16	-4.7
FB, FRGN HS*female	--	--	--	--			-0.56	0.26	-10.0
Constant	-0.40	0.03	40.1	-0.79	0.15	31.3	-0.91	0.16	28.8
-2 Log likelihood	7344.181			5549.253			5524.656		

Note. N = 5413 (unweighted), B = Coefficient, HS = High School, FB = Foreign Born, FRGN = Foreign, and *p < .10, **p < .05, ***p <.01. [a] Table 2 is derived from Bailey and Weininger (2002).

Adjusting Alpha Levels in Multiple Regression: An Example

Table 2 is an example from an article in the sample. In this case, the authors focused on foreign-born and native minority community college entrants at City University of New York. Stepwise logistic regression was used to predict the likelihood of entering a four-year or a two-year college program. The results show that non-native US students who immigrate to the US and graduate from a US high school are more likely than native US students to enroll in a four-year program. Additionally, those non-native US students who immigrated after high school (i.e., attended a non-US high school) are more likely to enroll in a two-year program (Bailey & Weininger, 2002).

The authors tested a total of three models. The third model is referred to as the full model and contained nearly 30 independent variables compared to only two independent variables in the initial model. Independent variables were added to the initial model until the full model was developed. All three models used the same sample. The authors report that the control variables, which were entered in model 2, had the expected influence. In particular, students who earned a GED, older students, those with jobs, those with childcare responsibilities, and those who did not aspire to a higher degree were more likely to enroll in a community college.

Interestingly, the authors report what they called a counterintuitive result concerning parental education. In the full regression model, all levels of parental education other than the highest level (i.e., attendance at graduate or professional school) exhibited negative coefficients even though only two categories were statistically significant with small effect sizes. The results suggested that if the parent had a high school degree or a college degree that those students were less likely to attend a four- year college. While these results were statistically significant, they were only marginally significant (i.e., $p < .10$). Had the researchers employed the Bonferroni correction, the testwise alpha level would have been set at .0167 and these variables would not have been statistically significant. Additionally, had the authors applied the modified Bonferroni approach proposed by Roozeboom et al. (2008), even fewer independent variables would maintain their level of significance. Roozeboom et al. posit that each time a new model is run and a decision concerning which independent variables are significant contributors is made, the Type I error rate is increased 2 to 6 times the nominal alpha level depending on the number of independent variables and their nonzero correlation with the dependent variable.

Continuing with the most basic application, given that there were three models tested in this example, employing the Bonferroni correction would result in having an adjusted alpha level of .02 (.05/3 = .017 rounded to .02) for the overall regression equation. As a result, two key interactions in the study (i.e., foreign born, US high school, reading score interaction; and foreign born, foreign high school, and female) would no longer be significant.

Regrettably, as was demonstrated, educational researchers regularly fail to adjust their testwise alpha levels when conducting HMLR/SMLR or simultaneous MLR consequently inflating their experimentwise alpha levels and Type I error rates. Zwick and Sklar (2005) is one of very few examples of authors who attempted to adjusted their testwise alpha levels. In their article published in AERJ, it was reported that "Statistical significance tests for individual predictors were conducted at an alpha level of .01 because of the large number of hypotheses being assessed" (p. 451).

Although Zwick and Sklar's (2005) article is a step in the right direction, Mundfrom et al. (2006) would suggest that they did not go far enough. The full model should evaluate each variable at the α/k level of significance, where α equals the beginning nominal alpha level and k equals the number of independent variables in the equation, which would result in an adjusted significance level of .05/30 or .002. To maintain power, the Roozeboom et al. (2008) modified Bonferroni technique could have also been applied as well.

At minimum, researchers should report the number of hypothesis tests they run (i.e., the number of models tested) so that it is clear whether the appropriate correction procedure is used. Zwick and Sklar (2005) used a common approach in education literature which is to just use .01, but if the authors had run 20 different models or hypothesis tests, the proper alpha could be .05/20 = .003. While an unusual occurrence, it is essential that researchers report the number of models or different hypothesis tests conducted.

In summary, our results demonstrated that educational researchers are not adhering to the basic standards of reporting when conducting MLR analyses (cf. Hoyt et al., 2006; Schafer, 1991). Moreover, researchers have adopted the trend to report a range of alpha levels, for example, from .05 to .01, never specifying their *a priori* alpha level. Additionally, in the discussion of their results, researchers will often report their results using a variety of significance levels; variables will be referred to as statistically significant whether at the .10, .05, or .01 level instead of maintaining a single *a priori* standard. As a result, inappropriate and possibly damaging recommendations for practice may flow from these studies.

Recommendations for Reporting Regression Results

Our recommendations for what should be included when reporting regression results are as follows: researchers should (a) describe the variables and the conceptual sets of variables (if distinct sets exist), (b) indicate if the sets are ordered, (c) describe what technique was used to adjust the alpha level when multiple models are run, (d) explicitly state the *a priori* alpha level, and (e) describe the research conclusions reached. Authors should include the following in separate tables: (a) descriptive statistics (i.e., means, standard deviations, and sample sizes), (b) correlation matrices of all continuous variables, and (c) regression results that include: the overall *F* ratio for each test, R^2, adjusted R^2 when comparing regression equations with different numbers of predictors and when using small sample sizes, standard error of estimate if the dependent variable has a meaningful metric, the change in R^2 and the associated significance for HMLR, regression coefficients and their associated *t* tests. Furthermore, we recognize page and word restrictions commonly in place in journals for authors seeking publication. In this event, we suggest authors make supplementary material available on websites or through other avenues provided through publication.

Conclusion

Results of this study indicated that researchers commonly fail to adjust alpha levels when implementing HMLR techniques. Instead, it is much more common to see authors report results using a range of alpha levels from .10 to .01, which leads to inflated Type I and experimentwise error rates. Consequently, many of the results and recommendations reported in the studies in AERJ and EEPA based on HMLR/SMLR techniques may be misleading. Additionally, about half of the articles failed to properly document research findings to ensure transparency and replication (e.g., correlation matrices). The omission of basic summary statistics not only prevents replication, but it violates the principle of transparency. Regression is a powerful statistical analysis tool when used correctly. We call for a common convention in reporting and a return to basic scientific research standards.

References

Altman, D. (2000, January). *A review of experimentwise type I error: Implications for univariate post hoc and for multivariate testing*. Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX.

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*(6), 33 – 40.

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Bailey, T., & Weininger, E. (2002). Performance, graduation, and transfer of immigrants and natives in City University of New York community colleges. *Educational Evaluation and Policy Analysis*, *24*(4), 359-77.

Clements, D., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal*, *45*(2), 443-494.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Courville, T. & Thompson, B. (2001). Use of structure coefficients in published multiple regressions articles: B is not enough. *Educational and Psychological Measurement*, *61*(2), 229-248.

Creswell, J.W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Daniel, L. (2001). *Changes in the APA "Publication Manual": How the new fifth edition will affect research reporting in the social sciences*. Retrieved from ERIC database.

Desimone, L, Smith, T., Baker, D., & Uano, K. (2005). Assessing barriers to the reform of U.S. mathematics instruction from an international perspective. *American Educational Research Journal*, *42*(3), 501-535.

Elmore, P. B., & Woehlke, P. L. (1996, April). *Research methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978 to 1995.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Gravetter, F. J., & Wallnau, L. B. (2007). *Statistics for the behavioral sciences* (7th ed.). Belmont, CA: Wadsworth Publishing.

Hoyt, W. T., Leierer, S., & Millington, M. J. (2006). Analysis and interpretation of findings using multiple linear regression techniques. *Rehabilitation Counseling Bulletin, 49*(4), 223-233.

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal, 38*(3), 499-534.

Keppel, G., & Wickens, T.D. (2004). *Design and analysis: A researcher's handbook* (4th Ed.). Upper Saddle River, NJ: Prentice Hall.

Love, G. (1988, November). *Understanding experimentwise error probability*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Louisville, KY.

McInerney, D.M., Roche, L., McInerney, V., & Marsh, H. W. (1997). Cultural perspectives on school motivation: The relevance and application of goal theory. *American Educational Research Journal*, *34*(1), 207-236.

Mundfrom, D. J., Perrett, J. J., Schaffer, J., Piccone, A., Roozeboom, M. (2006). Bonferroni adjustments in tests for regression coefficients. *Multiple Linear Regression Viewpoints*, *32*(1), 1 – 6.

Olejnik, S., Li, J., Supattathum, S., & Huberty, C.J. (1997). Multiple testing and statistical power with modified Bonferroni procedures. *Journal of Educational and Behavioral Statistics, 22*, 389-406.

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F.J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, *45*(2), 365-397.

Roozeboom, M., Mundfrom, D. J., & Perrett, J. (2008, April). A modified Bonferroni procedure for multiple tests. Paper presented at the annual meeting of the American Educational Research Association, New York.

Salkind, N. J. (2007). *Statistics for people who think they hate statistics* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Schafer, W. (1991). Reporting hierarchical regression results. *Measurement and Evaluation in Counseling and Development*, *24*(3), 98-100. (ERIC Document Reproduction Service No. EJ438926) Retrieved from ERIC database.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Company.

Sidak, Z. (1967). Rectangular confidence region for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62, 626-633.

Stern, D., Dayton, C., Paik, I. W., & Weisberg, A. (1989). Benefits and costs of dropout prevention in a high school program combining academic and vocational education: Third-year results from replications of the California peninsula academies. *Educational Evaluation and Policy Analysis*, *11*(4), 405-416.

Thompson, B. (1995). Stepwise regression and stepwise discriminate analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, *55*, 525-534.

Thompson, B. (2007, October). *Standards in conducting and publishing research in education*. Paper presented at the annual meeting of the Mid-Western Educational Research Association, St. Louis, MO.

Zientek, L. R., & Thompson, B. (2009). Matrix summaries improve research reports: Secondary analyses using published literature. *Educational Researcher, 38*(5), 343-352.

Zwick, R., & Sklar, J. C. (2005). Predicting college grades and degree completion using high school grades and SAT scores: The role of student ethnicity and first language. *American Educational Research Journal*, *42*(3), 439-464.

Send correspondence to:
Winona Burt Vesey
University of Houston, Clear Lake
Email: Vesey@UHCL.edu

What Makes a Winning Baseball Team and What Makes a Playoff Team?

Javier Lopez
New Mexico State University

Daniel J. Mundfrom
Eastern Kentucky University

Jay R. Schaffer
University of Northern Colorado

Team statistics from all 30 teams in Major League Baseball were analyzed to determine what makes a winning baseball team and what makes a playoff team. Thirty-two statistics in all, including batting, fielding, and pitching statistics, were used in a multiple linear regression and discriminant analyses. The regression procedure was to determine what makes a winning team, while the discriminant analyses were used to see what makes a playoff team. On-base percentage plus slugging (OPS) and earned run average (ERA) were fit on wins in the regression model with an $R^2 = 0.83$. The discriminant analyses distinguished different statistics in the National and American Leagues for discriminating between playoff and non-playoff teams. ERA and OPS were discriminating factors in the National League, while saves, on-base percentage, and earned run average were factors in the American League.

Every year, 30 Major League Baseball (MLB) teams strive to make the playoffs and World Series. Sixteen teams are from the National League and 14 are from the American League. Each league is split into three divisions, the East, Central, and West. In order to make the playoffs, a team must either win their division or win the wild card spot. That is, at the end of the 162 regular season games, they must either have the most wins in their division or have the most wins among the teams that did not win their division. Essentially, in order to be a successful team in MLB, you must win. General Managers (GM) in MLB can be fired because they are unable to build a winning baseball team. The question that every GM in MLB should be asking himself as they build their team each year is, "*What makes a winning baseball team*?"

Talsma (1999) answers this question in a simple and forthright manner. In his study, he determined that run differential (RDIFF = runs scored – runs allowed) is the most important statistic in determining how many wins a baseball team will have. The higher the run differential, the more wins a team will have. But if we are trying to build an MLB team, we cannot directly sign "runs" to our team. That is, we can only sign players. For instance, if we want to sign a free agent such as Manny Ramirez, there are many statistics that we can look at to determine his worth. However, the question may well be "which statistics are the ones that appear to contribute to winning and runs scored?" Should a GM sign him, and if so, on what should he base his decision if he cannot determine how many runs he is worth? In this study, we take 32 team statistics into account and attempt to find which ones best determine the wins a team will have in a given year. We do this by using past research to help determine which statistics should be used to fit a multiple linear regression model for all data collected between 1995 - 2009.

The second objective of this study is to determine which statistics are best for predicting which teams will make the playoffs and which teams will not. There have been some models to predict divisional winners. Barry and Hartigan (1993) used Markov chain sampling to determine future strengths of teams and future outcomes of games to predict divisional winners. This research was done before MLB split into 3 divisions and the addition of a wild card team in the playoffs. Barry and Hartigan only had a total of 4 playoff teams each year when this research was done, as opposed to the present day when we now have 8 playoff teams each year. For this reason we only use data after 1994 when the new wild card rule took effect. In this study, we use discriminant analysis to find out which statistics discriminate between teams that made and missed the playoffs.

Review of Literature

Talsma (1999) used a simple linear regression analysis to determine what made a winning baseball team. He initially explored the relationships between wins and hits, homeruns, runs, and opponents' runs. All four of these statistics are offensive statistics. After finding that the relationship between those four variables and winning was weak, he decided that maybe offense was not the only way a team could win. He found that run differential (RDIFF = runs scored – runs allowed) is the most important statistic in determining how many wins a baseball team will have. The higher the RDIFF, the more wins a team will have. His regression equation was:

$$\widehat{Wins} = 0.088 * RDIFF + 80.964$$

Talsma's results are vital information for use in this study. If run differential is the best predictor for wins, we would then need to find out which player and team statistics best predict runs scored and runs allowed.

There is a plethora of statistics that we can examine in baseball to evaluate performance. Batting average (BA), on-base percentage (OBP), home runs (HR), earned run average (ERA), strike outs (SO), on-base percentage plus slugging(OPS), and fielding percentage (Fld%) are just a few examples. But which ones are the most valuable?

A study by Cover and Keilers (1977) used a batter's cumulative statistics to determine which batters were the best of all time. They used at-bats, walks, singles, doubles, triples, homeruns, and outs to determine what they called an offensive earned run average (OERA). Cover and Keilers defined OERA as the number of earned runs per game that a player would score if he batted in all nine positions in the line-up. After getting the OERA for all players, they then could be compared to each other to determine which players were actually the best batters in the league. The higher the OERA, which means they would score more runs, the better the batter.

Another study by Koop (2002) used an output aggregator to compare players. This method weighed multiple batting statistics to output a value between 0 and 1, where 1 would be the best player and 0 the worst. These two methods of determining a player's performance were very different, but the consensus between these studies is that there were multiple statistics that contributed to the performance of a player.

There has been little published in regard to what statistics best predict playoff teams. As stated earlier, Barry and Hartigan (1993) used Markov chain sampling to determine future strengths of teams and future outcomes of games to predict divisional winners. They used information based on strength of teams to determine divisional winners. Our objective in the current study is to determine which baseball statistics may be able to be used to predict which teams will make the playoffs and which teams will not.

Table 1. Offensive and Defensive Statistics.

Offensive Statistics		
Abbreviation	Statistic	Mathematical Definition
PA	Plate Appearances	
AB	At-Bats	
H	Offensive Hits	
2B	Doubles Hit	
3B	Triples Hit	
HR	Home Runs Hit	
SB	Stolen Bases	
CS	Caught Stealing	
BB	Bases on Balls	
SO	Strike Outs	
BA	Batting Average	$\frac{\text{Hits}}{\text{AB}}$
OBP	On Base Percentage	$\frac{\text{(H+BB+HBP)}}{\text{(At-Bats+BB+HBP+SF)}}$
SLG	Slugging Percentage	$\frac{\text{(1B+2*2B+3*3B+4*HR)}}{\text{AB}}$
OPS	On Base Plus Slugging	OBP+SLG
E	Errors Committed	
DP	Double Plays Turned	
BatAge	Batter's Age	
Defensive Statistics		
Fld%	Fielding Percentage	$\frac{\text{(Putouts+Assists)}}{\text{(Putouts+Assists+Errors)}}$
ERA	Earned Run Average`	$\frac{\text{(Earned Runs*9)}}{\text{IP}}$
CG	Complete Games	
SHO	Shutouts	
SV	Saves	
IP	Innings Pitched	
H	Hits Allowed	
ER	Earned Runs Allowed	
HR	Home Runs Allowed	
BB	Bases on Balls Allowed	
SO	Strike Outs	
WHIP	Walks & Hits Per Innings Pitched	$\frac{\text{(BB+H)}}{\text{IP}}$
SO/9	Strike Outs Per 9 Innings	$\frac{\text{(9*Strikeouts)}}{\text{IP}}$
HR/9	Home Runs Per 9 Innings	$\frac{\text{(9*HR)}}{\text{IP}}$
PitchAge	Pitchers Age	

Methods

The following 32 statistics from Table 1 were collected for all teams' from the 1995 - 2009 seasons (Baseball-Reference.com, 2010). The data used in these analyses were the final season-ending values of each of these statistics for each team collectively for each of these 15 seasons.

Data were analyzed using SAS version 9.1.3 software using PROC GLM, PROC REG, PROC DISCRIM, and PROC STEPDISC. Data were initially explored graphically and, subsequent to model fitting, residual analysis was conducted. All 32 statistics for both American and National League teams were initially used in the multiple linear regression analysis. After eliminating variables to remove multicollinearity from the data, variables with large p-values were taken out of the model one at a time using the method described below. The statistical significance level for the regression was defined for $p < 0.05$.

Beginning with the full model that contained all 32 variables, an examination of the pairwise correlations and variance inflation factors (VIF) identified several pairs or groups of variables that were collinear (i.e., pairwise correlations > .70 and VIFs > 10). By removing variables one at a time and re-running the model with one fewer variable at each step, the multicollinearity was successfully removed with the elimination of 12 variables, leaving 20 predictor variables in the model with only one having a VIF slightly larger than 10 (R^2 = .889, adjusted R^2 = .885). The remaining predictors were plate appearances (PA), doubles (DB), triples (TR), stolen bases (SB), caught stealing (CS), offensive strikeouts (OSO), batting average (BA), on base plus slugging (OPS), errors (E), double plays, (DP), batter's age (BatAge), earned run average (ERA), complete games (CG), shutouts (SHO), saves (SV), hits (H), home runs (HR), walks (BB), defensive strikeouts (SO), and pitcher's age (PitchAge). With this model, the unique contribution of each variable to the explanation of the variance in the number of wins was examined (using Type III Sums of Squares in SAS) and the variable that made the smallest unique contribution (i.e., provided its p-value was less than .05) was dropped from the model.

The remaining 20 variables were re-analyzed and the variable with the smallest, non-significant unique contribution to the model was dropped. This process was continued, at each stage dropping only the one variable that made the smallest, non-significant, contribution to the explanation of the variation in the number of wins, until only the variables that made significant unique contributions remained. The final model, a multiple linear regression of OPS and ERA on Wins was fit.

Stepwise discriminant analyses were performed on the 32 statistics to determine which statistics discriminate between teams that made and missed the playoffs for both the National League and American Leagues separately. Separate analyses were used because the American League uses a designated hitter as opposed to the National League, which does not. The discriminant procedure was run three times for each league. Each time the significance level (i.e., statistical significance level for entry and significance level for removal were set to be equal) was set to be different. The three different significance levels used were $p < 0.1$, $p < 0.05$, and $p < 0.01$. This process was used to determine whether more predictors would lower the total probability of misclassification (TPM). If the TPM stayed the same when there were fewer predictors, the extra predictors were deemed unnecessary because they were not contributing to better predicting classification. The jack-knife (cross-validation) method was performed in SAS to estimate the TPM. Both linear and quadratic models for the jack-knife method were evaluated to determine which one performed better based on the TPM. Lower TPM was deemed to be better than higher.

Results

A model to predict the number of wins based on OPS and ERA was considered. Because these two variables were the only ones that survived the variable-identification process, it could be surmised that these two statistics are in some way representing offensive prowess (i.e, OPS) and defensive prowess (i.e., ERA). For a team to have enough wins at the end of the season to make the playoffs, they must be good both offensively and defensively so it made sense that these two variables would be good predictors of the number of games a team wins during the season. Some teams may win a lot of games because they have very potent offenses, and others may win fewer games because they are less productive at the plate. Also, some teams may win a lot of games because they have good pitching and defense, and others may win less often because they do not. It is also possible that some teams that may be only "average" on both offense and defense may win a lot of games because they are able to score just a few more runs than they

give up on enough occasions during the season so that they win a lot of games. Consequently, it seemed prudent to consider an interaction between these two variables as an addition to this model.

A model was subsequently fit that contained the two variables identified earlier, OPS and ERA, and their interaction to predict the number of wins. The results of this analysis showed that the interaction between OPS and ERA was not statistically significant ($p = 0.8991$. Consequently, it was dropped from the model. The final model from the multiple linear regression, containing only OPS and ERA as predictors, provided a good fit ($F = 1044$, $p < 0.0001$, $R^2 = 0.83$) and yielded the following estimated equation:

$$\widehat{Wins} = -10.45 + 213.62 * OPS - 15.98 * ERA$$

with standard errors for the intercept = 5.09, OPS = 6.5, and ERA = 0.45.

Results for the discriminant analysis varied with each league. Table 2 shows that the discriminant analysis did a credible job of classifying the American League teams properly. We see that 80% of the teams that made the playoffs were actually classified as making the playoffs and 89.12% of the teams that missed the playoffs were actually classified as missing the playoffs. Table 3 indicates that the discriminant analysis did not work quite as well in the National League as compared to the American League in terms of correctly classifying the teams that made the playoffs (66.3%). However, for the National League teams, 90.96% of the teams that missed the playoffs were classified correctly; just slightly more than in the American League.

Table 2. American League Classification Results.

American League	Classified as Type		
From Type	Made Playoffs	Missed Playoffs	Total
Made Playoffs	48	12	60
	80.00%	20.00%	100%
Missed Playoffs	16	131	147
	10.88%	89.12%	100%
Total	64	143	207
	30.92%	69.08%	100%

Table 3. National League Classification Results.

National League			
From Type	Made Playoffs	Missed Playoffs	Total
Made Playoffs	38	22	60
	63.33%	36.67%	100%
Missed Playoffs	16	161	177
	9.04%	90.96%	100%
Total	54	183	237
	22.78%	77.22%	100%

In Table 4, we see that for all cases, the linear function outperformed the quadratic. That is, the TPM was lower for the linear function in each case. The TPM stayed the same within league and model even when the stepwise significance level was decreased. The only change concerned the number of predictor variables identified as important discriminators.

Discussion

From our results, we found that statistics indeed can help us determine what makes a winning baseball team. OPS and ERA accounted for 82.57% of the variation in determining wins for a Major League baseball team. This result is consistent with what Talsma (1999) found in that both offense and defense are needed factors to succeed. Within our model, OPS is used as the offensive statistic and ERA is used as the defensive statistic. This fact may provide assistance to General Managers for potentially basing their decisions for personnel drafting and trading, as well as signing free agents. This is not to say that other statistics, such as stolen bases or strikeouts, do not contribute to winning; as all offense contributes to scoring runs and winning. However, when there are millions of dollars at stake, as there is in Major League Baseball, it is suggested by this analysis that General Manager's might benefit from looking at these two statistics before any others when making a decision to pursue a prospective player.

In the second set of analyses, we saw in both the National and American Leagues that even as the level of significance for selecting discriminating variables went down, the TPM remained the same. At $p < 0.1$, the TPM was the same at $p < 0.01$. What does change, is the number of discriminants, which even with this change, we can see that there is no change in the number of incorrectly classified teams. This result brought us to the conclusion that we could drop the extra variables because they were not contributing to the classification.

The next thing to notice in the results was that there were two different sets of statistics that discriminated between playoff teams and non-playoff teams. The National League followed our regression model; where the OPS and ERA were the best predictors in determining whether or not a team made the playoffs. However, the American League discriminant analysis indicated that saves (SV), on-base percentage (OBP), and earned run average (ERA) were the best predictors. The question then became, "Why were the results different between the two leagues?"

Table 4. Results from the Discriminant Analysis

Discriminant Analysis Cross-Validation Results				
League	**Model**	**TPM**	**Stepwise Sig.**	**Predictors**
NL	Linear	0.1595	p≤0.1	ERA,OPS,2B,Fld%
NL	Quad	0.2337	p≤0.1	ERA,OPS,2B,Fld%
NL	Linear	0.1595	p≤0.05	ERA, OPS, 2B
NL	Quad	0.2337	p≤0.05	ERA, OPS, 2B
NLl	Linear	0.1595	p≤0.01	ERA, OPS
NL	Quad	0.2337	p≤0.01	ERA, OPS
AL	Linear	0.1349	p≤0.1	SV,OBP,ERA,DP
AL	Quad	0.2676	p≤0.1	SV,OBP,ERA,DP
AL	Linear	0.1349	p≤0.05	SV,OBP,ERA
AL	Quad	0.2676	p≤0.05	SV,OBP,ERA
AL	Linear	0.1349	p≤0.01	SV,OBP,ERA
AL	Quad	0.2676	p≤0.01	SV,OBP,ERA

Let us first examine the difference between OBP and OPS. OBP is the percentage of times a batter gets on base by walking, hitting safely, or being hit by a pitch. OPS is equal to OBP plus slugging percentage (SLG%). Slugging percentage is the tell-tale statistic that takes into account a batter's power. A batter who hits mostly singles will have a lower SLG% than a batter who hits more doubles, triples, and home runs. These results indicate that OPS is more important in the National League and OBP is more important in the American League.

One reason there may be a difference is because the American League uses a designated hitter (DH) and the National League does not. That is, in the American League, the pitchers do not hit. The designated hitter is used in the lineup instead. Whereas, in the National League, the pitchers bat for themselves until a substitution is made that removes the pitcher from the game. Pitchers, in general, are the least proficient batters in the league. The use of the DH in the American League contributes to why most ERAs are lower in the National League. That is, National League pitchers face a lesser quality batter 1 out of every 9 batters in the lineup. It could be that it is for this reason OPS is more important in the National League.

Let us consider an example. We will assume that we have two outs and a man on first base. The 8th man in the lineup, who has a low slugging percentage, is up to bat and the pitcher is on deck to bat next. If the 8th man hits a single, the runner advances to 2nd base only. The next batter is the pitcher and in most cases, he is assumed to be an easy out. So if he strikes out or fails to hit safely, then the inning is over and 2 men are left on base. If, however, the 8th man in the order happens to be more of a power hitter and he hits a double, the runner may score a run. Then, once again, the pitcher, who is the next batter, strikes out or does not hit safely to end the inning. The difference being that a run scores with a double instead of a single. This scenario occurs daily in the National League.

In the American League, this same scenario does not happen as often because instead of having the pitcher bat, the league uses a DH who in most cases is a better batter. In the same hypothetical situation, if the 8th batter only hits a single, we would again have a man on first and second. Only this time, in the American League, the pitcher or "easy out" is not up to bat. Instead, we would have a higher quality batter up who even if he only singles, he drives in a run; whereas in the National League, the pitcher is much more likely to make an out. The OPS statistic takes into account a batter's power whereas OBP does not. We can see from the previous simple example, why power is an important aspect and, therefore, OPS was more important in the National League in the results.

Our results also show that saves are a significant factor in the American League and not in the National League. This result could also be due to the quality of batters that a pitcher must face in the American League. However, our previous example does not quite apply to this situation. Saves are credited to a pitcher when:

1. He is the finishing pitcher in a game won by his team;
2. He is not the winning pitcher;
3. He is credited with at least ⅓ of an inning pitched; and
4. He satisfies one of the following conditions:
   a. He enters the game with a lead of no more than three runs and pitches for at least one inning.
   b. He enters the game, regardless of the count, with the potential tying run either on base, at bat, or on deck.
   c. He pitches for at least three innings.

In the National League, a pinch hitter (PH) can bat for a pitcher only if the pitcher is substituted out of the game. After the PH hits, a different pitcher must come into the game to pitch. When it next becomes the pitcher's turn to bat, he must bat, or another PH can hit for him, and he will be substituted out. Once a pitcher is taken out of the game, he cannot come back in. The same goes for the PH. Any one player can only pinch hit once in a game (i.e., unless his turn comes around again in the same inning or he enters the game as a position player after completing his pinch hit). In the late innings of a game, the pitchers in the National League are usually substituted for to try and maximize runs scored. By having a better batter pinch hit for the pitcher, you eliminate what is often an "easy out" in the lineup.

Most likely, pitchers that are in save situations, will not have an easy out, but the quality of hitter may still differ from the National League to the American League. The DH in the American League hits every time around in the lineup. A pinch hitter in the National League sits most of the game until a substitution is needed. When a substitution is made, the pinch hitter usually goes up to bat one time and then his job is over for the day. For this reason, it is probably reasonable to say that it is more difficult to be a pinch hitter than a designated hitter. We may also guess that it is easier to get most PHs out than it is to get DHs out. A pinch hitter has sat most of the game and only has one chance to hit in a game whereas the designated hitter has probably already hit three or four times before a closing pitcher comes in to attempt a save. Since it is more difficult to get a designated hitter out, it becomes more crucial to have a good closer to get a save in the American League than in the National League. Consequently, the use of the DH in the American League versus the PH in the National League could be contributing to why our results show that SVs are more important in the American League than in the National League.

Conclusion

In conclusion, we can see that certain statistics are vital in determining wins for a baseball team and determining which teams will and will not make the playoffs. We can predict wins based on OPS and ERA. We can also determine that OPS and ERA are significant factors for predicting if a team makes the playoffs in the National League, whereas Saves, OBP, and ERA are significant factors in predicting whether a team makes the playoffs in the American League. Although our discriminant analysis procedure did predict 80% of teams correctly to make the playoffs in the American League, it only predicted 63.33% correctly for the National League. This outcome is not bad and is certainly better than simply guessing or flipping a coin. Perhaps future research can identify other variables that can improve this percentage.

References

Barry, D., & Hartigan, J. (1993). Choice models for predicting divisional winners in Major League Baseball. *Journal of the American Statistical Association, 88*(423), 766-774.

Baseball-Reference.com. (2010). *Major League Baseball statistics and history*. Retrieved from http://www.baseball-reference.com.

Cover, T., & Keilers C. (1977). An offensive earned-run average of baseball. *Operations Research, 25*(5), 729-740.

Koop, G. (2002). Comparing the performance of baseball players: A multiple-output approach. *Journal of the American Statistical Association, 97*(459), 710-721.

Talsma, G. (1999). Data analysis and baseball. *Mathematics Teacher, 92*(8), 738-742.

Send correspondence to:
Javier Lopez
New Mexico State University
Email: javmlopez@yahoo.com

A Note on Cost-Benefit Analysis

David A. Walker
Northern Illinois University

Following the framework presented by Leech and Onwuegbuzie (2004), results from an heuristic example added to the very limited scholarly literature in the area of cost-benefit analysis, and also served as a potential template related to the relative ease of implementation of some of cost-benefits' components that have shown initial properties of augmenting results affiliated with correlational designs and/or program evaluation.

In the research literaure for the social sciences, the idea of understanding and/or reviewing the cost effectiveness and also the benefit(s) derived from an intervention or program activity is an emergent concept with limited scholarship devoted to it. Of the research in this domain, programmatic overall cost analysis has been presented in the literature via estimated measures (King, 1994; Odden, 2000). Program cost effectiveness measured through meta-analysis has been proffered by Borman, Hewes, Overman, and Brown (2003) and Yeh (2008). A correlational study combining both cost (i.e., program) and benefit (i.e., increased student test scores) was conducted by Quinn, Van Mondfrans, and Worthen (1984). Barnett (1985) offered a cost (i.e., program) and benfit (i.e., social investment of a program) analysis of a preschool program.

Finding literature and guidelines that amalgamate both known, direct costs of a program and said program's tangible benefit(s), coupled with other measures such as effect sizes and practical effects of an intervention and/or program activity, is arduous. Two seminal sources in the literature that looked at both cost and benefit in terms of the effectiveness of intervention results were offered by Levin (1983) and Levin and McEwan (2001). These authors develed into this area by providing guidance related to how reviewing costs of an intervention given the outcome(s) derived may provide new and/or additional information pertaining to an intervention's effect. Related to the Levin and Levin and McEwan works, Leech and Onwuegbuzie (2004) coined the term 'economic significance' as the "economic value of the effect of an intervention" (p. 185). Their work yielded a typology of five economic-related indices used to measure cost in its various forms: effectiveness, benefit, utility, feasibility, and sensitivity. A major component of their indices was to incorporate the cost, either direct or estimated, along with the effect, typically measured as either *post-hoc* raw differences or standardized differences (i.e., effect sizes). Finally, along this same line of thought in the field of psychology, Wittmann (2004; 2007) proposed the use of *a priori* break-even effect sizes (i.e., standardized differences) to compare with known effects from the literature resultant from meta-analysis to assist in estimating a return on investment of an intervention. Wittman's work was an extension of earlier social science cost-benefit research completed on economic impact from workforce productivity studies (Schmidt, Hunter, & Pearlman, 1982).

Context

A school-university partnership between Northern Illinois University and Rockford, Illinois Public School District 205 has been in existence for the past decade. The focal point of this comprehensive partnership is to enhance student learning. As part of a partnership evaluation, a correctional design was employed to measure the relationships and effect sizes obtained from programming initiatives concerning student learning in the content areas of mathematics and reading at two school sites: a P-5 elementary school and a 6-8 middle school (MS). The concept of "student learning" for schools in the partnership was measured via data attained from their performance in mathematics and reading on the Illinois Standards Achievement Test (ISAT).

Cost-Benefit Design

Following the framework of a cost-benefit analysis presented by Leech and Onwuegbuzie (2004), two indices were implemented. The first index related to the cost per level of effectiveness (CE), where C was the direct cost of the program and E was the practical effect measured in terms of the raw difference in testing points (i.e., ISAT mathematics and reading). Note: the practical effect of the effect size measure in terms of testing points gained was based on average standard deviations from sample data trends found for ISAT mathematics elementary = 28.04, mathematics MS = 27.83, reading elementary = 27.51, and reading MS = 24.27 (Consortium on Chicago School Research, 2007).

$$CE = C / E \quad (1)$$

A second index measured the maximum effectiveness of a program per level cost (MCE), where cost (C) and effect (E) were continued in their use, but the idea of desired expernditure (D) was added; theoretically by the district as a form of sustainability after the initial program:

$$MCE = (E / C) \times D \quad (2)$$

Heuristic Example

An activity of the partnership to enhance student learning, National Board for Professional Teaching Standards (NBPTS) certification, lent itself to a cost-benefit analysis due to the ability to tally direct costs and relational effects for this endeavor. A NBPTS certification program was initiated at the P-5 and the middle school as a sustained mentoring and professional development plan focused on teacher instructional and curricular development, specifically in the subject areas of mathematics and reading, as well as a focus on the association between the practices of teaching and learning.

Effect Sizes

Figure 1 shows the average effect size in reading and mathematics for the P-5 and the middle school. The effect size used in this study was Cohen's d and employed benchmarks set at .20, .50, and .80 that represented small, medium, and large effects, respectively (Cohen, 1988). Results from research conducted by Lipsey and Wilson (1993) corroborated Cohen's .50 cut-point for a medium effect by finding, via meta-analysis, that the mean and median effects from over 300 studies were established at .50 and .47, correspondingly. Recently, Sawilowsky (2009) found in a review of the literature that the aforementioned effect size cut-points of .20, .50, and .80 could be conceptualized also as small, medium, and large; though as inclusive members of a more expanded d-based benchmark scheme. Throughout the duration of measuring the relational effects of the NBPTS initiative, certainly other factors in addition to it accounted for a percentage of the effect size results depicted in Figure 1. By comparative measures with the Cohen benchmarks and/or the Lipsey and Wilson and Sawilowsky values, both of the schools showed medium to large effect sizes in reading and mathematics scores, where the relational effect in mathematics, for instance at the MS, approximated a large effect contrasted against known criteria from the literature.

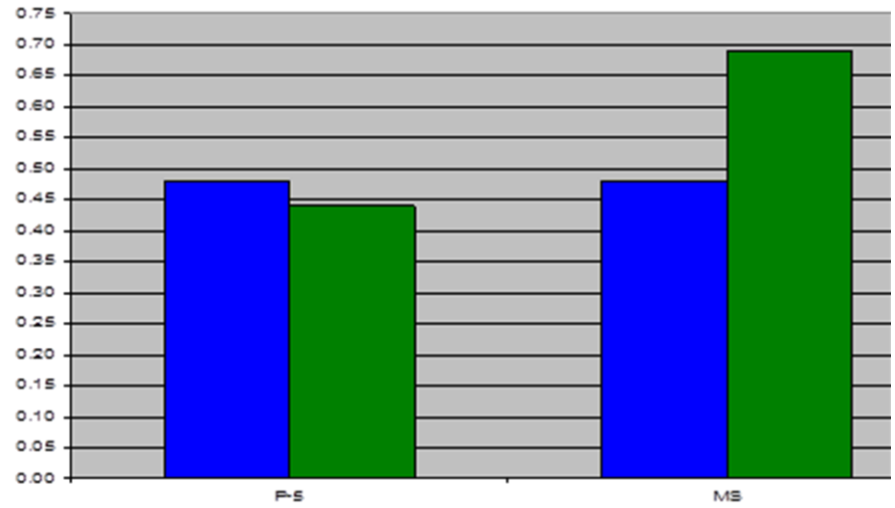

Figure 1. Average effect size. (Reading Scores in Blue; Mathematics Scores in Green)

Practical Effects

In conjunction with the effect size results, Figure 2 shows that in a practical sense for the amount of testing points gained, there was an increase at both schools. In fact, there was quite a substantial increase given that the ISAT test varies from a minimum of 120 to a maximum ranging from 340 to 411 based on grade level (Consortium on Chicago School Research, 2007).

Cost per Level of Effectiveness

The total direct cost for each of the 14 NBPTS certified teachers at the two schools was $12,571.43 with 7 teachers serving in both the P-5 and the MS. Figure 3 displays the cost per level of effectiveness. As examples, in the P-5 school, the relational effect in reading from the presence of 7 NBPTS certified teachers was an average gain of 13.20 ISAT points or a cost of $952.38 per each one point mean difference in reading scores (i.e., $12,571.43 / 13.20). For the NBPTS MS mathematics, the cost was $651.37 per each one point mean difference in math scores (i.e., $12,571.43 / 19.30).

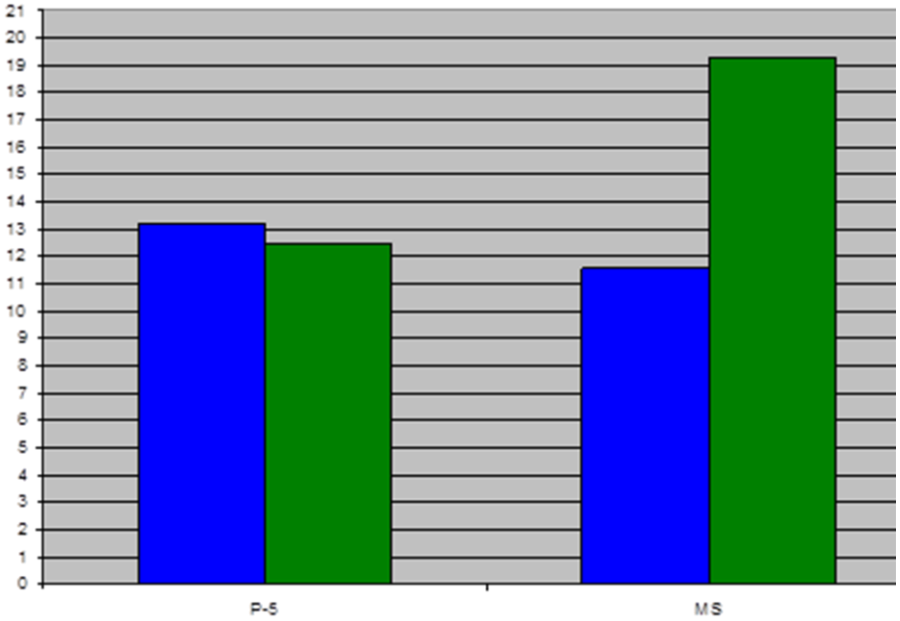

Figure 2. Average practical effect in testing points.
(Reading Scores in Blue; Mathematics Scores in Green)

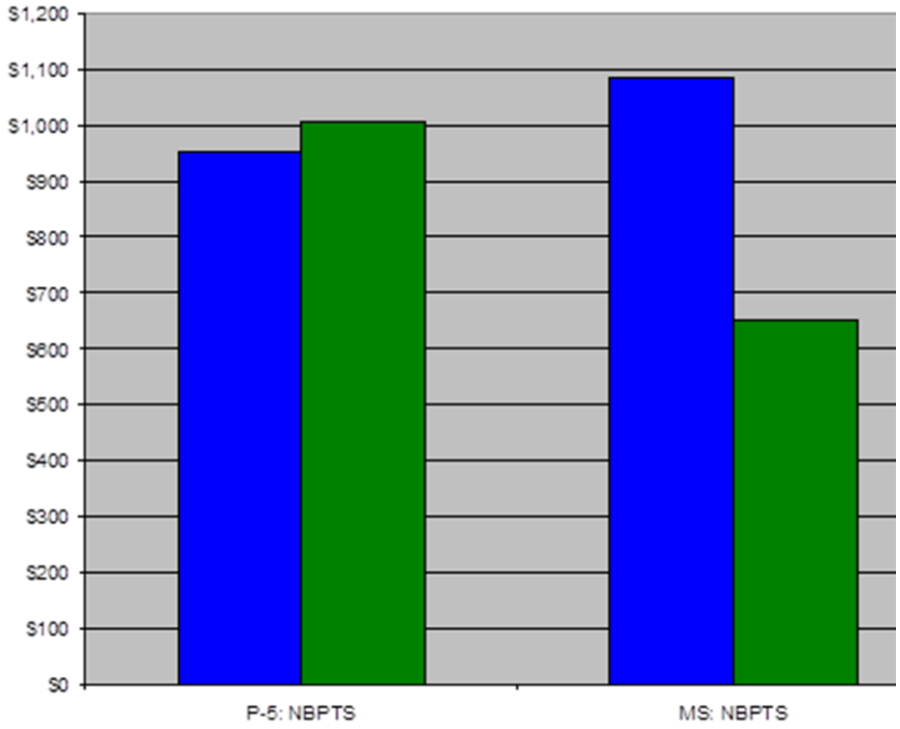

Figure 3. Cost per one point mean difference in test scores..
(Reading Scores in Blue; Mathematics Scores in Green)

Maximum Effectiveness of an Intervention per Level Cost

As a means to look at the potential sustainability of the NBPTS programs within the district, a maximum effectiveness of an intervention per level cost analysis was conducted to correspond with the previous findings. Given the medium to large effect sizes and the relatively low costs per one point mean difference in reading and mathematics scores for both school settings and each program, the question of effectiveness sustainability emerges. Thus, if after reviewing the positive, previously-mention results, the district were to allocate $50,000 a year for the continuation of the NBPTS programs (i.e., apportioning $25,000 to NBPTS x 2 schools).

Figure 4 indicates that they could predict quite large maximum mean test score increases. For instance at the MS NBPTS, increases of approximately 23 (i.e., (11.57 / $12,571.43) x $25,000) and 38 testing points for reading and mathematics, respectively may be predicted in the near future by continuing with the program of National Board certification of additional teachers within the school, however; with the caveat of having a reasonably similar student body and teacher ability as was accompanied with past results.

Conclusion

An importance of this research note is that it adds to the very limited scholarly literature in the area of cost-benefit analysis for the social sciences and serves as a potential template related to the relative ease of implementation of some of cost-benefits' components that have shown initial properties of augmenting results affiliated with, for example, correlational designs or program evaluation.

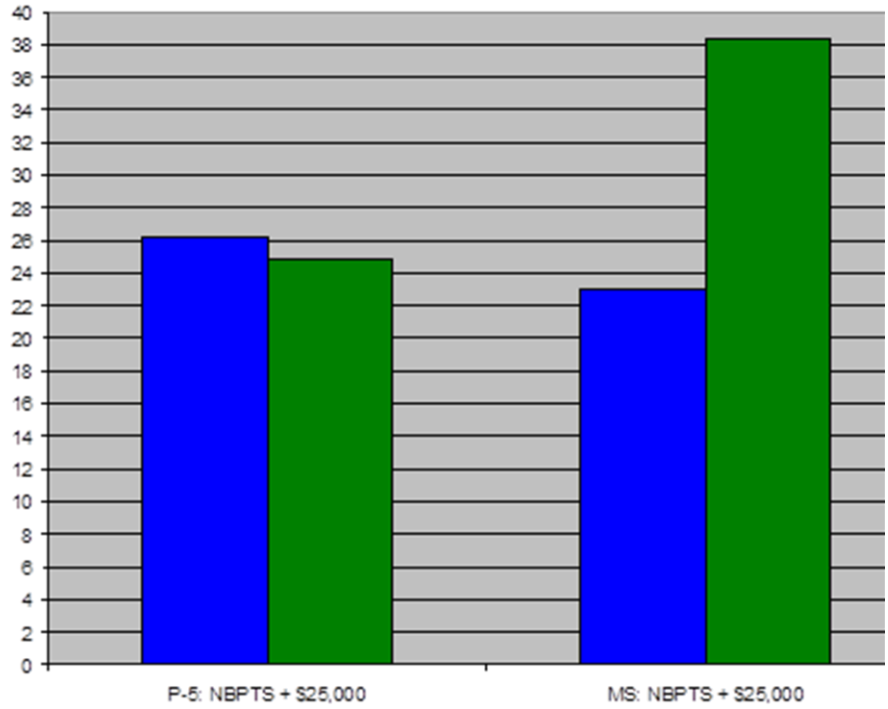

Figure 4. Maximum in mean test score gain.
(Reading Scores in Blue; Mathematics Scores in Green)

References

Barnett, W. S. (1985). Benefit-cost analysis of the Perry preschool program and its policy implications. *Educational Evaluation and Policy Analysis, 7*, 333-342.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73*, 125-230.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Consortium on Chicago School Research (2007). *2006 ISAT reading and math scores in Chicago and the rest of the state*. Retrieved from www-news.uchicago.edu/releases/07/pdf/070621.consortium.pdf

King, J. A. (1994). Meeting the educational needs of at-risk students: A cost analysis of three models. *Educational Evaluation and Policy Analysis, 16*, 1-19.

Leech, N. L., & Onwuegbuzie, A. J. (2004). A proposed fourth measure of significance: The role of economic significance in educational research. *Evaluation and Research in Education, 18*, 179-198.

Levin, H. M. (1983). *Cost effectiveness: A primer*. Newbury Park, CA: Sage.

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181-1209

Odden, A. (2000). The costs of sustaining educational change through comprehensive school reform. *Phi Delta Kappan, 81*, 433-438.

Quinn, B., Van Mondfrans, A., Worthen, B. R. (1984). Cost-effectiveness of two math programs as moderated by pupil SES. *Educational Evaluation and Policy Analysis, 6*, 39-52.

Sawilowsky, S. S. (2009). Very large and huge effect sizes. *Journal of Modern Applied Statistical Methods, 8*, 597-599.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1982). Assessing the economic impact of personnel programs on workforce productivity. *Personnel Psychology, 35*, 333-347.

Wittmann, W. W. (2004, November). *Practical significance as the economic impact of effect sizes*. Paper presented at the Evaluation Conference, Atlanta, GA.

Wittmann, W. W. (2007, May). *Meta-analysis, effect sizes, investment decisions, and economical implications*. Paper presented at the Seventh Annual International Campbell Collaboration Colloquium, London, England.

Yeh, S. S. (2008). The cost-effectiveness of comprehensive school reform and rapid assessment. *Education Policy Analysis Archives, 16*(13). Retrieved from http://epaa.asu.edu/epaa/v16n13

Send correspondence to:
David A. Walker
Northern Illinois University
Email: dawalker@niu.edu

Paying Attention to the Default Reference Category in Several SPSS Statistics Procedures: An Example of Coding Reversal

Hongwei Yang
University of Kentucky

This paper reviews two common approaches to handling categorical predictors in regression analysis and how they are implemented in several Statistical Package for the Social Sciences (SPSS) procedures. Through this review, the aim is to revisit a warning from the SPSS literature regarding dummy coding for categorical predictors with just two classes. In this problem, the original coding of a two-category dummy variable is reversed by the software program without any alert and such a change is likely to result in an incorrect interpretation of the regression coefficient estimate of the corresponding categorical predictor. Further, the paper discusses the generalizability of this conclusion to several other statistics programs: Statistical Analysis System (SAS), JMP, and STATA.

In regression analysis, categorical predictors are very commonly seen. A categorical predictor is usually presented as a single variable in one of two formats: 1) A string variable, and 2) a numeric variable. Here, the most commonly used two-class categorical predictor gender is used as an example. When it is presented as a string variable, its values are in the form of letters: male and female, for example. On the other hand, when it is presented as a numeric variable, its values are usually in the form of allocated (numeric) codes (Kutner, Nachtsheim, & Neter, 2004). The allocated codes are arbitrarily selected, and could be any sets of numbers. For the two values of the gender variable, possible sets of codes are (-1) and (-2), 0 and 1, 2 and 3, 99 and 100, 999 and 1000, etc. In fact, as long as two numbers are distinct from each other, they could be allocated to represent the two different values of the gender variable.

Although a categorical predictor could be presented as either a single string variable or a single numeric variable, in many cases neither format could be directly used as an input in a regression model without properly performing further coding of those (string/numeric) categories. For a string variable, the underlying reason is obvious because its values are non-numeric. For the other approach where a categorical predictor is allocated arbitrarily selected numeric codes, the problem is that those codes define a metric for the categories of the predictor that may not be reasonable. This is because the spacing of categories indicated by those allocated codes may not be in accord with the reality (Kutner et al., 2004). With that described, in order for a categorical predictor to be used as a direct input in a regression model, it should be properly recoded. And this process is known as dummy coding.

When dummy-coding a categorical predictor, the most common scheme is the 0-1 coding. Within the 0-1 scheme, for a categorical variable with c categories, a total of (c - 1) dummy/indicator variables are needed with each one representing one category of the predictor. The only ignored category for which there is no dummy variable created is the reference level or base level. All other coded categories are compared with the reference category in terms of the average change in the outcome when it moves from the reference category to that particular non-reference category. Unlike a single categorical predictor in either string or numeric format, its dummy-coded indicator variables can be used as direct inputs for a regression model.

The way the 0-1 coding scheme is implemented in the software Statistical Package for the Social Sciences (SPSS) varies from one regression procedure to another. Related procedures can generally be classified into two categories: 1) Those that are not capable of automatically creating dummy variables, and 2) those that have this particular capability. The REGRESSION procedure is an example from the first category, whereas the GENLIN, LOGISTIC REGRESSION, NOMREG procedures, etc. belong to the second category.

The paper focuses on the use of a categorical predictor in those SPSS procedures that can automatically perform dummy coding. A procedure in this category has to be informed of the categorical nature of a predictor before it automatically recodes the variable into one or more dummy variables in the background. For a categorical predictor presented in a string format, this is not an issue at all because its values are non-numeric. SPSS identifies all such non-numeric variables as categorical without having to

be told. However, when a categorical predictor is presented in numeric format using allocated codes, things could become more complex.

A particular confusing case is when the allocated codes for a two-category binary predictor are selected to be identical to its 0-1 dummy codes and, at the same time, the predictor is still specified as categorical in an SPSS procedure with an automatic dummy coding capability. Norusis (2003) provides a warning on this issue, saying that there is nothing to be gained by declaring such a predictor as categorical and since such a specification prevents the original coding (already in 0-1 format) from being preserved, it is not a recommended practice. However, some, like Field (2009), think that declaring a two-class binary predictor as categorical or non-categorical in a SPSS procedure should not make any difference when the variable is already coded as 0 and 1, which clearly violates this warning. Considering such an unfortunate fact, this paper elaborates on this warning and uses an example to demonstrate it with the hope of helping practitioner comprehension related to the issue.

As is known, numeric values associated with a categorical variable are often coded by the researcher in a manner that does not reflect substantive meaning. They are different from numeric values of a non-categorical predictor that are actual measures of an attribute. Unfortunately, SPSS is not capable of distinguishing the former from the latter without additional information from the outside. So, a SPSS procedure with an automatic dummy-coding capability needs to be informed of the categorical nature of a predictor before the procedure generates dummy indicators for it. When a single numeric categorical variable is entered into a SPSS procedure, it should usually be specified either as a factor (e.g., in the GENLIN procedure) or as a categorical covariate (e.g., in the binary logistic procedure). According to the default settings (although these default settings could be overridden or controlled in many cases), SPSS will then identify the "last" level of this predictor as the reference level and create dummy variables for all other categories of this predictor. By default, the "last" level is defined in ascending (from lowest to highest) order of the alpha-numeric coding. So, the highest numeric coding is the default "last" level, and it corresponds to the default reference category selected by a SPSS procedure with an automatic dummy coding capability (SPSS Inc., 2010).

Suppose the two-class categorical predictor gender is presented using two allocated codes: 99 (male) and 100 (female). After specifying this variable either as a factor or as a categorical covariate, SPSS will create (2-1) = 1 dummy variable for this categorical predictor. By default, it first identifies 100 as the last level of this predictor because 100 (for female) > 99 (for male), then it specifies this level (female) as the reference level by assigning values of 0 to the dummy variable, while the other level (male) is coded as 1. It is this newly created dummy variable for gender that is used in the regression model. Furthermore, it is this dummy variable (not the original gender variable) that has a regression coefficient to estimate. The interpretation of the regression coefficient estimate for this dummy-coded gender predictor variable should be made for the male category as is compared with the female category (i.e., reference level).

With the allocated codes of 99 and 100 described for the gender variable, this coding process and the final dummy coding result should remain the same regardless of its allocated codes: As the default option, the last level or, in ascending alpha-numeric order, the highest coding is selected as the reference level before the dummy variable is created for the other level of the two-class categorical predictor. This is even true when the allocated codes are 0 and 1; the same two values that are used in the 0-1 dummy coding scheme.

Suppose that the gender variable is originally allocated two numeric codes: 0 for male and 1 for female. This is similar to the previous example where 99 was allocated to male and 100 to female in the sense that the male category is always allocated the lower coding (i.e., 99 and 0, respectively), whereas the female category is allocated the higher coding (i.e., 100 and 1, respectively). For the second case, after this gender variable is designated as categorical in a SPSS regression procedure with an automatic dummy-coding capability, the procedure identifies the last level or by default, in ascending alpha-numeric order, the higher code of 1 (i.e., the female category) as the reference level and assigns values of 0 to the new dummy variable to be internally computer-generated. The other (non-reference) level of the dummy variable will be created for the other category (i.e., the male category) that is originally allocated the code of 0, so that the male category is now coded as 1 in the newly-generated dummy variable. So, in this new dummy variable, the coding of the gender variable is reversed from the original coding scheme. The dummy variable can then be used as a direct input for a regression model. When it comes to parameter

interpretation of the dummy variable, it should be made for the male category (i.e., coded as 1 in the new variable) as is compared with the female category (i.e., coded as 0 in the new variable).

Such a change in coding may be difficult to spot for practitioners of regression analysis, particularly those who are not familiar with SPSS procedures that can perform dummy-coding automatically. When a practitioner who knows about the 0-1 coding intentionally allocates 0 to male and 1 to female because the interest is in the female group rather than the male group, the above-described coding reversal in some SPSS procedures with an automatic dummy-coding capability causes the final parameter estimate to focus on the male group, instead. When this coding change goes unnoticed, the interpretation of the parameter estimate is very likely to be made still regarding the female group to stay consistent with the original coding, which of course is incorrect. An example follows that demonstrates this point.

Heuristic Example

The data analyzed here as an example comes from Kutner et al. (2004). The data are results from an economist who studied 10 mutual firms and 10 stock firms. The economist was most interested in the relationship between the elapsed time for the innovation to be adopted (Y), size of firm (X_1), and type of firm (X_{Type}). Type of firm (X_{Type}) is a two-class categorical predictor, so it has to be recoded into a numeric variable X_2. Y is expressed in number of months and X_1 in millions of dollars.

Figure 1 presents a screenshot of the data set that has a total of five columns. In the data set, the first two columns are the dependent variable (*ElapsedTime_y*) and one of the two predictors (*X1_Size*). They are not categorical, so no special recoding is needed for any of them. The next 3 columns are all about the other predictor variable; type of firm. The column called *X2_Type* presents the predictor variable in string format. And the other two columns present this predictor in numeric format with the column called *X2_Type99100* using the allocated codes of 99 and 100 and the column called *X2_Type01* using the allocated codes of 0 and 1. In both numeric columns, the lower code (99 and 0) is assigned to the mutual category and the higher code (100 and 1) is assigned to the stock category. Two distinct sets of allocated codes are used here for the purpose of comparing the final modeling results. It is anticipated that, if done properly, the results should be the same because the values allocated to represent company type are just numeric codes without any substantive meaning.

StockMutual.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

23 :

	ElapsedTime_y	X1_Size	X2_Type	X2_Type99100	X2_Type01	var
1	17	151	Mutual	99	0	
2	26	92	Mutual	99	0	
3	21	175	Mutual	99	0	
4	30	31	Mutual	99	0	
5	22	104	Mutual	99	0	
6	0	277	Mutual	99	0	
7	12	210	Mutual	99	0	
8	19	120	Mutual	99	0	
9	4	290	Mutual	99	0	
10	16	238	Mutual	99	0	
11	28	164	Stock	100	1	
12	15	272	Stock	100	1	
13	11	295	Stock	100	1	
14	38	68	Stock	100	1	
15	31	85	Stock	100	1	
16	21	224	Stock	100	1	
17	20	166	Stock	100	1	
18	13	305	Stock	100	1	
19	30	124	Stock	100	1	
20	14	246	Stock	100	1	
21						

Figure 1. A Screen Shot of Data.

To concisely present the information without having to burden the readers with unnecessary details, a simple, first-order regression model is fitted here that does not contain any complex terms like two-way interactions:

$$Mean\ Y = \alpha + \beta_1 * X_1 + \beta_2 * X_2 \quad (1)$$

where Y is the number of months for the innovation to be adopted, X_1 is the (non-categorical) company size variable measured in millions of dollars, and X_2 is the categorical company type variable. This company type variable could take one of the two forms: 1) A string variable in the form of letters (e.g., *X2_Type* in Figure 1), and 2) a numeric variable with allocated codes (e.g., *X2_Type99100*, and *X2_Type01* in Figure 1).

This model is estimated in SPSS in four different ways using two different procedures. Each way of fitting the model features a different approach to the categorical predictor representing type of company. The two procedures used here are the REGRESSION procedure and the GENLIN procedure. They are briefly compared below:

- The former procedure is designed for multiple linear regression based on general linear models. The procedure requires all inputs should be numeric, because it is not capable of automatically handling a categorical predictor in string format. Additionally, this procedure can analyze dummy indicators and/or numeric values of a non-categorical predictor that are actual measures of an attribute, but it does not allow the use of allocated codes representing classes of a categorical predictor. The only exception is when the allocated codes for the classes of a binary categorical predictor are selected to be the same as the dummy codes for that predictor.
- The latter procedure is designed for multiple regression based on generalized linear models that incorporate the type of models for the previous procedure as a special case. Not only is this procedure capable of handling a broader family of regression models, but it is also able to do the two things that the previous procedure cannot do: 1) Handling categorical predictors in non-numeric format, and 2) Handling allocated codes by automatically converting them to dummy codes in the form of 0 and 1.

With that described about the two procedures, they are used to estimate Equation 1 in four different ways. During the four analyses, the REGRESSION procedure uses the *X2_Type01* variable whereas the GENLIN procedure uses each of the three variables (*X2_Type*, *X2_Type99100*, and *X2_Type01*) as predictors in each of three regressions.

Note that when using the GENLIN procedure, both *X2_Type99100*, and *X2_Type01* are entered as Factors under the Predictors tab. This is so because this paper assumes the following: It is ***intuitive*** for a practitioner to think that either one of the two (*X2_Type99100*, and *X2_Type01*) has a categorical nature because it represents a categorical predictor: Type of company and, with that thinking in mind, he or she is likely to tell SPSS about the belief by entering either of the two into the program as a factor. Figures 2 and 3 are screenshots of the Predictors tab when entering *X2_Type99100*, and *X2_Type01*, respectively. In the interest of space, the screenshots of the other two (more straightforward) analyses are omitted from here.

The modeling results from all four analyses are presented in Table 1. The focus is on the parameter estimate for the predictor representing company type. As is indicated, all four analyses have produced similar results. The four estimates for the parameter of the corresponding predictor indicating company type are identical in absolute value. It is just that the parameter estimate from the REGRESSION procedure is (+8.055) whereas all others are (-8.055). Such a difference is due to the fact that the two procedures by default use different reference levels. The

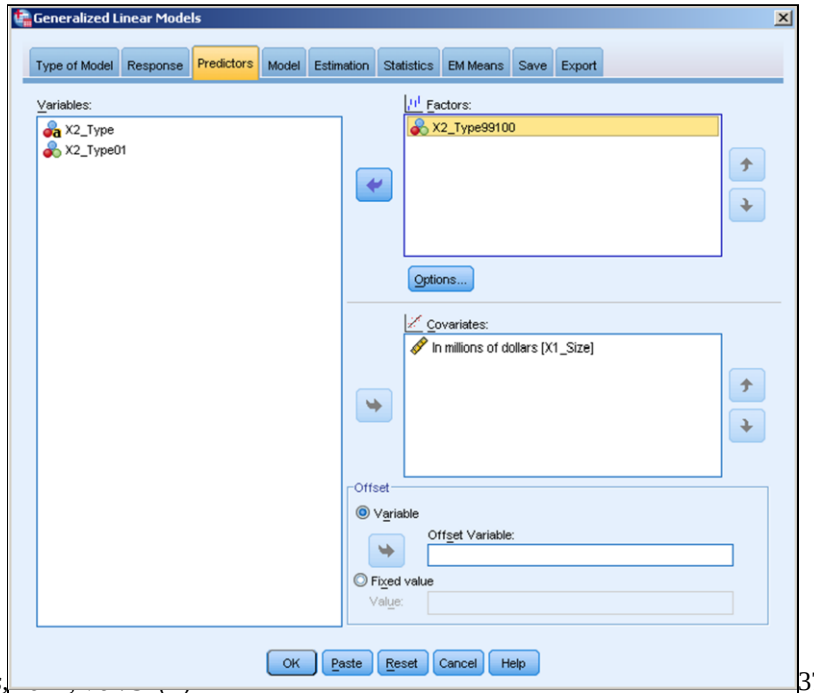

Figure 2. A Screenshot of Entering X2_Type99100.

REGRESSION procedure uses *X2_Type01*=0 (mutual companies) as the reference level and the parameter estimate for the company type variable measures the change in the outcome for stock companies as is compared to mutual companies. The GENLIN procedure does exactly the opposite by default because it uses the *X2_Type01*=1 (or *X2_Type99100*=100, *X2_Type* = Stock) as the reference level, or the last level given an ascending alpha-numeric order of categorical values. So, the parameter estimate for the company type variable of each analysis under the GENLIN procedure measures the change in the outcome for mutual companies as is compared to stock companies. Because the direction of change for the company type predictor is opposite to each other under the two procedures, the estimates for the corresponding parameter have opposite signs, but are identical in absolute value.

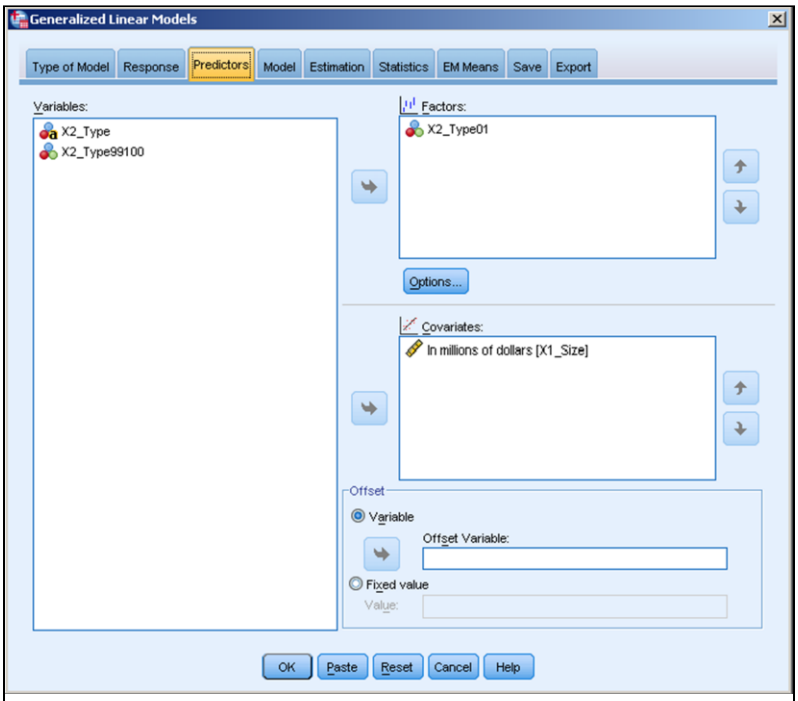

Figure 3. A Screenshot of Entering X2_Type01.

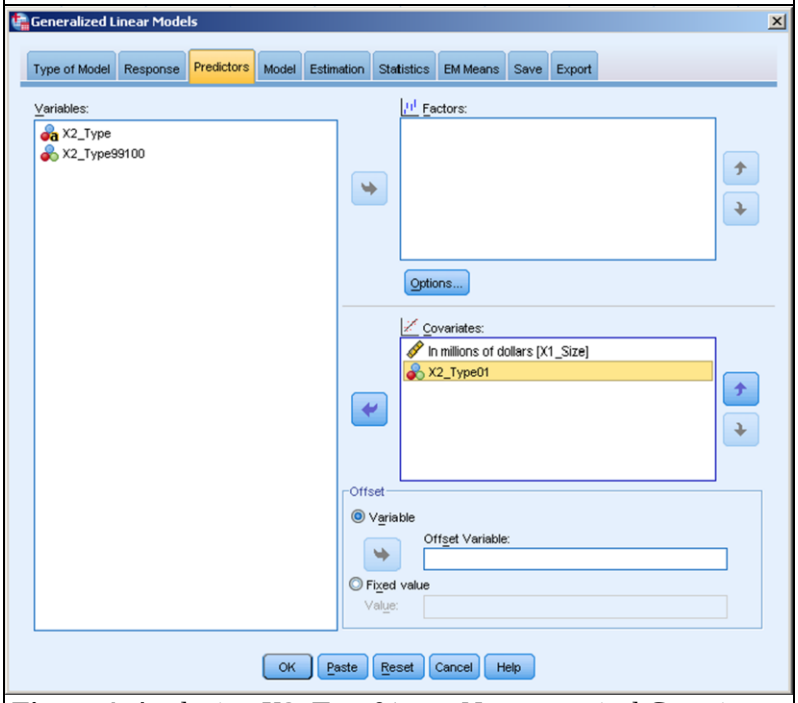

Figure 4. Analyzing X2_Type01 as a Noncategorical Covariate.

Special attention should be paid to the first and fourth analysis. In the first analysis, *X2_Type01* is analyzed in the REGRESSION procedure as an input for the Independent(s) box. In the fourth analysis, the *same* variable is analyzed in the GENLIN procedure (using its default settings) as an input in the Factors box under the Predictors tab. Although it is the same predictor (*X2_Type01*) that is used in both analyses that aim to fit the same regression model as is described by Equation 1, the parameter estimates are exactly opposite to each other due to the reasons outlined above: ((+8.055) in the first analysis versus (-8.055) in the fourth analysis). Both estimates are correct and are equivalent of each other, but they should be interpreted from different perspectives. That is, (+8.055) indicates that, on average, stock companies need 8.055 more months than mutual companies in adopting an innovation whereas (-8.055) suggests the average amount of time mutual companies take to adopt an innovation is 8.055 months less than stock companies.

In fact, it would have been unnecessary to declare *X2_Type01* as categorical because its allocated codes have already been selected to be identical to its dummy codes. In such a case, this variable could be just used as a (non-categorical) covariate in the GENLIN procedure. Figure 4 provides another analysis of the model in Equation 1 using the GENLIN procedure. In this analysis, the *X2_Type01* variable is entered

into the same box as the *X1_Size* variable, which is different from analyses 2, 3, and 4. This time, the parameter estimate for the company type variable becomes (+8.055), the same result as analysis 1. In this case, the interpretation should be made consistently, as noted with the original coding of the *X2_Type01* variable, because no coding reversal has been done.

Another point that is worth noting is, as long as the order of the coding remains the same, the choice of allocated codes for classes of a categorical predictor does not affect the parameter estimates. The third and the fourth analyses use different allocated codes for the two categories of the respective company type variable, but their parameter estimates are the same: (-8.055).

Table 1. Analysis Results from Two SPSS Procedures

	SPSS Statistics Procedure Used			
Items	**REGRESSION**	**GENLIN**		
Analysis	Analysis 1	Analysis 2	Analysis 3	Analysis 4
Variable	*X2_Type01*	*X2_Type*	*X2_Type99100*	*X2_Type01*
Data type	Numeric	String	Numeric	Numeric
Estimate	**(+8.055)**	**(-8.055)**	**(-8.055)**	**(-8.055)**
Base level	*X2_Type01*=0	*X2_Type*=Stock	*X2_Type99100*=100	*X2_Type01*=1

Discussion

The paper focuses on the coding reversal issue that happens to binary categorical predictors in some SPSS procedures that have an automatic dummy coding capability. The paper alerts practitioners of regression analysis to this issue, particularly when the allocated codes for the binary predictor are selected to be the same as the codes for the 0-1 dummy coding scheme. Although there is nothing wrong to think of this binary predictor as categorical in this situation, it requires special attention to find out what category it is that is being used as the reference level by SPSS if the predictor is indeed declared as categorical in the computer program.

When analyzing a two-class categorical predictor that has already been dummy-coded in a SPSS procedure with an automatic dummy-coding capability, the best strategy is not to let the program recode it again. To prevent the program from performing the recoding automatically, the categorical predictor under discussion should be used as a (non-categorical) covariate but not as a factor or a categorical covariate.

With the conclusion drawn based on SPSS, it may also be generalized to other statistics programs. Like SPSS, a few Statistical Analysis System (SAS) procedures by default take a similar approach to a declared categorical predictor; selecting its last category in ascending alpha-numeric order as the reference level, coding it into 0, and using 1 to represent all other non-reference categories. Therefore, the aforementioned analyses 2 to 4 where a two-level predictor is declared as categorical can be duplicated using such SAS procedures that include PROC GLM and PROC GENMOD, which allow the specification of a predictor as categorical using the CLASS statement (SAS Documentation, 2010). In these three analyses using either SAS procedure, the issue of coding reversal as described in this paper also exists. Analysis 1 where a two-level categorical predictor presented in 0 and 1 is analyzed as a non-categorical covariate can be duplicated using either of the above SAS procedures (without declaring the predictor as categorical) or using another one called PROC REG; the counterpart of the REGRESSION procedure in SPSS.

However, there are also procedures in SAS that work differently than PROC GLM or PROC GENMOD, and among them is PROC LOGISTIC for logistic regression analysis. With PROC LOGISTIC, after declaring a two-level predictor already in the form of 0 and 1 as categorical, the interpretation of its parameter estimate should be made relative to the average effect across both levels rather than relative to an internal, computer-generated 0 category that in fact does not exist with PROC LOGISTIC. This is so because PROC LOGISTIC uses a different dummy coding scheme than the other two procedures. Whereas PROC GLM and PROC GENMOD use the 0-1 coding, PROC LOGISTIC performs the (-1)-1 coding where the last category in ascending alpha-numeric order is coded as (-1), instead of 0 (SAS Documentation, 2010). Further, the last category is no longer the reference level with which the other category is compared. It is the average effect across both levels of the categorical

predictor that serves as the baseline of comparison for its two categories. Another statistics program that handles a declared categorical predictor in the same manner as PROC LOGISTIC in SAS is JMP (SAS Institute Inc., 2008). Several of JMP's regression modules (like the Fit Model module) by default also perform the (-1)-1 coding and code into (-1) the last category in ascending alpha-numeric order. Although the issue around the (-1)-1 coding for a declared categorical predictor in a statistics program does not quite fall into the coding reversal issue as discussed in this paper, both cases are likely to cause the original coding of the categorical predictor to change without any alert. Therefore, cautions should be taken when interpreting such parameter estimates.

Finally, the last statistics program that should be discussed here is STATA because it is almost as comprehensive and popular as SPSS and SAS. STATA provides a xi command that is capable of using the 0-1 coding scheme to automatically create dummy variables that can next be analyzed by the regress command for regression modeling. Unlike SPSS, whose default setting in many of its procedures is to pick up the last category in ascending alpha-numeric order as the reference, the xi command in STATA by default does exactly the opposite by selecting the first category as the reference (Hamilton, 2004). Therefore, the coding reversal issue for a declared two-level categorical predictor already in the form of 0 and 1 as described in the paper does not exist with the xi command in STATA.

References

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage Publications Inc.

Hamilton, L. C. (2004). *Statistics with STATA: Updated for version 8*. Belmont, CA: Thomson Learning.

Kutner, M., Nachtsheim, C., & Neter, J. (2004). *Applied linear regression models* (4th ed.). New York: McGraw-Hill.

Norusis, M. J. (2003). *SPSS 12.0 statistical procedures companion*. Chicago: SPSS, Inc.

SAS Documentation. (2010). *SAS/STAT 9.22 user's guide*. Cary, NC: SAS Institute, Inc.

SAS Institute Inc. (2008). JMP (Version 8) [Compute software]. Cary, NC: SAS Institute Inc.

SPSS Inc. (2010). *IBM SPSS Advanced Statistics 19*. Chicago, IL: SPSS, Inc.

Send correspondence to:
Hongwei Yang
University of Kentucky
Email: hya222@uky.edu