11

Sample Size Planning with Applications to Multiple Regression: Power and Accuracy for Omnibus and Targeted Effects

Ken Kelley and Scott E. Maxwell

ABSTRACT

When designing a research study, sample size planning is one of the key factors to consider. One aspect of sample size planning is whether the primary goal of the research study is to reject a false null hypothesis, the power analytic approach. Another primary goal may be to obtain a confidence interval that is sufficiently narrow, the accuracy in parameter estimation approach. Some questions of interest may pertain to a collection of parameters (i.e. an omnibus effect), whereas other questions may pertain to only a single parameter (i.e. a targeted effect). The issue of power or accuracy and the issue of an omnibus effect or a targeted effect leads to a two-by-two conceptualization for planning sample size. The power analytic and accuracy in parameter estimation approaches are discussed in the context of multiple regression for the squared multiple correlation coefficient (an omnibus effect) and for a specific regression coefficient (a targeted effect). A discussion of statistical significance testing and confidence interval construction for the parameters of interest is provided. Whereas the power analytic approach is largely reviewed from existing literature, developments are made for the accuracy in parameter estimation approach.

At the heart of scientific research is the desire for understanding. Even though many methods exist for attempting to gain a better understanding of the phenomenon or phenomena of interest, statistical methods have proven to be the most useful way of extracting information from data. Given that the use of statistical methods is so

vital to scientific research, ensuring that the statistical methods chosen provide the information of interest is an important step for scientific progress. Even though science is often laborious and slow, by designing a well-planned study researchers can be in the best position to maximize their chances for success, where the ultimate goal is gaining a better understanding of the phenomenon of interest.

Designing research studies is arguably the most important single phase of research. With a poorly designed study, little or no understanding of the phenomenon of interest may be gained. Given the high economic and professional costs of poorly designed research, motivation of the researcher should clearly be on the side of beginning an investigation with a well-designed study.

Many facets exist to research design and each one deserves attention. At a minimum, the following points must be considered when designing studies in the behavioral, educational, and social sciences:

(a) the question(s) of interest must be determined;
(b) the population of interest must be identified;
(c) a sampling scheme must be devised;
(d) selection of independent and dependent measures must occur;
(e) a decision regarding experimentation versus observation must be made;
(f) statistical methods must be chosen so that the question(s) of interest can be answered in an appropriate and optimal way;
(g) sample size planning must occur so that an appropriate sample size given the particular scenario, as defined by points a through f, can be used;
(h) the duration of the study and number of measurement occasions need to be considered;
(i) the financial cost (and feasibility) of the proposed study calculated.

Sample size planning (Point g) as it relates to the question(s) of interest (Point a) of an investigation is the focus of this chapter. Although sample size planning is an important part of research design, sample size planning cannot occur without some question of interest first being defined. There are multiple ways to plan sample size for a single study. The way in which sample size is planned depends heavily on the question(s) of interest that the investigator has defined. Thus, not defining the question of interest implies that a method for choosing sample size, and thus the sample size itself, cannot adequately be defined[1].

For example, suppose a researcher wishes to examine the relationship between five regressor variables and a criterion variable in a multiple regression context. However, the process of deciding on an appropriate sample size cannot begin until the question of interest has been clearly defined. There are *at least* four scenarios in which sample size planning can proceed in a multiple regression context:

(a) desired degree of statistical power for the overall fit of the model (i.e. power for the squared multiple correlation coefficient);
(b) desired degree of statistical power for a specific regressor variable (i.e. power for the test of a particular population regression coefficient);
(c) statistical accuracy for the overall fit of the model (i.e. a narrow confidence interval for the population squared multiple correlation coefficient);
(d) statistical accuracy for a specific regressor variable (i.e. a narrow confidence interval for one or more population regression coefficients)[2].

Thus, an appropriate sample size depends very much on the goals of the researcher. Not surprisingly, given the fundamental differences between power and accuracy for omnibus and targeted effects, necessary sample size can be very different in the four scenarios. More general than the multiple regression example, sample size planning can be conceptualized in a two-by-two table, where the effect of interest, either an omnibus or a targeted effect, is on one dimension and the goal, either power or accuracy, is on the other dimension. Such a conceptualization is given in Table 11.1 for sample size planning

Table 11.1 Two-by-two conceptualization of possible scenarios when statistical power is crossed with statistical accuracy

Goal	Effect: Omnibus	Effect: Targeted
Power	a	b
Accuracy	c	d

for points a–d, where the effect is represented by the column dimension and the goal is represented by the row dimension.

Even though Table 11.1 has four cells, none of the cells are mutually exclusive, nor is any specific one necessary. That is to say, a researcher could have the goal of achieving power for the omnibus effect (cell a) and a specific effect (cell b). Likewise, a researcher could have the goal of accuracy for the omnibus effect (cell c) and a specific effect (cell d). A researcher most interested in the omnibus effect could desire both its power (cell a) and its accuracy (cell c). A researcher most interested in a specific effect could desire both power (cell b) and accuracy (cell d). Another possibility is for a researcher to desire a high degree of power for the omnibus effect (cell a) and to desire accuracy for a specific effect (cell d). Conversely, a researcher might desire a high degree of accuracy for the omnibus effect (cell c) and a high degree of power for a specific effect (cell b). Any combination of the cells in the table is possible, and given the goals of the researcher, multiple cells in Table 11.1 might be relevant.

What may not be obvious from Table 11.1 is that the sample size necessary to fulfill one of the scenarios of interest might also be large enough to fulfill one or more other scenario(s). We will discuss methods of planning sample size for each of the scenarios in upcoming sections of the chapter in the context of multiple regression. The next three sections provide overviews and rationales of statistical power, statistical accuracy, and multiple regression, respectively. The overview sections are followed by methods for planning sample size given the goals of statistical power and statistical accuracy for omnibus and targeted effects, respectively, in the context of multiple regression analysis. The computer program R (R Development Core Team, 2007) is used throughout the article with the MBESS package (Kelley, 2007). R is a comprehensive statistics environment and language with powerful graphics capabilities. MBESS is an add-on package for R that has, among other things, numerous functions for assisting researchers planning an appropriate sample size. Both R and MBESS are Open Source and freely available[3]. The R code used throughout the chapter is distinguished from text by using a non-serif font (such as this). R examples are typeset in a gray box with '`R >`' denoting an executable R command as follows:

```
R > mean (data)
```

which returns the mean of the values contained in the object 'data.'

We have synthesized a large amount of work done in the sample size planning literature and packaged it in what we hope is a conceptually appealing and readily comprehensible presentation, complete with easy to use computer commands for planning necessary sample size in each of the four scenarios described.

RATIONALE OF STATISTICAL POWER ANALYSIS

Statistical power is a function of four things: (a) the size of the effect; (b) the model error variance; (c) the Type I error rate (α); and (d) sample size (N)[4]. Power is defined as one minus the probability of a Type II error[5]. In most cases the size of the effect, Type I error rate (e.g. $\alpha = 0.01$ or $\alpha = 0.05$), and often the model error variance are considered fixed, leaving only the sample size as a quantity that is in the control of the researcher[6]. Given that power is in part

a function of sample size, the sample size can be manipulated so that a desired degree of power is reached. Power has been discussed in numerous book length treatments for many statistical tests (e.g. Kraemer & Thiemann, 1987; Cohen, 1988; Lipsey, 1990; Murphy & Myors, 1998).

The use of null hypothesis significance testing has been under fire for some time (e.g. Nickerson, 2000, for a review; the works contained in Rozeboom, 1960; Bakan, 1966; Morrison & Henkel, 1970; Meehl, 1978; Cohen, 1994; Schmidt, 1996). Even though we sympathize with many of the critiques leveled against the use of null hypothesis significance testing, null hypothesis significance testing has its place in science and there is little question that it will continue to be widely used (e.g. Chow, 1996; Hagen, 1997; Harris, 1997; Wainer, 1999; Mogie, 2004). There are two main reasons why null hypothesis significance tests are valuable in research: they help researchers decide if the population value of some effect differs from a specified quantity (generally zero), and for many tests they allow the researcher to decide the direction of the effect. For some questions of interest, the use of null hypothesis significance tests is not especially helpful. In those situations other techniques can be used.

One common alternative to null hypothesis significance testing is the use of effect sizes and their corresponding confidence intervals (e.g. Schmidt, 1996; Thompson, 2002; Smithson, 2003; Hunter & Schmidt, 2004; Steiger, 2004; Grissom & Kim, 2005). Effect sizes and their corresponding confidence intervals can better address issues involving the magnitude of an effect than can null hypothesis significance tests. However, some research questions do not lend themselves to being framed as an effect where the magnitude is meaningful and of interest. This is especially true with some multiparameter and multivariate hypotheses, as such tests are more difficult to transform into an effect size and corresponding confidence interval that is readily interpretable. For example, multivariate analysis of variance and covariance have omnibus effect sizes that are generally not easy to interpret. One option is to reduce such multivariate effects into simpler effects (e.g. pairwise, simple main effects, specific effects, etc.) and then report their corresponding effect sizes and confidence intervals. Even though such effects are readily comprehensible, such simplified hypotheses generally fail to consider the complexity and multivariate nature of the original research question, requiring the questions to be addressed with multivariate techniques that may not have readily interpretable effect sizes. We will discuss the benefits of confidence interval formation in the next section, but we acknowledge that confidence intervals are not adequate for addressing all substantively interesting questions. In cases where a research question is best addressed with a null hypothesis significance test, the a priori power of the test should be as important as the obtained probability value.

Even though the conceptual rationale of power analysis is generally well understood, not often discussed are the implications and importance of mapping a power analysis onto the research question(s) of interest. In a given study, there are often numerous statistical hypotheses evaluated. Given a particular sample size and holding everything else constant, each of the potential statistical tests has a population effect size and model error (or simply a standardized effect size which simultaneously considers both) that must be estimated, and an associated level of statistical power. Sample size can thus be determined so that power is at some desired level for one or several tests. If power is set to a value, such as 0.85, it is likely that a different sample size would be necessary for each of the statistical tests of interest. Depending on the exact question of interest (i.e. for which test is the appropriate sample size determined), necessary sample size to achieve some desired goal will generally be different. Thus, before sample size planning from a power analytic approach can proceed, the exact question of interest must be specified (Point a from the designing research list).

When statistical tests are conducted in situations of low power, the literature of an area can become awash with contradictory results (e.g. Sedlmeier & Gigerenzer, 1989; Rossi, 1990; Hunter & Schmidt, 2004; Maxwell, 2004). For example, suppose several researchers each replicate the same previously reported study using multiple regression with several regressor variables. Further suppose that the power was low for each of the several regressors. It is entirely possible that each of the researchers obtained a different set of statistically significant regression coefficients, none of which mirror the previously reported study! By having low power across multiple parameters, there is often a high probability of obtaining statistical significance somewhere (Kelley et al., 2003), but a small probability of replicating the same set of statistically significant regressors (Maxwell, 2000, 2004). Consistency of research findings is thus difficult if power is low for some or all of the effects examined. Without ensuring that an adequate degree of power is achieved, low-powered studies riddled with Type II errors can permeate the literature and scientific growth can falter because of inconsistencies regarding statistically significant effects across multiple studies that examine the same effects (Rosenthal, 1993; Schmidt, 1996; Kraemer et al., 1998; Hunter & Schmidt, 2004, chapter 1).

Many times when a study has important implications, such as those often conducted in the behavioral, educational, social, and medical sciences, ignoring issues of power is irresponsible and potentially even unethical. This is true, for example, when individuals are subjected to an inferior treatment condition in a study with low power. The individuals in such studies are put at risk with little chance of determining whether some treatments are truly superior to others. A more tangible reason for seriously considering power analysis is that grant funding review boards now generally require explicit consideration of design and power in grant proposals in order to receive funding (e.g. Allison et al., 1997; Kraemer et al., 1998). Thus, not only can ignoring power issues lead to a study with little chance of achieving statistical significance for the parameter(s) of interest, it can prevent the study from even being conducted because funding is not secured.

Power analysis is also an important tool for protecting valuable resources. For example, suppose a study was conducted with a sample size of $N = 20$. Further suppose that the statistical test on the parameter of interest did not yield a statistically significant result. Such a result might be disappointing, but such a result might have also been avoided. Suppose that a power analysis (e.g. based on an independent group t-test where the population standardized mean difference is thought to be 0.40 with the Type I error rate set to 0.05) would have revealed that a sample size of 100 would be necessary in order for the power to equal 0.80, the researcher's operational definition of 'adequate power.' Had such a power analysis been conducted by the researcher a priori, the researcher would have had at least three choices: (a) perform the study with $N = 20$ anyway, with the caveat that there would be only a small probability (specifically 0.23 under the anticipated effect size) of achieving statistical significance (i.e. low power); (b) modify the original design so that the sample size was changed to $N = 100$ in order for the researcher to have an adequate degree of power for detecting the effect of interest; or (c) realize that $N = 100$ is not practical given the difficulty of collecting data and conclude that the cost/benefit ratio is not worth conducting the study at the present time. Points b and c are both enlightening from a resource standpoint, because it may become apparent that $N = 20$ is not adequate and thus using a sample size of only 20 may not be a wise use of resources given the low probability of finding statistical significance.

RATIONALE OF ACCURACY IN PARAMETER ESTIMATION

In order for a piece of information to be meaningful, it is generally desirable for that piece of information to be accurate. In the context of parameter estimation, accuracy is

defined in terms of the (square) root of the mean square error (RMSE), and is a function of precision and bias. Formally, the accuracy of an estimate $\hat{\theta}$ is defined as

$$\begin{aligned}\text{RMSE} &= \sqrt{\text{E}\left[\left(\hat{\theta}-\theta\right)^2\right]} \\ &= \sqrt{\text{E}\left[\left(\hat{\theta}-\text{E}\left[\hat{\theta}\right]\right)^2\right]+\left(\text{E}[\hat{\theta}-\theta]\right)^2} \\ &= \sqrt{\sigma^2_{\hat{\theta}}+\text{B}^2_{\hat{\theta}}}, \end{aligned} \quad (1)$$

where $\text{E}[\cdot]$ is the expected value of the quantity in brackets, θ is the parameter of interest with $\hat{\theta}$ as its estimate, $\sigma^2_{\hat{\theta}}$ is the population variance of the estimator $\left(\text{i.e. E}\left[\left(\hat{\theta}-\text{E}\left[\hat{\theta}\right]\right)^2\right]\right)$, and $\text{B}_{\hat{\theta}}$ is the bias of the estimator $\left(\text{i.e. E}\left[\hat{\theta}-\theta\right]\right)$ (Rozeboom, 1966, p. 500). Whereas precision reflects the repeatability of measurements and is thus inversely related to the sample-to-sample variability, bias is the systematic (i.e. average) discrepancy between an estimate and the parameter it estimates. Notice that when the bias equals zero, the estimate is unbiased and accuracy and precision are equivalent concepts[7]. However, precision alone does not imply an accurate estimate[8].

A narrow confidence interval has a tightly clustered set of plausible parameter values that will contain the parameter of interest with the degree of confidence specified. These plausible parameter values are those that *cannot* be rejected as the value of the population parameter. In the long run when the assumptions of the model are satisfied for an exact confidence interval procedure, (1–α)100% of the confidence intervals formed under the same conditions will contain θ (Hahn & Meeker, 1991, p. 31). Holding the confidence level constant, the narrower the confidence interval width, the more values can be excluded from the plausible set of parameter values. The effect of this is a homing in on the population parameter. Because an appropriately constructed confidence interval will always contain the observed parameter estimate and will contain the parameter (1–α)100% of the time, as the width of the interval decreases the expected accuracy of the estimate improves (i.e. the RMSE is reduced).

The effect of increasing sample size potentially has two effects on accuracy. First, the larger the sample size generally the more precision the estimate will have (i.e. its variance decreases as N increases)[9]. For unbiased estimates, improving the precision necessarily improves accuracy. Estimators that are biased will many times become less biased as sample size increases. Indeed, for consistent estimators, regardless of whether the estimator is biased or unbiased, as sample size tends to infinity the probability that the sample estimate differs from the population quantity by any value tends to zero (Stuart et al., 1994, chapter 17). Thus, above and beyond any effect of precision, decreasing bias also improves accuracy. In fact, even for biased estimates, decreasing the confidence interval width can still be desirable. In such a scenario the point estimate itself might be biased but the range of plausible parameter values sufficiently small[10].

Sample size planning is almost always regarded as being synonymous with power analysis. However, as previously discussed, sample size planning can also proceed with the goal of obtaining a sufficiently narrow confidence interval. We call this method of sample size planning *accuracy in parameter estimation* (AIPE; Kelley & Rausch, in press; Kelley et al., 2003; Kelley & Maxwell, 2003; Kelley, 2006), because when the width of the (1–α)100% confidence interval decreases — implying that there is a smaller range of plausible parameter values at a given confidence level — the expected accuracy of the estimate necessarily increases. Because accuracy can almost never be calculated for a single estimate, due to the fact that it depends on unknown population values, minimizing the confidence interval width to some acceptable value serves as a way to operationally define the expected accuracy of the estimate. Our usage of the term 'accuracy in parameter estimation' is consistent with that

used by Neyman in his seminal work on the theory of confidence intervals: 'the accuracy of estimation corresponding to a fixed value of 1-α may be measured by the length of the confidence interval' (1937, p. 358, notation changed to reflect current system).

It can be argued that obtaining an estimate that has a narrow confidence interval is more beneficial scientifically than obtaining an estimate that reaches statistical significance. It has even been recommended that statistical significance tests be banned and replaced with point estimates and their corresponding confidence intervals (Schmidt, 1996, p. 116). In many situations, especially in observational research, it is known a priori that the null hypothesis is almost always false (Bakan, 1966; Meehl, 1967; Cohen, 1994; Schmidt, 1996; Harris, 1997), and as such situations reaching statistical significance is simply a function of having a large enough sample size (of course, the direction of some effects is often of interest and importance; see our discussion in the previous section)[11]. However, when an effect is of interest, learning as much as possible about the size of the effect is almost always beneficial, and many times it can be more beneficial than learning only the direction and statistical significance of the parameter. Embracing the AIPE approach to sample size planning will help to facilitate the accumulation of scientific knowledge by yielding more accurate information about the parameter. Indeed, as Rosenthal (1993) discusses, there are really two results of interest: (a) the estimate of the magnitude of the effect; and (b) an indication of the accuracy of the effect 'as in a confidence interval around the estimate' (p. 521). Thus, rather than simply asking if an effect differs from some specified null value, in most cases it seems better to address the size of the effect, realizing that the more accurate the estimate of the effect the more information is learned.

Suppose there is no treatment effect in a two-group situation (i.e. the null hypothesis is true). Assuming its assumptions are met, the *t*-test will yield a *p*-value greater than α on (1–α)100% of occasions. The corresponding confidence interval will, on (1–α)100% of occasions, have its lower bound less than zero and its upper bound greater than zero (and thus the null value of zero is contained within the interval and cannot be rejected). Further suppose that the confidence interval contains zero, yet is wide relative to the scale of the measurement. Even though the null hypothesis of zero cannot be rejected, a large range of other plausible values (i.e. those values contained in the confidence limits) can also not be rejected. Contrast such a situation with one where zero is contained within the interval and the width of the confidence interval is narrow. In such a situation it is possible to exclude a wide range of values as being plausible (i.e. those not contained within the confidence limits) and thus narrow the range of plausible values.

When one wishes to show support for the null hypothesis (Greenwald, 1975), the accuracy of the obtained estimate as judged by the width of the corresponding confidence interval should be of utmost concern. The 'good enough' principle can be used and a corresponding 'good enough belt' can be formed for the null value, where the limits of the belt would define what constituted a nontrivial effect (Serlin & Lapsley, 1985, 1993). Suppose that not only is the null value contained within the good enough belt, but so too are the confidence limits. This would be a situation where all of the plausible values would be smaller in magnitude than what has been defined as a trivial effect (i.e. the confidence limits are contained within the good enough belt). In such a situation the limits of the (1–α)100% confidence interval would exclude all effects of any 'meaningful' size. If the parameter is less in magnitude than what is minimally important, then learning this can be very valuable. This information may or may not support the theory of interest, but what is important is that valuable information about the size of the effect, and thus the phenomenon of interest, has been gained. Illuminating the size of the effect is something a null hypothesis test in and of itself cannot do. Furthermore, in order for future researchers

to incorporate the study into a meta-analysis, the size of the effect is required (e.g. Hunter & Schmidt, 2004).

OVERVIEW OF MULTIPLE REGRESSION

Let Y_i be an observed score on some criterion variable for the ith individual ($i = 1, \ldots, N$) and X_{ij} be the observed score for the jth regressor variable ($j = 1, \ldots, p$) for the ith individual[12,13]. The general univariate linear model can be written as

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ip}\beta_p + \varepsilon_i, \tag{2}$$

where β_0 is the population intercept, β_j is the regression coefficient for the j^{th} regressor, and ε_i is the error in prediction for the ith individual generally assumed to be normally distributed with mean zero and variance σ_ε^2 [14]. The matrix analog of Equation 2 can be written as

$$\mathbf{y} = \beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3}$$

where $\mathbf{y}$ is an N length vector of observed criterion variables, β_0 is the intercept, $\mathbf{1}$ is an N length column vector of 1s, $\mathbf{X}$ is an N by p matrix of fixed regressor variables, $\boldsymbol{\beta}$ is a p length vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is an N length vector of errors[15]. The p regression coefficients in the vector $\boldsymbol{\beta}$ can be obtained by manipulation of the normal equations as

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\sigma}_{XY} = \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\sigma}_{YX}^{'}, \tag{4}$$

where $\boldsymbol{\Sigma}_{XX}$ is the p by p covariance matrix of the regressor variables with a minus one power representing the inverse of the matrix, $\boldsymbol{\sigma}_{XY}$ is the p length column vector of covariances of the p regressors with Y and $\boldsymbol{\sigma}_{YX}$ is the p length row vector of covariance of Y with the p regressors ($\boldsymbol{\sigma}_{XY}^{'} = \boldsymbol{\sigma}_{YX}$, where prime denotes transposition). The intercept is defined as

$$\beta_0 = \mu_Y - \boldsymbol{\mu}_{\mathbf{X}}^{'}\boldsymbol{\beta}, \tag{5}$$

where μ_Y is the population mean of Y and $\boldsymbol{\mu}_X$ is the p length vector of population means for the regressor variables (see, for example, Graybill, 1976; Darlington, 1990; Pedhazur, 1997; Rancher, 2000; Cohen et al., 2003 for comprehensive coverage of multiple regression and the general linear model).

Throughout the chapter we assume that the regressor variables are fixed, which implies that in theoretical replications of the study the same $\mathbf{X}$ matrix would be obtained. This would be the case, for example, when the $\mathbf{X}$ matrix is literally developed as part of the study design. Theoretical replications of the study would then have the same $\mathbf{X}$ matrix and the only variation would be the values of the criterion variables (and thus the error). When the regressors are random, and thus in theoretical repetitions of the study different $\mathbf{X}$ matrices would be obtained, the discussion that follows would need to be modified to take into consideration the increased randomness of the design (e.g. Sampson, 1974; Gatsonis & Sampson, 1989; Rancher, 2000).

Often of interest in a multiple regression context is the squared multiple correlation coefficient, sometimes termed the coefficient of determination. Recall that the squared multiple correlation coefficient is the proportion of variance in Y that is accounted for by the p regressor variables. The population multiple correlation coefficient, denoted with an uppercase Greek rho, squared, is defined as

$$P_{Y\cdot\mathbf{X}}^2 = \frac{\boldsymbol{\sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\sigma}_{XY}}{\sigma_Y^2}, \tag{6}$$

which is equivalent to the population squared product moment correlation coefficient between the observed scores (Y_i) and the predicted scores ($\hat{Y}_i$; i.e. $P_{Y\cdot\mathbf{X}}^2 = \rho_{Y\hat{Y}}^2$) [16].

Equations 2–6 have used only population parameters. In practice, of course, only the sample means, variances, and covariances are known. The means and the variance/covariance matrix of the $p + 1$ variables (the outcome variable and the p regressor variables) are estimated with the usual unbiased estimates and substituted into Equations 4–6. The estimate of $\boldsymbol{\beta}$ corresponding to the p

regressor variables, $\mathbf{b}$, can be obtained by substituting $\mathbf{s}_{YX}$ or $\mathbf{s}_{XY}$ and $\mathbf{S}_{XX}$ for their population analogs into Equation 4:

$$\mathbf{b} = \mathbf{S}_{XX}^{-1}\mathbf{s}_{XY} = \mathbf{S}_{XX}^{-1}\mathbf{s}_{YX}^{'}. \quad (7)$$

Likewise, the estimate of β_0 can be obtained by substituting the sample means for the population means and the vector of sample regression coefficients in Equation 5:

$$b_0 = \overline{Y} - \overline{\mathbf{X}}'\mathbf{b}. \quad (8)$$

The estimate of $P^2_{Y\cdot\mathbf{X}}$, $R^2_{Y\cdot\mathbf{X}}$, is obtained by substituting the sample estimates of the parameters into Equation 6:

$$R^2_{Y\cdot\mathbf{X}} = \frac{\mathbf{s}_{YX}\mathbf{S}_{XX}^{-1}\mathbf{s}_{XY}}{s^2_Y}. \quad (9)$$

An obtained estimate will almost certainly not equal its population value. What is generally of interest is knowing if the population value differs from some specified null value (generally zero) or determining the plausible values of the parameter (i.e. the values contained within the $(1-\alpha)100\%$ confidence interval). The next two sections discuss null hypothesis significance testing and confidence interval formation, respectively, first for the squared multiple correlation coefficient and then for regression coefficients. Null hypothesis significance tests and confidence interval formation are briefly discussed for regression parameters in order to form a basis for the methods of sample size planning that will be discussed in later sections of the chapter.

NULL HYPOTHESIS SIGNIFICANCE TESTS FOR REGRESSION PARAMETERS

The idea of a null hypothesis significance test is to infer if values at least as extreme as the observed value are sufficiently unlikely if in fact the population value were equal to the specified null value (usually zero). Of course, when claiming statistical significance, there is always the possibility of a Type I error, but that is the price of rejecting a population value based on a sample value. The next two subsections discuss the two most common null hypotheses that are tested in the context of multiple regression: the test that $P^2_{Y\cdot\mathbf{X}} = 0$ and the test that β_j=0.

The test of the null hypothesis that the squared multiple correlation coefficient equals zero

When $P^2_{Y\cdot\mathbf{X}}$ is zero, by implication $\boldsymbol{\beta}$ is a p-length vector of zeros (i.e. $\boldsymbol{\beta} = \mathbf{0}_p$). Of course, in any particular sample, $R^2_{Y\cdot\mathbf{X}}$ will almost certainly be greater than zero. It is a task of the researcher to evaluate if enough evidence exists to reject the idea that $P^2_{Y\cdot\mathbf{X}}$ is zero. When the null hypothesis that $P^2_{Y\cdot\mathbf{X}} = 0$ is true, a test statistic can be formed from $R^2_{Y\cdot\mathbf{X}}$that follows a central F-distribution. The statistic that is used to test the null hypothesis for the squared multiple correlation coefficient is

$$F = \frac{R^2_{Y\cdot\mathbf{X}}/p}{\left(1 - R^2_{Y\cdot\mathbf{X}}\right)/(N-p-1)}, \quad (10)$$

where the F-value has p and $N - p - 1$ degrees of freedom. Of course, this F-statistic has an associated probability value, and if the obtained p-value is less than the adopted Type I error rate (i.e. the α level), then the null hypothesis can be rejected.

When $P^2_{Y\cdot\mathbf{X}}$is not zero, implying that $\boldsymbol{\beta} \neq \mathbf{0}_p$ (i.e. at least one element of the vector of regression coefficient is non-zero), the distribution of the F-statistic from Equation 10 follows a noncentral F-distribution, whereas the F-statistic when the null hypothesis is true follows a central F-distribution (the central F-distribution is the standard 'F-distribution' discussed in introductory and intermediate level statistics books). Rather than having only two parameters, the numerator and denominator degrees of freedom like the central F-distribution, the noncentral F-distribution also has a noncentrality parameter. The noncentrality

parameter indexes the magnitude of the difference between the null and alternative hypotheses. The larger the difference between the null and alternative hypotheses, the larger is the noncentrality parameter.

It can be shown that the noncentrality parameter of the sampling distribution for the F-statistic of Equation 10 is given as

$$\Lambda = f^2 N, \tag{11}$$

where

$$f^2 = \frac{\mathrm{P}^2_{Y\cdot\mathbf{X}}}{1 - \mathrm{P}^2_{Y\cdot\mathbf{X}}} \tag{12}$$

and where f^2 has an interpretation as the signal-to-noise ratio (Cohen, 1988; Stuart et al., 1999; Rancher, 2000; Smithson, 2001). As can be seen, Λ is a function of $\mathrm{P}^2_{Y\cdot\mathbf{X}}$ and N. As either of these quantities becomes larger, so too does Λ. The effect of a larger Λ is that the sampling distribution of the F-statistic in Equation 10 has a larger mean and for fixed sample size values will be more positively skewed. Thus, a larger proportion of the noncentral distribution will be larger than the critical value under the null hypothesis. This idea will become important in the discussion of power and for confidence interval formation.

The test of the null hypothesis that a regression coefficient equals zero

Let $\mathrm{P}^2_{Y\cdot\mathbf{X}_{-j}}$ be the population squared multiple correlation coefficient when Y is predicted from $p-1$ regressor variables with X_j excluded. Researchers are often interested in knowing if a specific regressor variable adds a statistically significant amount to the fit of the model, which translates into a test of $\mathrm{P}^2_{Y\cdot\mathbf{X}}$ being larger than $\mathrm{P}^2_{Y\cdot\mathbf{X}_{-j}}$. Such a test is equivalent to the test of the regression coefficient for X_j when all of the p variables are included in the model.

One of the ways to test the hypothesis that $\boldsymbol{\beta}_j$ is non-zero is to conduct a t-test directly on b_j from the full model. A null hypothesis significance test for a regression coefficient, evaluated against a null value of zero, is based on a t-value with $N-p-1$ degrees of freedom, and is given as

$$t = \frac{b_j}{s_{b_j}}, \tag{13}$$

where s_{b_j} is given as

$$s_{b_j} = \sqrt{\frac{1 - R^2_{Y\cdot\mathbf{X}}}{\left(1 - R^2_{\mathrm{X}_j\cdot\mathbf{X}_{-j}}\right)(N-p-1)}}\left(\frac{s_Y}{s_{\mathrm{X}_j}}\right), \tag{14}$$

with $R^2_{\mathrm{X}_j\cdot\mathbf{X}_{-j}}$ being the squared multiple correlation coefficient using the jth regressor as the criterion on the remaining $p-1$ regressors. $R^2_{\mathrm{X}_j\cdot\mathbf{X}_{-j}}$ is also indirectly available from $\mathbf{S}_{\mathrm{XX}}$ as

$$R^2_{\mathrm{X}_j\cdot\mathbf{X}_{-j}} = 1 - \left(s_j^2 \mathrm{c}_{jj}\right)^{-1}, \tag{15}$$

where s_j^2 is the variance for the j^{th} regressor and c_{jj} is the jth diagonal element of $\mathbf{S}^{-1}_{\mathrm{XX}}$ (Harris, 2001).

Similar to the situation described previously when the null hypothesis that $\mathrm{P}^2 = 0$ is false and the F-statistic of Equation 10 follows a noncentral distribution, so too does the test statistic of Equation 13 when $\beta_j = 0$. It can be shown that when the null hypothesis that $\beta_j = 0$ is false, the t-statistic in Equation 13 has a noncentrality parameter which can be written as

$$\lambda_j = f_j\sqrt{N}, \tag{16}$$

where

$$f_j = \beta_j \sqrt{\frac{1 - \mathrm{P}^2_{\mathrm{X}_j\cdot\mathbf{X}_{-j}}}{1 - \mathrm{P}^2_{Y\cdot\mathbf{X}}}}\left(\frac{\sigma_{X_j}}{\sigma_Y}\right). \tag{17}$$

Because β_j can be written (e.g. Hays, 1994) as

$$\beta_j = \sqrt{\frac{\mathrm{P}^2_{Y\cdot\mathbf{X}} - \mathrm{P}^2_{Y\cdot\mathbf{X}_{-j}}}{1 - \mathrm{P}^2_{\mathrm{X}_j\cdot\mathbf{X}_{-j}}}}\left(\frac{\sigma_Y}{\sigma_{\mathrm{X}_j}}\right), \tag{18}$$

f_j from Equation 17 can be rewritten as

$$f_j = \sqrt{\frac{\mathrm{P}^2_{Y\cdot\mathbf{X}} - \mathrm{P}^2_{Y\cdot\mathbf{X}_{-j}}}{1 - \mathrm{P}^2_{Y\cdot\mathbf{X}}}}. \tag{19}$$

As mentioned, the test of a specific regression coefficient is equivalent to the test of no change in $P^2_{Y\cdot\mathbf{X}}$ when the jth regressor is removed from the regression equation $\left(\text{i.e. } P^2_{Y\cdot\mathbf{X}} - P^2_{Y\cdot\mathbf{X}_{-j}} = 0\right)$. This is in turn equivalent to the test of the squared semi-partial (part) correlation of Y with the jth regressor being zero. Let $P^2_{Y\cdot(X_j\cdot\mathbf{X}_{-j})}$ be the correlation of Y with the independent part of X_j (i.e. the squared semi-partial correlation between Y and X_j). The definition of $P^2_{Y\cdot(X_j\cdot\mathbf{X}_{-j})}$ is given as

$$P^2_{Y\cdot(X_j\cdot\mathbf{X}_{-j})} = P^2_{Y\cdot\mathbf{X}} - P^2_{Y\cdot\mathbf{X}_{-j}}\,. \tag{20}$$

Similar to the test of $P^2_{Y\cdot\mathbf{X}}$ from Equation 10, the test of $P^2_{Y\cdot(X_j\cdot\mathbf{X}_{-j})}$ can be written as an F-statistic with 1 and $N - p - 1$ degrees of freedom:

$$\begin{aligned} F &= \frac{\left(R^2_{Y\cdot\mathbf{X}} - R^2_{Y\cdot\mathbf{X}_{-j}}\right)/(p-(p-1))}{\left(1-R^2_{Y\cdot\mathbf{X}}\right)/(N-p-1)} \\ &= \frac{R^2_{Y\cdot(X_j\cdot\mathbf{X}_{-j})}}{\left(1-R^2_{Y\cdot\mathbf{X}}\right)/(N-p-1)}, \end{aligned} \tag{21}$$

The F-statistic of Equation 21 is the square of the t-statistic in Equation 13. The reason for rewriting the t-statistic for β_j as an F-test for the change in $P^2_{Y\cdot\mathbf{X}}$ when X_j is removed $\left(\text{i.e. } P^2_{Y\cdot\mathbf{X}} - P^2_{Y\cdot\mathbf{X}_{-j}}\right)$ from the prediction equation is to show the relationship between the omnibus F-statistic of Equation 10 and the targeted F-statistic of Equation 21. This relationship will become important later when discussing power.

It should be noted that the noncentrality parameter of the test of a single regression coefficient is very similar to the noncentrality parameter of the test of all regression coefficients tested simultaneously (i.e. the test of $P^2_{Y\cdot\mathbf{X}} = 0$). The signal-to-noise ratio for the change in $P^2_{Y\cdot\mathbf{X}}$ when the jth regressor is removed is given as

$$f^2_{-j} = \frac{P^2_{Y\cdot\mathbf{X}} - P^2_{Y\cdot\mathbf{X}_{-j}}}{P^2_{Y\cdot\mathbf{X}}}, \tag{22}$$

implying the noncentrality parameter for the jth regressor is

$$\Lambda_{-j} = f^2_{-j}N. \tag{23}$$

It should be kept in mind that all derivations have been for the case where the regressors are considered fixed. This and the previous section laid out the formal distributional theory of $R^2_{Y\cdot\mathbf{X}}$ and b_j. The derivations given in this section allow them to be used in a future section that deals with statistical power for the squared multiple correlation coefficient.

CONFIDENCE INTERVAL FORMATION FOR REGRESSION PARAMETERS

In order to understand how well an observed estimate represents its corresponding parameter, confidence intervals are necessary. Confidence intervals for some effects are simple and involve only the estimate, the standard error of the estimate, and the critical value from the test of the null hypothesis (e.g. the critical t, F, or χ^2 value). However, in certain cases the confidence interval is more complicated and involves the use of noncentral distributions.

Noncentral distributions, as will be discussed in a future section, are important for determining sample size in a power analytic context. These distributions are also important for confidence interval formation for certain effects, especially those that have been standardized or when the sampling distribution of the statistic does not follow a central distribution or a mean-shifted central distribution[17]. Effects that are standardized will *not* generally follow a central distribution, because such effects are not pivotal. Stuart et al. (chapter 23, 1999) provide a technical discussion of pivotal quantities, but in the context of effect sizes, a pivotal quantity is one where the confidence interval is a simple rearrangement of the test statistic (Cumming and Finch, 2001). Effects such as the squared multiple correlation coefficient (e.g. Smithson, 2003), the standardized mean difference (e.g. Steiger &

Fouladi, 1997; Cumming & Finch, 2001; Kelley, 2005), and standardized regression coefficients all require the use of noncentral distributions. The following subsection will discuss methods of forming confidence intervals when noncentral distributions are required.

Forming noncentral confidence intervals: Applications to regression parameters

Confidence intervals based on noncentral distributions are computed in a different manner than typical confidence intervals based on central distributions. Two principles, or their equivalent, are necessary and are described below. The description given here is largely based on Steiger and Fouladi (1997) and Steiger (2004).

The *confidence interval transformation* principle is beneficial for forming a confidence interval on a parameter that is monotonically related to another parameter, when the latter has a tractable method of obtaining the confidence interval whereas the former might not. Let $f(\theta)$ be a monotonic transformation of θ, some parameter of interest, with θ_L and θ_U being the lower and upper $(1-\alpha)100\%$ ($\alpha = \alpha_L + \alpha_U$; generally $\alpha_L = \alpha_U = \alpha/2$) confidence limits for θ, where α_L and α_U define the lower and upper proportion of the distribution beyond the lower θ_L and upper θ_U, respectively. The $(1-\alpha)100\%$ confidence limits for $f(\theta)$ are $f(\theta_L)$ and $f(\theta_U)$,

$$prob.[f(\theta_L) \leqslant f(\theta) \leqslant f(\theta_U)] = 1-(\alpha_L+\alpha_U),$$

where *prob.* represents probability. Thus, for monotonic transformations the confidence interval for the transformed population quantity is obtained by applying the same transformation to the limits of the confidence interval for the population quantity (Steiger & Fouladi, 1997; Steiger, 2004).

The *inversion confidence interval principle* states that if $\hat{\theta}$ is an estimate of θ with a cumulative distribution that depends on some Ψ, the probability of observing an estimate of θ smaller than that obtained is given as $p(\hat{\theta}|\Psi)$. Calculation of a confidence interval for θ based on the inversion confidence interval principle involves finding θ_L such that $p(\hat{\theta}|\theta_L) = 1-\alpha_L$ for the lower limit and θ_U such that $p(\hat{\theta}|\theta_U) = \alpha_U$ for the upper limit. The confidence interval for θ has coverage of $1-(\alpha_L+\alpha_U)$ and is given as

$$prob.[\theta_L \leqslant \theta \leqslant \theta_U] = 1-(\alpha_L+\alpha_U).$$

The confidence interval is general and need not have equal rejection regions. For example, a one-sided confidence interval is obtained by setting α_L or α_U (whichever is appropriate for the specific situation) to zero (Steiger & Fouladi, 1997; Steiger, 2004).

The real benefit from the confidence interval transformation and inversion confidence interval principles, is that when the two principles are combined, confidence intervals for quantities that are not pivotal can be determined. In the context of effect sizes, Cumming & Finch (2001) describe pivotal quantities to be those that are of the form

$$\frac{\hat{\theta}-\theta^*}{s_{\hat{\theta}}},$$

where $\hat{\theta}$ is the estimate of the population quantity θ, $\theta^*$is the null value of interest (usually zero), and $s_{\hat{\theta}}$ is the standard deviation of the sampling distribution of $\hat{\theta}$ (i.e. its standard error). What can be done in order to form confidence intervals for non-pivotal quantities is to use the inversion confidence interval principle to find a confidence interval for some noncentrality value (i.e. what values of the noncentrality parameter lead to the observed noncentrality parameter being the $1-\alpha/2$ and $\alpha/2$ quantiles?). When these values are found, the noncentrality parameters (i.e. the confidence bounds of the noncentral value) are transformed into the statistic of interest, which then yields a $(1-\alpha)100\%$ confidence interval for the parameter of interest. Stated another way, confidence intervals for non-pivotal quantities are found by determining the values of the noncentrality parameter that would lead to the observed noncentral

value having probability $1 - \alpha/2$ and $\alpha/2$ for the lower and upper confidence limits, respectively. The values of the noncentrality parameter that would lead to the observed values occurring with the specified probabilities are then transformed into the quantity of interest. The resultant limits form the $(1 - \alpha)100\%$ confidence interval for the population quantity of interest. Although true for confidence intervals based on central distributions when $\alpha_L = \alpha_U$, there is no requirement that the lower confidence interval width, $\hat{\theta} - \theta_L$, will equal the upper confidence interval width, $\theta_U - \hat{\theta}$ for confidence intervals based on noncentral distributions. Throughout the chapter, 'width' refers to the full confidence interval width, $\theta_U - \theta_L$.

Confidence interval for the squared multiple correlation coefficient

The squared multiple correlation coefficient is one of the most widely used statistics. $R^2_{Y\cdot\mathbf{X}}$ is almost always reported in the context of multiple regression, but in its various forms $R^2_{Y\cdot\mathbf{X}}$ can be used to describe the proportion of variance accounted for in a wide variety of situations (e.g. between subjects analysis of variance and covariance designs; as a measure of cross validation; as an index of comparison in meta-analyses, etc.). As Steiger states, 'confidence intervals for the squared multiple correlation are very informative yet are not discussed in standard texts, because a single simple formula for the direct calculation of such an interval cannot be obtained in a manner that is analogous to the way one obtains a confidence interval for the population mean' (2004, p. 167). However, confidence intervals for the population squared multiple correlation coefficient are available with certain software (e.g. R2, an MS-DOS program written by Steiger and Fouladi, 1992; MultipleR2, a Mathematica package written by Mendoza and Stafford, 2001; MBESS, an R package written by Kelley (2007); and indirectly with SAS and SPSS, Smithson, 2003). Difficulties arise when forming a confidence interval for $P^2_{Y\cdot\mathbf{X}}$ because when $P^2_{Y\cdot\mathbf{X}} \neq 0$, the test statistic given in Equation 10 follows a noncentral F-distribution with noncentrality parameter Λ, as given in Equation 11. In accord with the inversion confidence interval principle, $R^2_{Y\cdot\mathbf{X}}$ must be converted into the estimated noncentrality parameter and then noncentral parameters must be found such that

$$p\left(\hat{\Lambda}|\Lambda_L\right) = 1 - \alpha/2 \tag{24}$$

and

$$p\left(\hat{\Lambda}|\Lambda_U\right) = \alpha/2, \tag{25}$$

where $\hat{\Lambda}$ is the observed noncentrality parameter, Λ_L and Λ_U are the noncentral values that have at their $1–\alpha/2$ and $\alpha/2$ quantiles $\hat{\Lambda}$ and are thus the lower and upper confidence limits, respectively (e.g. Mendoza and Stafford, 2001; Smithson, 2003; Steiger 2004).

The MBESS R package includes a function, ci.R2(), for confidence interval formation for $P^2_{Y\cdot\mathbf{X}}$, for fixed (or random) regressor variables. Although other options can be specified, a straightforward call to the ci.R2() function for fixed regressor variables would be of the form

```
R > ci.R2(R2 = R²_Y.X, N = N, p = p,
    conf.level = 1 − α,
    Random.Regressors = FALSE)
```

where $R^2_{Y\cdot\mathbf{X}}$, N, p, and $1–\alpha$ are defined in the function in the same way as they have been defined previously and Random.Regressors identifies if the regressors are random (TRUE) or fixed (FALSE). For example, suppose a researcher conducts a study with five regressor variables on 145 individuals and obtains a multiple correlation of $R^2_{Y\cdot\mathbf{X}} = 0.7854$[18]. The ci.R2() function for 95% confidence interval coverage could be specified as

```
R > ci.R2(R2 = 0.7854, N = 145,
    p = 5, conf.level = 0.95,
    Random.Regressors = FALSE)
```

which yields a confidence interval of $CI_{0.95} = [0.7165 \leqslant P^2_{Y \cdot \mathbf{X}} \leqslant 0.8206]$, where $CI_{0.95}$ represents a 95% confidence interval with the limits given in the brackets for the parameter on interest. Thus, we can be 95% confident that the population squared multiple correlation coefficient in this situation is somewhere between 0.7165 and 0.8206.

Confidence interval for a regression coefficient

Before forming a confidence interval for a regression coefficient, the distinction has to be made whether or not the regression coefficient will be standardized. An unstandardized regression coefficient is a pivotal quantity, whereas a standardized regression coefficient is a non-pivotal quantity (in an analogous fashion as the difference between two group means is pivotal but the standardized difference between two group means is nonpivotal). Thus, a confidence interval for an unstandardized regression coefficient requires only a critical value from a central distribution whereas a standardized regression coefficient requires the critical values to be obtained from a noncentral distribution (analogous to forming a confidence interval for $P^2_{Y \cdot \mathbf{X}}$). The following two sections discuss confidence intervals for unstandardized and standardized regression coefficients.

Confidence intervals for an unstandardized regression coefficient

The t-test for the unstandardized regression coefficient, Equation 11, is a pivotal quantity implying that the test statistic can be manipulated into a confidence interval. The confidence interval for the unstandardized regression coefficient is thus given as

$$prob.[b_j - t_{(1-\alpha/2;N-p-1)} s_{b_j} \leqslant \beta_j \leqslant b_j + t_{(1-\alpha/2;N-p-1)} s_{b_j}] = 1 - \alpha. \quad (26)$$

The confidence interval given above is the confidence interval given in standard textbooks that discuss multiple regression.

The MBESS R package includes a function, ci.reg.coef(), for confidence interval formation for β_j. A confidence interval for an unstandardized regression coefficient can be obtained by specifying the standard deviations of the variables (with the arguments s.Y and s.X) and specifying Noncentral = FALSE. In the situation described for the unstandardized regression coefficients ($b_j = 4.4245$), where $s_Y = 150.0734$ and $s_{X_j} = 9.3605$, the ci.reg.coef() function could be specified as

```
R > ci.reg.coef( b.j = 4.4245,
    R2. Y_X = 0.7854,
    R2.j_X.without.j = 0.3607, N = 145, p = 5,
    s.Y = 150.0734, s.X = 9.3605,
    conf.level = 0.95, Noncentral = FALSE)
```

which yields a confidence interval of $CI_{0.95} = [2.8667 \leqslant \beta_j \leqslant 5.9823]$, where b.j is the unstandardized regression coefficient for the jth regressor variable, R2.Y_X is the squared multiple correlation coefficient, R2.j_X.without.j is the squared multiple correlation coefficient when the jth regressor variables are predicted from the remaining $p - 1$ regressor variables, conf.level is the confidence level specified (i.e. $1 - \alpha$), and Noncentral is an indicator of whether or not the noncentral method should be used (FALSE for unstandardized and TRUE for standardized regression coefficients).

Confidence intervals for a standardized regression coefficient

When a regression coefficient is standardized, the unstandardized regression coefficient is multiplied by the quantity $\frac{s_{X_j}}{s_Y}$ in order to remove the scale of X_j and Y. Such a quantity is no longer pivotal because of the process of standardization, implying that the confidence interval necessarily depends on a noncentral t-distribution. The difficulties that arise when forming a confidence interval for ${}_s\beta_j$, the population standardized regression coefficient for the jth regressor, arise because

b_j is multiplied by $\frac{s_{X_j}}{s_Y}$ (in order to obtain ${}_sb_j$, the sample standardized regression coefficient for variable j). The distribution of ${}_sb_j$ is not pivotal and it is necessary to form confidence intervals based on noncentral t-distributions. In accord with the inversion confidence interval principle, ${}_sb_j$ must be converted into the observed noncentrality parameter (via, Equation 13), and then the noncentral parameters must be found such that

$$p\left(\hat{\lambda}|\lambda_L\right) = 1 - \alpha/2 \tag{27}$$

and

$$p\left(\hat{\lambda}|\lambda_U\right) = \alpha/2, \tag{28}$$

where λ_L and λ_U are the lower and upper confidence limits for ${}_s\beta_j$and are noncentrality parameters from t-distributions.

The MBESS R package includes a function, ci.reg.coef(), for confidence interval formation for ${}_s\beta_j$, technically assuming fixed regressor variables. Although other options can be specified, a straightforward call to the ci.reg.coef() function would be of the form

R > ci.reg.coef (b.j = ${}_sb_j$, R2.Y_X = $R^2_{Y\cdot\mathbf{X}}$,
R2.j_X.without.j = $R^2_{X_j\cdot\mathbf{X}_{-j}}$, N = N, p = p,
conf.level = $1-\alpha$, Noncentral = TRUE).

For example, in the previous example where $N = 145$ and $R^2_{Y\cdot\mathbf{X}} = 0.7854$, suppose that ${}_sb_j = 0.2760$ and $R^2_{X_j\cdot\mathbf{X}_{-j}} = 0.3607$. The ci.reg.coef() function for 95% confidence interval coverage could be specified as

```
R > ci.reg.coef(b.j = 0.2760,
  R2.Y_X = 0.7854,
  R2.j_X.without.j = 0.3607, N = 145, p = 5,
  conf.level = 0.95, Noncentral = TRUE)
```

which yields a confidence interval of $CI_{0.95} = [0.1739 \leqslant {}_s\beta_j \leqslant 0.3771]$. Notice the asymmetry between the confidence limits and the estimate for the standardized regression coefficient, whereas it was symmetric for the unstandardized regression coefficient.This asymmetric property about the point estimate generally holds for confidence intervals based on noncentral distributions.

SAMPLE SIZE PLANNING FOR MULTIPLE REGRESSION GIVEN THE GOAL OF STATISTICAL POWER

This section discusses methods to plan sample size for statistical power in multiple regression. We begin with an overview of sample size planning for a desired power for the omnibus effect (i.e. $P^2_{Y\cdot\mathbf{X}}$) and then provide an overview of sample size planning for a desired power for a targeted effect (i.e. β_j or ${}_s\beta_j$).

Power for omnibus effects in multiple regression: Obtaining statistical significance for the squared multiple correlation coefficient

When interest concerns the omnibus effect of the model, recall that the noncentrality parameter was previously shown (Equations 11–12) to equal

$$\Lambda = \left(\frac{P^2_{Y\cdot\mathbf{X}}}{1 - P^2_{Y\cdot\mathbf{X}}}\right) N. \tag{29}$$

This implies that sample size is given as

$$N = \Lambda\left(\frac{1 - P^2_{Y\cdot\mathbf{X}}}{P^2_{Y\cdot\mathbf{X}}}\right). \tag{30}$$

Thus, given $P^2_{Y\cdot\mathbf{X}}$ and Λ, sample size can be determined. Once $P^2_{Y\cdot\mathbf{X}}$ is specified, Λ is the only unknown parameter since N is unknown. If the Λ that satisfies a desired degree of power can be determined, then the equation can be solved for necessary sample size.

Power is based on Λ and the degrees of freedom, which are in turn based on N. Even though, N is unknown, it is the value of interest when planning a study with a desired degree of power. The way to plan an appropriate sample size is to use different values of N to update Λ and the degrees of freedom until the desired level of power is achieved for the test that $P^2_{Y\cdot\mathbf{X}} = 0$. This process of using different

values of N, which occurs essentially by systematic trial and error, can be implemented using tabled values (e.g. Kraemer & Thiemann, 1987; Cohen, 1988; Murphy & Myors, 1998; Lipsey, 1990) or with a noncentral F computer routine (see also Gatsonis & Sampson, 1989; Green, 1991; Dunlap et al., 2004). The general idea of the power analysis procedure is to determine the sample size so that the proportion of the alternative distribution beyond the critical value under the null distribution is at or greater than the desired degree of power.

The ss.power.R2() function from MBESS can be used to determine sample size for the omnibus effect of the regression model $\left(\text{i.e., } \rho^2_{Y\cdot\mathbf{X}}\right)$. For example, suppose a researcher wishes to determine necessary sample size when it is believed $\rho^2_{Y\cdot\mathbf{X}} = 0.25$ for the test of the null hypothesis that the squared multiple correlation coefficient is zero in order to have power of 0.80 when the Type I error rate is specified at $\alpha = 0.05$. The basic way in which the ss.power.R2() function from MBESS would be used is as follows:

```
R > ss.power.R2(Population.R2 = 0.25,
    alpha.level = 0.05,
    desired.power = 0.80, p = 5)
```

where Population.R2 is the (hypothesized) value of $\rho^2_{Y\cdot\mathbf{X}}$, alpha.level is the Type I error rate, desired.power is the desired degree of power, and *p* is the number of regressor variables. Applying this function to the example yields a necessary sample size of 45.

Power for targeted effects in multiple regression: Obtaining statistical significance for a regression coefficient of interest

When the effect of interest concerns a single regression coefficient, the noncentrality parameter from the noncentral t-distribution was previously shown (Equations 16–19) to equal

$$\lambda_j = \beta_j \sqrt{\frac{1-\rho^2_{X_j\cdot\mathbf{X}_{-j}}}{1-\rho^2_{Y\cdot\mathbf{X}}}}\left(\frac{\sigma_{X_j}}{\sigma_Y}\right)\sqrt{N} = \sqrt{\frac{\rho^2_{Y\cdot\mathbf{X}}-\rho^2_{Y\cdot\mathbf{X}_{-j}}}{1-\rho^2_{Y\cdot\mathbf{X}}}}\sqrt{N}. \quad (31)$$

This implies that sample size is given as

$$N = \left(\frac{\lambda_j}{\beta_j}\right)^2\left(\frac{1-\rho^2_{Y\cdot\mathbf{X}}}{1-\rho^2_{X_j\cdot\mathbf{X}_{-j}}}\right)\left(\frac{\sigma^2_Y}{\sigma^2_{X_j}}\right) = \lambda_j^2\left(\frac{1-\rho^2_{Y\cdot\mathbf{X}}}{\rho^2_{Y\cdot\mathbf{X}}-\rho^2_{Y\cdot\mathbf{X}_{-j}}}\right). \quad (32)$$

Thus, given the population parameters and λ_j, sample size can be determined. However, in order to plan an appropriate sample size, once the population parameters and the desired degree of certainty are specified, λ_j is the only unknown parameter because N is unknown. If the λ_j that satisfies a desired degree of power can be determined, then the equation can be solved for necessary sample size.

Power is based on λ_j and the degrees of freedom, which in turn are based on N. Different values of N can be used to update λ_j and the degrees of freedom until the desired level of power is achieved for the test that $\beta_j = 0$. As before, this process can be implemented with tabled values (e.g. Kraemer & Thiemann, 1987; Cohen, 1988; Lipsey, 1990; Murphy & Myors, 1998; see also Maxwell, 2000 for a comprehensive review) or with a noncentral t (or F) computer routine.

The ss.power.reg.coef() function from MBESS can be used to determine sample size for a targeted regression coefficient. For example, suppose a researcher believes that $\rho^2_{Y\cdot\mathbf{X}} = 0.40$ and when the regressor of interest is removed $\rho^2_{Y\cdot\mathbf{X}_{-j}} = 0.30$. Thus, the regressor of interest uniquely explains 0.10 of the proportion of variance in the criterion variable. Although several possibilities exist, the basic way that the ss.power.reg.coef()

function from MBESS can be specified is as follows:

```
R > ss.power.reg.coef(Rho2.Y_X = 0.40,
   Rho2.Y_X.without.j = 0.30, p = 5,
   desired.power = 0.80, alpha.level = 0.05)
```

where Rho2.Y_X is the population squared multiple correlation coefficient predicting Y from $\mathbf{X}$ and Rho2.Y_X.without.j is the population squared multiple correlation coefficient predicting Y from $\mathbf{X}_{-j}$. The necessary sample size in this example is 50.

SAMPLE SIZE PLANNING FOR MULTIPLE REGRESSION GIVEN THE GOAL OF STATISTICAL ACCURACY

AIPE for the omnibus effect in multiple regression: Obtaining a narrow confidence interval for the population squared multiple correlation coefficient

The way in which sample size can be determined in order for the expected width of the confidence interval for $\rho^2_{Y\cdot\mathbf{X}}$ to be sufficiently narrow is quite involved. The method is computationally tedious and can only be carried out with the use of an iterative computer routine that uses noncentral F-distributions. As elsewhere in the chapter, we have restricted the discussion to regressors that are fixed. The case of random regressors is fully developed in Kelley (2006)[19]. It should be noted that two methods are discussed. The first method discussed provides necessary sample size for the expected confidence interval width. The confidence interval width is a random variable that will vary from sample to sample. A modified approach will also be discussed so that the width will be sufficiently narrow with no less than some specified degree of certainty.

The values that must be specified in order to determine the necessary sample size given an expected confidence interval width that is sufficiently narrow are $\rho^2_{Y\cdot\mathbf{X}}$, p, and α. The idea is to first use $\rho^2_{Y\cdot\mathbf{X}}$, p, and α in order to determine the width of the confidence interval given some minimal sample size. If the width is larger than desired, the current estimate of N is incremented by 1 and then the expected width is determined again. This iterative process continues until the sample size is just large enough so that the expected confidence interval width is sufficiently narrow. Two caveats with such an approach arise: $R^2_{Y\cdot\mathbf{X}}$ is a positively biased estimate of $\rho^2_{Y\cdot\mathbf{X}}$and the sample size calculated is only for the expected width.

Even though $R^2_{Y\cdot\mathbf{X}}$ is the sample estimate of $\rho^2_{Y\cdot\mathbf{X}}$, $R^2_{Y\cdot\mathbf{X}}$ is positively biased. However, the confidence limits for $\rho^2_{Y\cdot\mathbf{X}}$, and thus its width, are based on $R^2_{Y\cdot\mathbf{X}}$. Even though the bias of $R^2_{Y\cdot\mathbf{X}}$ decreases as N increases, holding everything else constant, basing the necessary sample size on $\rho^2_{Y\cdot\mathbf{X}}$ directly would lead to inappropriate estimates of necessary sample size because the width of the computed confidence interval in part depends on $R^2_{Y\cdot\mathbf{X}}$. The way in which this complication is overcome is by using the expected value of $R^2_{Y\cdot\mathbf{X}}$ in place of $\rho^2_{Y\cdot\mathbf{X}}$. The expected value of $R^2_{Y\cdot\mathrm{X}}$ given ρ^2_{Y}, N, and p when regressors are fixed does not have a known derivation. However, the expected value of $R^2_{Y\cdot\mathrm{X}}$ given $\rho^2_{Y\cdot\mathbf{X}}$, N, and p when regressors are random is known and is used as an approximation to the case where predictors are fixed, which is given as

$$\begin{aligned}\mathrm{E}\left[R^2_{Y\cdot\mathbf{X}}\middle|\left(\rho^2_{Y\cdot\mathbf{X}}, N, p\right)\right] &= 1 - \frac{N-p-1}{N-1}\left(1-\rho^2_{Y\cdot\mathbf{X}}\right) \\ &\quad\times \mathrm{H}\left[1;1;\frac{N+1}{2};\rho^2_{Y\cdot\mathbf{X}}\right], \qquad (33)\end{aligned}$$

where H is the hypergeometric function (Stuart et al., 1999, section 28.32; Johnson et al., 1995).

The sample size procedure is based on the expected value of $R^2_{Y\cdot\mathbf{X}}$ because it is the value expected to be obtained in the study. For a given α, p, and N, the confidence interval width depends only on $R^2_{Y\cdot\mathbf{X}}$. Thus, the expected confidence interval width can be

determined by forming a confidence interval with the expected $R^2_{Y\cdot\mathbf{X}}$. The expected confidence interval width can be made sufficiently narrow by increasing sample size, implying that the expected value of $R^2_{Y\cdot\mathbf{X}}$ changes, until the expected confidence interval width is equal to or just narrower than the desired width. Once the sample size is found so that the expected confidence interval width is sufficiently narrow, using the sample size in a study will ensure that the expected width of the confidence interval will be sufficiently narrow.

For example, suppose a researcher wishes to determine necessary sample size so that the expected width of a 95% confidence interval for $\mathrm{P}^2_{Y\cdot\mathbf{X}}$ is 0.20 for 5 regressor variables in a situation where $\mathrm{P}^2_{Y\cdot\mathbf{X}} = 0.5$. The **ss.aipe.R2()** function from MBESS would be used as

```
R > ss.aipe.R2(Population.R2 = 0.50,
   conf.level = 0.95, width = 0.20, p = 5,
   Random.Regressors=FALSE),
```

which returns a necessary sample size of 152. Thus, using a sample size of 152 would provide an expected width for the confidence interval of 0.20.

Since the width of the confidence interval is a random variable, having a sample size such that the expected width is sufficiently narrow does not ensure that any particular sample will have a confidence interval that is sufficiently narrow (e.g. see Hahn & Meeker, 1991, or Kupper & Hafner, 1989, for a discussion of these issues in simpler situations). What can be done is to specify some desired degree of certainty that the obtained confidence interval will in fact be sufficiently narrow. The way in which this additional step proceeds is by using the sample size obtained from the previously discussed procedure and from two $\gamma 100\%$ one-sided confidence intervals for $\mathrm{P}^2_{Y\cdot\mathbf{X}}$, where γ is the desired degree of certainty that the obtained interval will be sufficiently narrow. The limits from the $\gamma 100\%$ confidence intervals are then used to plan an appropriate sample size as before, but now using the confidence limits in place of $\mathrm{P}^2_{Y\cdot\mathbf{X}}$ from the first procedure. The rationale of this approach is to base the sample size procedure on the largest and smallest plausible value for the obtained $R^2_{Y\cdot\mathbf{X}}$ based on the original sample size and the degree of certainty specified.

The reason the upper and lower confidence limits are used is because, unlike many effects where the larger the noncentrality parameter the wider the confidence interval (holding everything else constant), there is a nonmonotonic relationship between $R^2_{Y\cdot\mathbf{X}}$ and the confidence interval width. Depending on the particular situation, a larger sample size may be necessitated by the lower limit or the upper limit from the two $\gamma 100\%$ one-sided confidence limits (or a value in between). The relationship between $R^2_{Y\cdot\mathbf{X}}$ and the corresponding confidence interval width is illustrated in Figure 11.1 for 95% confidence intervals where $p = 5$ and $N = 100$. The lack of monotonicity between the size of $R^2_{Y\cdot\mathbf{X}}$ and the confidence interval width implies that, depending on the particular situation, the upper limit, the lower limit, or values in-between the two one-sided $\gamma 100\%$ confidence interval limits will yield wider confidence intervals for $\mathrm{P}^2_{Y\cdot\mathbf{X}}$. Even though Figure 11.1 is helpful to illustrate why upper and lower limits are required, recall that the procedure always uses the expected value of $R^2_{Y\cdot\mathbf{X}}$. Thus, an analog to the figure presented, and what is actually used in the procedure, is one where the values on the ordinate are a function of basing confidence interval width on the expected values of $R^2_{Y\cdot\mathbf{X}}$ for corresponding values of $\mathrm{P}^2_{Y\cdot\mathbf{X}}$.

Two issues arise when basing the sample size procedure on limits from the $\gamma 100\%$ one-sided confidence intervals. First, it is possible that the point estimate itself requires a larger sample size than either of the confidence limits (e.g. suppose the corresponding point estimate is 0.35 from the figure). Second, the maximum confidence interval width could be between the limits (e.g. suppose the corresponding confidence limits are 0.2 and 0.6 from the figure). To ensure that an appropriate sample size is determined, an

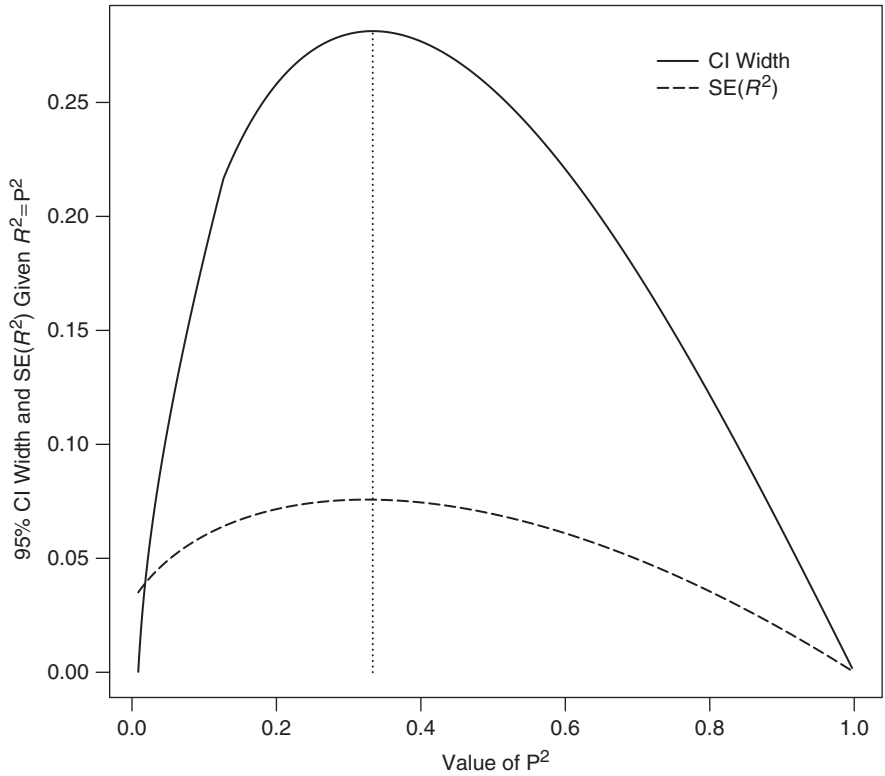

Figure 11.1 Relationship between the observed width of the 95% confidence interval for the population squared multiple correlation coefficient ($P^2_{Y \cdot \mathbf{X}}$) as a function of the observed squared multiple correlation coefficient ($R^2_{Y \cdot \mathbf{X}}$) when the total sample size is 100 and there are five regressors

optimization routine is used to determine if there is a value within the confidence limits that leads to a wider confidence interval than either of the limits. If not, the larger of the two sample sizes is used. If so, if the value that leads to the widest confidence interval is the value on which the original sample size is based, then the original sample size is used. If it is some other value between the confidence limits, then the $R^2_{Y \cdot \mathbf{X}}$ value that occurs with probability $(1 - \gamma)/2$ less than the value leading to the maximum confidence width and the value that occurs with probability $(1 - \gamma)/2$ more than the maximum confidence width are used. The probabilities are determined from the appropriate noncentral F-distributions. Of the contending sample sizes, the largest one is used. Doing so ensures that no less than $\gamma 100\%$ of the confidence intervals widths will be sufficiently narrow (Kelley, 2006, provides more detail on the procedure in the case of random regressors). It is important to remember that at every stage, the expected value of $R^2_{Y \cdot \mathbf{X}}$ is used based on the particular population value. Depending on the particular situation, incorporating a degree of certainty parameter can yield only a small or a large increase in necessary sample size.

The method discussed in order to obtain a narrow confidence interval with some degree of certainty can be readily implemented with the `ss.aipe.R2()` function. Realizing that having only an expected width of 0.20 is not sufficient, further suppose that the researcher incorporates a 99% degree of certainty that the obtained confidence interval will be no wider than 0.20 units. The way in which the

ss.aipe.R2() function is used in order to ensure a desired degree of certainty of 0.99 is given as follows:

```
R > ss.aipe.R2(Population.R2 = 0.50,
   conf.level = 0.95, width = 0.20, p = 5,
   degree.of.certainty = 0.99,
   Random.Regressors = FALSE),
```

which yields a necessary sample size of 189.

AIPE for targeted effects in multiple regression: Obtaining a narrow confidence interval for the population regression coefficient

Recall that when regression coefficients are unstandardized, the way in which confidence intervals are obtained is based on the central t-distribution. However, confidence intervals based on standardized regression coefficients require the use of noncentral distributions (since ${}_s b_j$ is not a pivotal quantity). Thus, the appropriate procedures are different for the two scenarios. The first procedure discussed will be for unstandardized regression coefficients followed by a procedure for standardized regression coefficients.

AIPE for unstandardized regression coefficients

Kelley and Maxwell (2003) discussed AIPE for a targeted regression coefficient. We will base the present discussion largely on an updated account of that work in the context of unstandardized regression coefficients. Recall from Equation 26 that the confidence interval for β_j is straightforward to calculate given b_j, s_{b_j}(which is a function of N, p, $R^2_{Y\cdot\mathbf{X}}$, $R^2_{X_j\cdot\mathbf{X}_{-j}}$), N, p, and α. The population variance for the jth regression coefficient is given as

$$\sigma^2_{b_j} = \left(\frac{\left(1-\mathrm{P}^2_{Y\cdot\mathbf{X}}\right)}{\left(1-\mathrm{P}^2_{X_j\cdot\mathbf{X}_{-j}}\right)/(N-p-1)}\right)\left(\frac{\sigma^2_Y}{\sigma^2_{X_j}}\right). \tag{34}$$

Given $\sigma^2_{b_j}$, the sample size can be solved for, yielding the necessary sample size in order for the expected width to be sufficiently narrow:

$$N = \left(\frac{t_{(1-\alpha/2;N-p-1)}}{\omega/2}\right)^2 \left(\frac{1-\mathrm{P}^2_{Y\cdot\mathbf{X}}}{1-\mathrm{P}^2_{X_j\cdot\mathbf{X}_{-j}}}\right) \times \left(\frac{\sigma^2_Y}{\sigma^2_{X_j}}\right) + p + 1, \tag{35}$$

where ω is the desired full width of the confidence interval. A complication is that the desired N is implicitly involved on the right side of the equation since the degrees of freedom of the t-value depend on N. It is thus necessary to solve Equation 35 iteratively.

Because the confidence interval width is itself a random variable, obtained values of $s^2_{b_j}$ larger than the population value used in the calculation of N will lead to confidence intervals wider than desired. In order to avoid obtaining a confidence interval wider than desired, the $\gamma 100\%$ confidence limit for the standard error can be used in place of the population standard error when solving for N. The $\gamma 100\%$ upper confidence limit for the population standard error of the jth regression coefficient, based on a chi-square distribution with $N - p - 1$ degrees of freedom, can then be substituted for the population variance from Equation 34. Doing so will ensure that the obtained confidence interval will be sufficiently narrow no less than $\gamma 100\%$ of the time. Since the only way for a confidence interval to be wider than desired is to obtain a standard error larger than the population standard error, using the upper $\gamma 100\%$ confidence limit of the standard error will ensure that the confidence interval will be sufficiently narrow no less than $\gamma 100\%$ of the time.

The way in which the upper limit for the variance of the regression coefficient is determined is given as

$${}_\gamma\sigma^2_{b_j} = \frac{\left(1-\mathrm{P}^2_{Y\cdot\mathbf{X}}\right)}{\left(1-\mathrm{P}^2_{X_j\cdot\mathbf{X}_{-j}}\right)/(N-p-1)}\left(\frac{\sigma^2_Y}{\sigma^2_{X_j}}\right) \times \left(\frac{\chi^2_{(\gamma;N-1)}}{N-p-1}\right), \tag{36}$$

where $\chi^2_{(\gamma;N-1)}$ is the γth quantile from a χ^2 distribution with $N-1$ degrees of freedom and ${}_{\gamma}\sigma^2_{b_j}$ is the upper limit of the $\gamma 100\%$ confidence interval for $\sigma^2_{b_j}$. Substituting ${}_{\gamma}\sigma^2_{b_j}$ from Equation 36 for $\sigma^2_{b_j}$ from Equation 34 yields the modified sample size,

$$N_{\gamma} = \left(\frac{t_{(1-\alpha/2;N-p-1)}}{\omega/2}\right)^2 \left(\frac{1-\mathrm{P}^2_{Y\cdot\mathbf{X}}}{1-\mathrm{P}^2_{X_j\cdot\mathbf{X}_{-j}}}\right) \times \left(\frac{\sigma^2_Y}{\sigma^2_{X_j}}\right)\left(\frac{\chi^2_{(\gamma;N-1)}}{N-p-1}\right) + p + 1, \tag{37}$$

where N_{γ} is the modified sample size so that there is $\gamma 100\%$ certainty that the obtained confidence interval will be sufficiently narrow.

The methods discussed can be readily implemented with the MBESS R function ss.aipe.reg.coef(). Suppose that $\mathrm{P}^2_{Y\cdot\mathbf{X}} = 0.50$ and $\mathrm{P}^2_{X_j\cdot\mathbf{X}_{-j}} = 0.20$, $\sigma^2_Y = 50$, $\sigma^2_{X_j} = 5$, $p = 5$, and $\beta_j = 3$. Further suppose that the desired width for the 95% confidence interval is 2 for the regressor of primary importance (the estimate plus and minus 1 unit). The way in which the ss.aipe.reg.coef() function can be used is given as

```
R > ss.aipe.reg.coef(Rho2.Y_X = 0.5,
  Rho2.j_X.without.j = 0.2, p = 5, b.j = 3,
  width = 2, sigma.Y = 50, sigma.X = 5,
  conf.level = 0.95.)
```

with the result of the function being 250. Further suppose that the researcher would like to be 85% certain that the 95% confidence interval is no larger than 2 units wide. The modified sample size can be obtained by specifying the degree of certainty parameter:

```
R > ss.aipe.reg.coef(Rho2.Y_X = 0.5,
  Rho2.j_X.without.j = 0.2, p = 5, b.j = 3,
  width = 2, sigma.Y = 50, sigma.X = 5,
  conf.level = 0.95,
  degree of certainty = 0.85)
```

which yields a necessary sample size of 278.

AIPE for standardized regression coefficients

Similar to the sample size for the expected confidence interval width being sufficiently narrow for an unstandardized regression coefficient, the sample size necessary in order for the expected width of a noncentral confidence interval for ${}_s\beta_j$ can be solved iteratively. Because the critical value cannot be written analytically since it is based on a noncentral t-distribution, the iterative nature for the necessary sample size of the standardized regression coefficients must also include a step for determining the expected confidence interval width given the particular sample size. Thus, the iterative nature necessary to determine the expected width is more difficult for standardized regression coefficients than it is for their unstandardized counterparts due to the necessary employment of the noncentral t-distribution. Although this requires a great deal more work in the actual algorithm to determine sample size, there is no conceptual difference compared to the method for the unstandardized regression coefficient.

The method has been implemented in the ss.aipe.reg.coef() function from MBESS when Noncentral=TRUE has been specified. For the situation described in the previous section, sample size for the standardized analog can be obtained as

```
R > ss.aipe.reg.coef(Rho2.Y_X = 0.5,
  Rho2.j_X.without.j = 0.2, p = 5, b.j = 0.3,
  width = 0.2, sigma.Y = 1, sigma.X = 1,
  conf.level = 0.95, Noncentral = TRUE)
```

which yields a necessary sample size of 264.

As in the unstandardized case, the confidence interval width is itself a random variable. At the present time, there has not been a satisfactory method developed for determining necessary sample size for confidence intervals for ${}_s\beta_j$ that incorporates a desired degree of certainty. The complication in developing such a method stems from the fact that the noncentrality parameter is based on two parameters: ${}_s\beta_j$ and $\sigma^2_{s_{b_j}}$.

Thus, an analog for the way a desired degree of certainty is incorporated into the unstandardized regression coefficient, where the confidence interval width depends on only one parameter, $\sigma^2_{b_j}$, is necessarily more difficult in the standardized case. Even though we believe that a method can and will be developed, at the present time a brute-force trial and error simulation-based method can be implemented in order to plan an appropriate necessary sample size. Such an approach would proceed by specifying the population parameters and simulating data based on a particular sample size. From there, confidence intervals could be performed for standardized regression coefficients as previously discussed. The proportion of confidence intervals that are less than the desired width can be determined for different sample size values. This could be done until the minimum sample size is found that yields no less than the desired degree of certainty specified.

The function ss.aipe.reg.coef.sensitivity() contained in the MBESS R package can be used to determine the appropriate sample size as well as perform general sensitivity analyses. When an estimated set of population parameters is specified (that differs from the true set), the sample size used is based on the estimated values, but the simulation is conducted based on the properties of the true set of parameter values. This allows one to perform a sensitivity analysis, where the effects of mis-specifying population parameters by varying amounts on the typical width and the percentage of confidence intervals narrower/wider than desired can be evaluated. Alternatively, a specific sample size can be used in order to evaluate the properties of the situation described by the true set of parameter values at the specified value of sample size. Using the specified sample size approach, one can run the simulation with different values of sample size until the percentage of confidence interval widths less than the desired width is equal to the degree of certainty of interest. Although generally more time consuming, the brute force method described works very well when one wants to incorporate a desired degree of certainty parameter into the sample size procedure for standardized regression coefficients and for sensitivity analyses in general[20].

DISCUSSION

In the context of multiple regression, the question 'What size sample should I use?' does not have a simple answer. As this chapter has demonstrated, the answer is best addressed with the two-by-two conceptualization presented in Table 11.1. Specifically, the sample size that should be used depends on the goals of the study. If the goal is for the overall fit of the model, then interest concerns $\rho^2_{Y\cdot\mathbf{X}}$; if the goal is for a targeted effect, then interest concerns β_j(or ${}_s\beta_j$). Of course, both $\rho^2_{Y\cdot\mathbf{X}}$ and β_j(or ${}_s\beta_j$) might be of interest, which implies that the larger of the two sample sizes from the situations of interest should be used.

However, identifying only that one is interested in $\rho^2_{Y\cdot\mathbf{X}}$ and/or β_j(or ${}_s\beta_j$) is still not enough to determine the necessary sample size. It is also necessary to determine if the goal is to reject the null hypothesis that the effect is zero in the population or if the goal is to obtain an accurate parameter estimate via a narrow confidence interval for the population parameter (possibly both). In multiple regression, although the idea is much more general, choosing an adequate sample size is not generally possible until a particular cell in Table 11.1 has been identified as the scenario of interest. Once the particular scenario from the two-by-two conceptualization has been determined, then and only then can an appropriate sample size be planned (recall Point f from the designing research studies list in the introduction of the chapter).

Even after the scenario has been determined, it is still necessary to use an appropriate value of an effect size parameter. One thing that has been conspicuously absent from the chapter is ways to choose an appropriate value for the effect size parameter so that all the sample size procedures can

be implemented. The effect size has been termed the 'problematic parameter' due to the difficulty in estimating this unknown but necessary quantity (Lipsey, 1990). Options include basing population values on values obtained in previous research, possibly using meta-analytic techniques, performing a pilot study to estimate the necessary population quantities, or basing the population values on a reasonable exchangeable correlation structure. An exchangeable correlation structure is one where the correlation between each regressor and the criterion is the same and the correlation among the regressors is the same (but the two correlation values may be different; Maxwell, 2000). Even though this may seem simplistic, it is often a reasonable alternative unless obvious reasons exist for why it should not be used (Maxwell, 2000; see also Green, 1977). Given the difficulty of estimating the effect size parameter, combined with the nonlinear relationship between the necessary sample size and the desired degree of power or accuracy, sensitivity analyses are almost always helpful.

The chapter has made use of the Open Source and freely available computer package MBESS for the R statistical language and environment. We believe that the user-friendly functions contained in this package will be helpful for researchers planning sample size for multiple regression from any of the cells within Table 11.1. Alternatively, when there are multiple goals, choosing the larger of the necessary sample sizes is suggested as a way to achieve the multiple goals.

It is our hope that this chapter has been helpful in synthesizing four very different methods of planning sample size. The correct choice, of course, depends on the goal(s) of the researcher. Before determining sample size, a necessary but not a sufficient task is to clearly identify the particular question of interest that the study would ideally accomplish. Unless the question of interest is clearly identified, sample size cannot be adequately planned. Perhaps the best answer to the question 'What size sample should I use?' is, 'Well, it depends.'

NOTES

1 One important complication not addressed in this chapter is the total financial cost of conducting a study. Some studies may require a necessary sample size so large that the cost of conducting the study with that sample size becomes prohibitively expensive (e.g. Kraemer, 1991; Allison et al., 1997).

2 With regards to the statistical power or accuracy of regression coefficients, we have approached the chapter as if interest is restricted to *either* the omnibus effect *or* a single targeted regression coefficient. Of course, a researcher might be interested in more than one regression coefficient or potentially all regression coefficients. When interest includes more than one regression coefficient or all regression coefficients, issues of multiple and simultaneous inference become important. These issues are beyond the scope of the present chapter and are not discussed.

3 R and MBESS, along with their respective manuals, can be downloaded from the following Internet address: http://www.cran.r-project.org/.

4 Some sources state that power is a function of only three things, but in those cases the work generally refers to the standardized effect size, which involves both the (unstandardized) effect size and the model error variance. An example of such a situation is when planning sample size to detect the difference between two independent group means. Either the mean difference and the common variance or the standardized mean difference, which is defined as the mean difference divided by the square root of the common variance, can be specified.

5 A Type I error occurs when the null hypothesis is true but the null hypothesis is rejected (this occurs with probability α). A Type II error occurs when the null hypothesis is false but the null hypothesis fails to be rejected.

6 At times the data analysis procedure can be modified so as to reduce the model error variance yet still address the same research question, which potentially increases power and/or accuracy. For example, analysis of covariance can be used instead of an analysis of variance in a randomized design. The same question is addressed (are there differences among the population group means?), yet the model error variance is reduced by an amount related to the squared correlation between the covariate and the dependent variable (e.g. Huitema, 1980; Cox & McCullagh, 1982; Maxwell & Delaney, 2004).

7 It should be noted that the terms accuracy and precision have often been (incorrectly) used synonymously in the literature, which has at times caused confusion (Stallings & Gillmore, 1971). We believe the definition used here is optimal, in the sense that accuracy is clearly a function of precision and bias. The term accuracy in parameter estimation, the term we use for planning sample size with the desire to have a narrow confidence interval, is also thought to be

ideal, as it conveys the goal of achieving a parameter estimate that is close to its population value.

8 As an extreme example, suppose that regardless of the observed data, a researcher always estimates the parameter to be a value that corresponds to an a priori theory irrespective of any observed data. In such a case there would be a high degree of precision but the accuracy would likely be poor due to the effect of bias in the estimation procedure unless the theory is perfect. Precision is thus a necessary but not a sufficient condition for achieving accurate parameter estimates.

9 A counter example is the Cauchy distribution, where the precision of the location estimate is the same regardless of the sample size used to estimate it (Stuart et al., 1994, pp. 2–3).

10 Some population parameters are typically estimated with biased estimators but have exact confidence interval procedures. Even though the estimator is biased, the point estimate may be necessary for calculation of the (exact) confidence interval, where the values within the interval represent plausible values and will contain the parameter with $(1 - \alpha)100\%$ confidence. Many such population parameters also have unbiased (or more unbiased) estimators. Examples include the standardized mean difference (e.g. Hedges & Olkin, 1985), the squared multiple correlation coefficient (e.g. Algona & Olenek, 2000), the standard deviation (e.g. Hays, 1994, for the confidence interval method and Boltzmann, 1950, for the unbiased estimate), and the coefficient of variation (e.g. Johnson & Welch, 1940 for the confidence interval method and Social & Baumann, 1980, for its nearly unbiased estimate). A strategy in such cases is to report the exact confidence interval and the unbiased estimate of the population parameter.

11 The direction of an effect is known if the upper and lower limits of the confidence interval are both in the same direction (i.e. both are positive or both are negative). Furthermore, the confidence limits determine whether or not a particular null hypothesis (such as zero) can be rejected. Confidence limits provide the same information as an infinite set of hypothesis tests. The values within the confidence limits are the values of the null hypothesis that would not be rejected. The values outside of the confidence limits are the values of the null hypothesis that would be rejected.

12 The term 'regressors' has been used throughout the chapter as a generic term for the A_x variables. A regressor variable is termed independent, explanatory, predictor, or concomitant variable in other contexts. The term criterion is used as a generic term for the Y variable. The criterion variable is termed dependent, outcome, or predicted variable in other contexts.

13 Notice that the regressor variables (i.e. the A_x variables) are not italicized in any of the equations. This is because we will regard the regressors as fixed throughout the chapter. Even though the distinction between fixed and random regressors is not often made in applied work, the sampling distribution of an estimated regression coefficient tends to depend on whether the regressors are fixed or random (e.g. Stuart et al., 1999; Rancher, 2000). Many applications of multiple regression implicitly or explicitly take the view 'given this **X**' so that the **X** variables can be considered fixed for purposes of the study (e.g. O'Brien & Mueller, 1993, p. 23). O'Brien and Mueller (1993) make the argument that the distinction is not important in the context of sample size planning for power in multiple regression by stating that 'the practical discrepancy between the two approaches disappears as the sample size increases' (p. 23). O'Brien and Mueller (1993) go on to say that 'because the population parameters are conjectures or estimates, strict numerical accuracy of the power computations is usually not critical' (p. 23). We will say more about the distinction between fixed and random regressors elsewhere in the chapter.

14 We use both standardized and unstandardized regression coefficients in various parts of the chapter. Observed standardized regression coefficients have at times been referred to as 'beta weights' in the behavioral and educational sciences. We will use β_j to represent the unstandardized population regression coefficient of variable j with b_j as its estimate. We use ${}_s\beta_j$ to represent the standardized population regression coefficient of variable j with ${}_sb_j$ as its estimate.

15 Notice that we have not used the standard general linear model equations, where the intercept is contained within β and X contains a vector of ones for the intercept. The notation used here is equivalent to the standard general linear model equations, but it is especially helpful for presenting the necessary information for each of the four approaches to sample size planning for multiple regression.

16 Throughout the chapter, multiple correlation coefficients will be denoted with a subscript that identifies the variable being predicted separated by a dot from one or more regressor variables. Thus, the criterion variable is on the left of the dot and the regressor variable(s) are to the right of the dot, where the dot can literally be read as 'regressed on,' 'predicted from' or 'explained by.'

17 A mean-shifted central distribution is one that follows a central distribution after subtracting the population value. For example, when comparing two independent group means, if there is a population mean difference between the two groups a priori, then that difference can be subtracted from the observed difference: $(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)$, where $\bar{Y}_1$ and $\bar{Y}_2$ are the observed means for groups one and two, respectively, and μ_1 and μ_2 are the population means for groups 1 and 2,respectively.

18 The illustrative data from Holzinger and Swineford's (1939) Grant-White School data (available in MBESS), where the criterion variable, *total*

score (the sum of all of the 26 measured variables included in the dataset), is modeled as a function of the regressor variables *flags*, *wordm*, *addition*, *object*, and *series*. The standardized and unstandardized regression coefficients, presented in the next section, are for the *series* variable, which was a test that measured students' ability to complete mathematical/numeric series. Notice that the squared multiple correlation coefficient is quite large by most behavioral, educational, and social science standards. The large squared multiple correlation coefficient is because the dependent variable is a sum of five positively correlated measures, where the zero-order correlations among the measures tended to be large.

19 Even though only fixed regressors are discussed in the chapter, The ss.aipe.R2() function in MBESS can be used for regressors that are fixed *or* random by specifying Random.Predictors=TRUE (for random predictors) or Random.Predictors=FALSE (for fixed regressors).

20 In addition to the ss.aipe.reg.coef.sensitivity() function described, there is also a ss.power.reg.coef.sensitivity() function that allows the effects of parameter mis-specification or selected sample size to be specified in order to assess empirical power, and other properties, for a targeted regression coefficient. These functions for confidence interval width and power have analogs for omnibus effect with the ss.aipe.R2.sensitivity() and the ss.power.R2.sensitivity() functions.

REFERENCES

Algona, J., & Olenek, S. (2000). Determining sample size for accurate estimation of the squared multiple correlation coefficient. *Multivariate Behavioral Research*, *35*, 119–136.

Allison, D. B., Allison, R. L., Faith, M. S., Paultre, F., & F. X. Pi-Sunyer. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods*, 2, 20–33.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423–437.

Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility*. Newbury Park, CA: Sage Publications.

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

Cohen, J. (1994, December). The earth is round ($p < 0.05$). *American Psychologist*, *49*, 997–1003.

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.

Cox, D. R., & McCullagh, P. (1982). Some aspects of analysis of covariance. *Biometrics*, 541–561.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532–574.

Darlington, R. B. (1990). *Regression and linear models*. New York, NY: McGraw-Hill.

Dunlap, W. P., Xin, X., & Myers, L. (2004). Computing aspects of power for multiple regression. *Behavior Research Methods, Instruments, & Computers*, *36*, 695–701.

Gatsonis, C., & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin*, *106*, 516–524.

Graybill, F. A. (1976). *Theory and application of the linear model*. Pacific Grove, CA: Brooks/Cole.

Green, B. F. (1977). Parameter sensitivity in multivariate methods. *Multivariate Behavioral Research*, *12*, 263–288.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, *26*, 499–510.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*(1), 15–24.

Hahn, G., & Meeker, W. (1991). *Statistical intervals: A guide for practitioners*. New York, NY: John Wiley & Sons, Inc.

Harris, R. J. (1997). Significance tests have their place. *Psychological Science*, *8*, 8–11.

Harris, R. J. (2001). *A primer of multivariate statistics* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Hays, W. L. (1994). *Statistics* (5th ed.). Belmont, CA: Wadsworth Publishing.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL:Academic Press.

Boltzmann, W. H. (1950). The unbiased estimate of the population variance and standard deviation. *American Journal of Psychology*, *63*, 615–617.

Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution.* Chicago, IL: The University of Chicago.

Huitema, B. E. (1980). *The analysis of covariance and alternatives*. New York, NY: Wiley.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (Vol. 2). New York, NY: John Wiley & Sons, Inc.

Johnson, N. L., & Welch, B. L. (1940). Applications of the noncentral *t*-distribution. *Biometrika*, *31*, 362–389.

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals for the standardized mean difference: Bootstrapping as an alternative to parametric confidence intervals. *Educational and Psychological Measurement*, *65*(1), 51–69.

Kelley, K. (2006). Sample size planning for the squared multiple correlation coefficient: Accuracy in parameter estimation via narrow confidence intervals. *Manuscript under review*.

Kelley, K. (2007). MBESS version 0.0.9: An R package. [computer software and manual]. Retrievable from http://www.cran.r-project.org/

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*, 305–321.

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision: Delineating methods of sample size planning. *Evaluation and the Health Professions*, *26*, 258–287.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11*, 363–385.

Kraemer, H., Gardner, C., Brooks, J. O., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods*, *3*, 23–31.

Kraemer, H. C. (1991). To increase power in randomized clinical trials without increasing sample size. *Psychopharmacology Bulletin*, *27*, 217–224.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects?* Beverly Hills, CA: Sage.

Kupper, L. L., & Hafner, K. B. (1989). How appropriate are popular sample size formulas? *American Statistician*, *43*, 101–105.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Maxwell, S. E. (2000). Sample size and multiple regression. *Psychological Methods*, *5*, 434–458.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective*. (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Meehl, P. E. (1967). Theory testing in psychology and in physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Mendoza, J. L., & Stafford, K. L. (2001). Confidence intervals, power calculations, and sample size estimation for the squared multiple correlation coefficient under the fixed and random regression models: A computer program and useful standard tables. *Educational and Psychological Measurement*, *61*, 650–667.

Mogie, M. (2004). In support of null hypothesis significance testing. *Proceedings of the Royal Society of London, Series B, Biology Letters*, *271*, 82–84.

Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy: A Reader*. Chicago, IL: Aldine Publishing Company.

Murphy, K. R., & Myors, B. (1998). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ: Erlbaum.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transaction of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *236*, 333–380.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

O'Brien, R., & Mueller, K. E. (1993). A unified approach to statistical power for *t*-tests to multivariate models. In L. Edwards (Ed.), *Applied analysis of variance in behavioral sciences* (pp. 297–344). New York, NY: Marcel Dekker.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). New York, NY: Harcourt Brace College Publishers.

R Development Core Team. (2007). R version 2.5.0: A language and environment for statistical computing [computer software and manual], R foundation for statistical computing.

Rancher, A. C. (2000). *Linear models in statistics*. New York, NY: John Wiley & Sons, Inc.

Rosenthal, R. (1993). Cumulative evidence. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 519–559). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*(5), 646–656.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416–428.

Rozeboom, W. W. (1966). *Foundations of the theory of prediction*. Homewood, IL: The Dorsey Press.

Sampson, A. R. (1974). A tale of two regressions. *Journal of the American Statistical Association*, *69*, 682–689.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115–129.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316.

Serlin, R., & Lapsley, D. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*, 73–83.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of methodological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *Methodological and quantitative issues in the analysis of psychological data* (pp. 199–228). Mahwah, NJ: Lawrence Earlbaum Associates.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*, 605–632.

Smithson, M. (2003). *Confidence intervals*. Thousand Oaks, CA: Sage Publications.

Social, R. R., & Baumann, C. A. (1980). Significance tests for coefficients of variation and variability profiles. *Systematic Zoology*, *29*, 50–66.

Stallings, W. M., & Gillmore, G. M. (1971). A note on 'accuracy' and 'precision'. *Journal of Educational Measurement*, *8*, 127–129.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*, 164–182.

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

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

Stuart, A., & Ord, J. K. (1994). Kendall's advanced theory of statistics: Distribution theory (6th ed.). New York, NY: John Wiley & Sons.

Stuart, A., Ord, J. K., & Arnold, S. (1999). *Kendall's advanced theory of statistics: Classical inference and the linear model* (6th ed., Vol. 2A). New York, NY: Oxford University Press.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25–32.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *4*(2), 212–213.