

Multiple Correlation: Exact Power and Sample Size Calculations

Constantine Gatsonis

Department of Biostatistics, Harvard School of Public Health

Allan R. Sampson

Department of Mathematics and Statistics,
University of Pittsburgh

This article discusses power and sample size calculations for observational studies in which the values of the independent variables cannot be fixed in advance but are themselves outcomes of the study. It reviews the mathematical framework applicable when a multivariate normal distribution can be assumed and describes a method for calculating exact power and sample sizes using a series expansion for the distribution of the multiple correlation coefficient. A table of exact sample sizes for level .05 tests is provided. Approximations to the exact power are discussed, most notably those of Cohen (1977). A rigorous justification of Cohen's approximations is given. Comparisons with exact answers show that the approximations are quite accurate in many situations of practical interest. More extensive tables and a computer program for exact calculations can be obtained from the authors.

It is common in the behavioral and social sciences to have studies in which the multiple variables for each experimental unit cannot be controlled and are only available after observation. In these studies, the questions of interest are oftentimes addressed via the multivariate statistical techniques of regression and correlation analysis. Briefly stated, the goal is to study the relation of a dependent variable Y to p independent factors $X(1), \dots, X(p)$, which are collectively denoted in this article by \mathbf{X} . Inferences are based on a set of N observations, $(Y_1, \mathbf{X}_1), \dots, (Y_N, \mathbf{X}_N)$. A good general review of relevant statistical methods for observational studies is given by McKinlay (1975).

Throughout this article we assume that the $p + 1$ variables $Y, X(1), \dots, X(p)$ have a joint multivariate normal distribution. Traditionally, there are two points of view for approaching the statistical modeling and analysis of such multivariate studies, namely, *unconditional* and *conditional*. The former approach treats the N observations of the $p + 1$ variables, $Y, X(1), \dots, X(p)$ as a random sample from the underlying multivariate normal distribution and accounts for the consequent variability in all analyses. The latter approach treats the observations on the independent variables as fixed and known; thus, the only variability in the model pertains to Y , the distribution of which depends on the specific values observed for the \mathbf{X} s.

These two approaches are known to lead to mathematical

formulations that share a common nomenclature and give parallel sets of results (e.g., Sampson, 1974). The conceptual difference between them is primarily one of interpretation and generalizability of the conclusions. The unconditional approach recognizes and accounts for the extra variability stemming from the fact that, in another replication of the same experiment, different values for the independent variables will be obtained. The conditional approach is not concerned with this extra variability. From a theoretical perspective, the results of a conditional analysis would be specific to the particular values of the independent variables that are observed. Thus, in settings in which the independent variables are themselves outcomes of the study, the appropriate approach is the unconditional one. The conditional approach is only appropriate for settings in which the independent variables are preset by the experimenter.

In practice, the distinction between the two approaches becomes important when power and sample size calculations are to be made. These calculations are carried out during the planning stages of the study, when the values of the independent variables are not available.

In the unconditional approach, one would typically begin by setting a minimal desirable level of dependence between Y and $X(1), \dots, X(p)$, expressed in terms of the underlying population multiple correlation coefficient. (When studying the relationship of Y to a subset of $X(1), \dots, X(p)$, keeping the remaining independent variables fixed, the multiple partial correlation coefficient is used.) Having set the alternative hypothesis, one would then proceed to compute a sample size that ensures a certain power for the test that will be used.

In the conditional approach, the parameters of interest are usually the regression coefficients. However, the power function depends also on the yet to be observed \mathbf{X} s and on the (conditional) variance of Y . Because of the nature of this dependence, it is difficult to specify practically meaningful alternative hypotheses in advance of seeing the data. Under some extra assumptions, which we discuss in the next section, it is still possible to index the alternative hypotheses using the multiple correlation coefficient for the underlying multivariate normal

We thank Hui Yu and Bob Wilkinson for their help in computer programming and the Department of Statistics, Carnegie Mellon University, for generous availability of the computing facilities. We also thank Joel Greenhouse, Saul Shiffman, and the two referees for commenting on an earlier draft of this article.

Constantine Gatsonis's research was supported by National Institute of Mental Health Grant No. MH 15758 at the Department of Statistics, Carnegie Mellon University. Allan R. Sampson's research was sponsored by the Air Force Office of Scientific Research under Contract No. AROSR-84-0113.

Correspondence concerning this article should be addressed to Constantine Gatsonis, Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115.

distribution. We note that this is essentially the approach taken in the well-known texts of Cohen and Cohen (1975) and Cohen (1977).

The primary aims of this article are twofold: One is to present an expository discussion of the effects of the two modeling viewpoints on power analyses and related sample size considerations. In doing so, we underscore the fact that because in our setting the values of the independent variables cannot be specified in advance, power calculations for these multivariate normal observational studies must be done in the unconditional framework. The other aim of our article is to present new tables and a computer program to implement the exact calculations for the unconditional approach. In the next section, we present the mathematical formulation for both the unconditional and conditional models and describe the corresponding power functions. The distinct features of the two models are highlighted in that section. Exact power calculations require the noncentral distribution of the sample multiple correlation coefficients. A series expansion, given by Lee (1972), is the basis for our numerical calculations. These calculations are described in the third section, in which we present a table of sample sizes for tests of level .05. (More complete tables, and an interactive computer program, can be obtained from us.) Several approximations to the unconditional power can be made. We briefly discuss them in the fourth section and devote most of that section to the presentation of a rigorous argument that justifies the power computations of Cohen (1977) as approximations to the unconditional power. In the same section, we report on a detailed comparison of the exact results with the approximations obtained via Cohen's tables. A brief discussion of our results constitutes the fifth section.

Unconditional Versus Conditional Power: Theory

Unconditional Viewpoint

Formally, we suppose that we are studying the relation between a random variable Y and a p -dimensional random vector \mathbf{X} , where (Y, \mathbf{X}) have a joint $(p + 1)$ -dimensional multivariate normal distribution with a mean of (μ_Y, μ_X) and a positive definite covariance matrix of

$$\begin{bmatrix} \sigma_Y^2 & \Sigma'_{YX} \\ \Sigma_{YX} & \Sigma_X \end{bmatrix}.$$

The population multiple correlation coefficient between Y and \mathbf{X} , ρ_{YX} , is defined (e.g., Anderson, 1984, Section 2.5.2) by

$$\rho_{YX} = [\Sigma'_{YX} \Sigma_X^{-1} \Sigma_{YX} / \sigma_Y^2]^{1/2}.$$

For the underlying normal population, the statement that the independent factors \mathbf{X} provide no information that can be used in the linear prediction of the dependent variable Y is equivalent to $\Sigma_{YX} = 0$, which in turn is equivalent to $\rho_{YX} = 0$. The standardly used estimator of ρ_{YX} , based on a random sample $(Y_1, \mathbf{X}_1), \dots, (Y_N, \mathbf{X}_N)$, is

$$R_{YX} = [\hat{\Sigma}'_{YX} \hat{\Sigma}_X^{-1} \hat{\Sigma}_{YX} / \hat{\sigma}_Y^2]^{1/2},$$

where

$$\hat{\sigma}_Y^2 = \sum_{i=1}^N (Y_i - \bar{Y})^2 / (N - 1),$$

$$\hat{\Sigma}_{YX} = \sum_{i=1}^N (Y_i - \bar{Y})(\mathbf{X}_i - \bar{\mathbf{X}})' / (N - 1),$$

and $\hat{\Sigma}_X = \sum_{i=1}^N (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' / (N - 1)$. A commonly used test statistic for the null hypothesis $\rho_{YX} = 0$ is

$$[(N - 1 - p) / p] R_{YX}^2 / (1 - R_{YX}^2). \tag{1}$$

In order to give a concise description of the distribution of this statistic, we first introduce the notion of a chi-square random variable, with random degrees of freedom. Specifically, if L is a random variable taking positive integer values, we denote by χ_L^2 the random variable with the following cumulative distribution function:

$$P(\chi_L^2 \leq x) = \sum_{l=0}^{\infty} P(\chi_l^2 \leq x) P(L = l).$$

In a computer simulation, in order to generate an observation from χ_L^2 , one would first generate an observation l from L and then an observation from χ_l^2 . It can be shown that if the population multiple correlation coefficient has an arbitrary value ρ_{YX} , the distribution of the test statistic in Equation 1 is the same as the distribution of the ratio

$$p^{-1} \chi_{p+2K}^2 / [(N - 1 - p)^{-1} \chi_{N-1-p}^2], \tag{2}$$

where the numerator and denominator are multiples of independent chi-square random variables, with random degrees of freedom for the numerator. The distribution of the random variable K is a negative binomial, that is,

$$\text{Prob}(K = k) = \Gamma(k + (N - 1) / 2) [\Gamma(k + 1)]^{-1} \times [\Gamma((N - 1) / 2)]^{-1} (\rho_{YX}^2)^k (1 - \rho_{YX}^2)^{(N-1)/2}. \tag{3}$$

For details of the proof, see Sampson (1974) or Anderson (1984, Section 4.4.3). Under the null hypothesis that $\rho_{YX} = 0$, the random variable K is always 0, and then the distribution of the statistic in Equation 1 is simply $F_{p, N-1-p}$. This leads to rejecting the null hypothesis at level α if $[(N - 1 - p) / p] R_{YX}^2 / (1 - R_{YX}^2) \geq F_{p, N-1-p}^{1-\alpha}$. The power of this test for a true multiple correlation ρ_{YX} is

$$\text{Prob}[p^{-1} \chi_{p+2K}^2 / [(N - 1 - p)^{-1} \chi_{N-1-p}^2] \geq F_{N-1-p}^{1-\alpha}], \tag{4}$$

where K has the negative binomial distribution given by Equation 3. The probability in Equation 4 is the unconditional power function of the test. It was derived by taking into account the randomness of the independent variables.

Another hypothesis-testing problem that can be handled by essentially the same unconditional analysis is testing whether a multiple partial correlation coefficient is zero. If we partition the p independent variables into two groups of size p_1 and $p - p_1$, respectively, denoted by $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, then we want to test for no relation between Y and $\mathbf{X}^{(1)}$, for fixed values of the $\mathbf{X}^{(2)}$ variables.

Let $\hat{\Sigma}_{YX^{(1)}}$, $\hat{\Sigma}_{YX^{(2)}}$, $\hat{\Sigma}_{X^{(1)}}$, $\hat{\Sigma}_{X^{(2)}}$, and $\hat{\Sigma}_X$ with respective dimensions $p_1 \times 1$, $(p - p_1) \times 1$, $p_1 \times p_1$, $p_1 \times (p - p_1)$,

and $(p - p_1) \times (p - p_1)$ be defined by the following matrix equality:

$$\begin{bmatrix} \hat{\sigma}_Y^2 & \hat{\Sigma}'_{YX^{(1)}} & \hat{\Sigma}'_{YX^{(2)}} \\ \hat{\Sigma}_{YX^{(1)}} & \hat{\Sigma}_{X^{(1)}} & \hat{\Sigma}'_{X^{(1)}X^{(2)}} \\ \hat{\Sigma}_{YX^{(2)}} & \hat{\Sigma}'_{X^{(1)}X^{(2)}} & \hat{\Sigma}_{X^{(2)}} \end{bmatrix} = \begin{bmatrix} \hat{\sigma}_Y^2 & \hat{\Sigma}'_{YX} \\ \hat{\Sigma}_{YX} & \hat{\Sigma}_X \end{bmatrix}.$$

Now define

$$\begin{bmatrix} \hat{\sigma}_{Y \cdot X^{(2)}}^2 & \hat{\Sigma}'_{YX^{(1)} \cdot X^{(2)}} \\ \hat{\Sigma}_{YX^{(1)} \cdot X^{(2)}} & \hat{\Sigma}_{X^{(1)} \cdot X^{(2)}} \end{bmatrix} = \begin{bmatrix} \hat{\sigma}_Y^2 & \hat{\Sigma}'_{YX^{(1)}} \\ \hat{\Sigma}_{YX^{(1)}} & \hat{\Sigma}_{X^{(1)}} \end{bmatrix} - \begin{bmatrix} \hat{\Sigma}'_{YX^{(2)}} \\ \hat{\Sigma}'_{X^{(1)}X^{(2)}} \end{bmatrix} \hat{\Sigma}_X^{-1} \begin{bmatrix} \hat{\Sigma}_{YX^{(2)}} \\ \hat{\Sigma}_{X^{(1)}X^{(2)}} \end{bmatrix}.$$

The sample multiple partial correlation coefficient between Y and $X^{(1)}$, controlling for $X^{(2)}$, is defined by

$$R_{YX^{(1)} \cdot X^{(2)}} = [\hat{\Sigma}'_{YX^{(1)} \cdot X^{(2)}} \hat{\Sigma}_{X^{(1)} \cdot X^{(2)}}^{-1} \hat{\Sigma}_{YX^{(1)} \cdot X^{(2)}} / \hat{\sigma}_{Y \cdot X^{(2)}}^2]^{1/2}.$$

Again, the statistic used for testing the null hypothesis, which is that the analogously defined population partial multiple correlation $\rho_{YX^{(1)} \cdot X^{(2)}}$ is zero, is $[(N - 1 - p)/p_1] R_{YX^{(1)} \cdot X^{(2)}}^2 / (1 - R_{YX^{(1)} \cdot X^{(2)}}^2)$. This test statistic has a null hypothesis distribution $F_{p_1, N-1-p}$ and an alternative distribution given by $p_1 X_{p_1+2K}^2 / [(N - p - 1)^{-1} X_{N-p-1}^2]$, with the distribution of K given by Equation 3 with ρ_{YX}^2 replaced by $\rho_{YX^{(1)} \cdot X^{(2)}}^2$ and $N - 1$ replaced by $N - 1 - p + p_1$.

Conditional Viewpoint

For the conditional analysis, we consider the values of X_1, \dots, X_N as given and work with the conditional distribution of Y_1, \dots, Y_N , given these X s. Specifically, we suppose that $(Y_1, \dots, Y_N)'$ has a multivariate normal distribution with mean vector $\alpha(1, \dots, 1)' + [X_1' : \dots : X_N']\beta$ and covariance matrix $\sigma^2 I$. The vector, β , of regression coefficients corresponds to $\Sigma_X^{-1} \Sigma_{YX}$ in the unconditional model; the intercept, α , corresponds to $\mu_Y - \Sigma'_{YX} \Sigma_X^{-1} \Sigma_{YX}$; and the variance, σ^2 , corresponds to $\sigma_Y^2 - \Sigma'_{YX} \Sigma_X^{-1} \Sigma_{YX}$. In the context of the conditional model, lack of linear association between the independent variables X and the dependent variable Y translates into $\beta = 0$. The usual (least squares) estimator, $\hat{\beta}$, of β can be written as $\hat{\beta} = \hat{\Sigma}_X^{-1} \hat{\Sigma}_{YX}$.

The F statistic for testing $\beta = 0$ is equal to the statistic used for testing $\rho_{YX} = 0$ in the unconditional model, namely, $F = [(N - 1 - p)/p] R_{YX}^2 / (1 - R_{YX}^2)$. For the conditional model, for an arbitrary value of β , this statistic has a noncentral F distribution, which can be represented in a form similar to Equation 2, namely, as the distribution of the ratio

$$p^{-1} X_{p+2K^*}^2 / [(N - 1 - p)^{-1} X_{N-1-p}^2], \tag{5}$$

where the numerator and denominator random variables are independent and K^* has a Poisson distribution with parameter $\lambda = \{\beta' [(N - 1) \hat{\Sigma}_X] \beta\} / (2\sigma^2)$. In terms of the multivariate normal parameters,

$$\lambda = [(N - 1)/2] \Sigma'_{YX} \Sigma_X^{-1} \hat{\Sigma}_X \Sigma_X^{-1} \Sigma_{YX} / (\sigma_Y^2 - \Sigma'_{YX} \Sigma_X^{-1} \Sigma_{YX}). \tag{6}$$

Thus, when designing a study from the conditional viewpoint, one must technically consider alternative values λ that depend not only on the values of the unknown parameters β and σ^2 but also on the observed value of $\hat{\Sigma}_X$, which is available only after the completion of the study. To circumvent this design difficulty, one can assume that $\hat{\Sigma}_X = \Sigma_X$, that is, that the observed covariance matrix of X will be equal to the population covariance matrix. It then follows from Equation 6 that $\lambda = [(N - 1)/2] \rho_{YX}^2 / (1 - \rho_{YX}^2)$, which is a population parameter and does not depend on the data. In planning the experiment now, the power function that should be used is

$$\text{Prob} \{ p^{-1} X_{p+2K^*}^2 / [(N - 1 - p)^{-1} X_{N-1-p}^2] \geq F_{p, N-1}^{\lambda, \sigma} \}, \tag{7}$$

where K^* has the Poisson distribution with

$$\lambda^* = [(N - 1)/2] \rho_{YX}^2 / (1 - \rho_{YX}^2). \tag{8}$$

A more detailed discussion in this framework of unconditional power, conditional power, and their interrelation can be found in Sampson (1974).

On the one hand, the assumption that $\hat{\Sigma}_X = \Sigma_X$ essentially discards the inherent randomness in the independent variables. On the other hand, it allows the conditional power to be computed for alternatives indexed by the underlying population multiple correlation coefficient. Thus, this assumption provides a rigorous basis for interpreting the power calculations of authors who work in the conditional model, notably Cohen and Cohen (1975) and Cohen (1977). Because of the reasons we noted earlier, we would argue that, for the multivariate normal observational study, the conditional power assumptions are untenable in the design phase and that experimental planning should be approached from the unconditional point of view. We hasten to add that this does not negate the usefulness of the calculations based on Cohen's tables. As we show in a later section, these calculations provide good approximations to the unconditional power.

Computations for the Unconditional Analysis

In this section we discuss the computation of exact sample size tables and give such a table for $\alpha = .05$. Although the computation of the conditional power can be accomplished using noncentral F tables with the noncentrality parameter $2\lambda^*$, the computation of the unconditional power requires the use of the distribution of $R_{YX}^2 / (1 - R_{YX}^2)$ or, equivalently, the distribution of R_{YX}^2 . Lee (1972) provided the following computationally effective form of the distribution of R_{YX}^2 :

$$\text{Prob} (R_{YX}^2 \leq R_0^2) = \{ \text{Be} (1; p/2, (N - 1 - p)/2) \}^{-1} (1 - \rho_{YX}^2)^{(N-1)/2} \sum_{k=0}^{\infty} C_K,$$

where

$$C_K = \{ ((N - 1)/2)_k \}^2 \rho_{YX}^{2k} \text{Be} \left(R_0^2; \frac{p}{2} + k, \frac{N - 1 - p}{2} \right) / [(p/2)_k k!],$$

where $(l)_0 = 1, (l)_j = (l)(l + 1) \dots (l + j - 1)$ and $\text{Be} (x; \alpha, \beta) = \int_0^x t^{\alpha-1} (1 - t)^{\beta-1} dt$ is the incomplete beta function. To

put Equation 4 into this computational form is to note that $[(N - 1 - p)/p]R_{YX}^2/(1 - R_{YX}^2) \geq F_{p, N-1-p}^{1-\alpha}$ is equivalent to $R_{YX}^2 \geq F_{p, N-1-p}^{1-\alpha}/[F_{p, N-1-p}^{1-\alpha} + (N - 1 - p)/p]$.

As was suggested by Lee (1972), the series was summed up to the term of order k , where k is the first integer for which

$$A_k = k^{-1}[(N - 1)/2 + k - 1]^2 \rho_{YX}^2 / (p/2 + k - 1) < 1$$

and

$$C_k / (1 - A_k) < 10^{-5}.$$

Thus the truncation error is less than 10^{-5} , and the estimated power is accurate to at least four decimal digits. The sample size corresponding to a specific choice of α , p , ρ_{YX} , and the desired power is the smallest integer guaranteeing that the actual power is at least equal to the desired one.

Table 1 presents these sample sizes for tests of level $\alpha = .05$, for values of $u = p$ ranging between 1 and 10, with $u = 15$ and $u = 20$ as well. For each value of u , ρ ranges between .1 and .9, and the desired power ranges between .25 and .99. To obtain the actual sample size N from Table 1, one must add $p + 1$ to the tabled value. Table 1 can be used for sample sizes calculations based on either the multiple correlation coefficient or the multiple partial correlation coefficient by choosing $\rho = \rho_{YX}$ or $\rho_{YX^{(1), X^{(2)}}}$, respectively. Specifically, it follows from the discussion in the second section that if the multiple partial correlation of Y with a subset of $p_1 X$ s is of interest, Table 1 can be used with $u = p_1$, and then Sample size = Table entry + $(p + 1)$.

A complete set of tables for $\alpha = .01, .05, .1$ and $u = 1, 2, \dots, 25, 28, 30, 32, 40, 50, 60, 80, 100$ is available from us. An interactive program is also available from us that carries out these sample size and power calculations. The program is written in FORTRAN and uses double-precision International Mathematical and Statistical Library (IMSL) subroutines for the inverse F distribution, as well as the logarithm, exponential, log-gamma, and incomplete beta functions.

Approximations to the Unconditional Power

An obvious first step in approximating the unconditional power of Equation 4 is suggested by the fact that the negative binomial random variable K in Equation 3 can be approximated by the Poisson random variable K^* with parameter λ^* given in Equation 8. The effect on this approximation is to allow the critical values of Equation 4 to be computable from a non-central F distribution. We note that this approximation is valid for large values of N and for small values of $\rho_{YX}^2/(1 - \rho_{YX}^2)$ (see Johnson & Kotz, 1969, p. 127). The mean and variance of K are $[(N - 1)/2]\rho_{YX}^2/(1 - \rho_{YX}^2)$ and $[(N - 1)/2] \times [\rho_{YX}^2/(1 - \rho_{YX}^2)](1 - \rho_{YX}^2)^{-1}$, respectively, and K^* has the same mean as K and $\text{var}(K^*) = (1 - \rho_{YX}^2) \text{var}(K)$, that is, $\text{var}(K^*) < \text{var}(K)$. The effects of these moment results on the quality of the approximation are not clear and require further distributional research.

Cohen's Approximation

The tables and methodology of Cohen (1977, chap. 9) can be viewed as providing another approximation to the unconditional power, where in part this approximation uses the one

noted previously. Specifically, Cohen's (1977, section 10.9) tables are based on the noncentral χ^2 distribution, but the exact details of the approximation are not discussed in his text. Strictly speaking, these tables are an approximation to the conditional power of Equation 7. However, a justification of their use for approximating the unconditional power can be developed in the following way: We begin with the approximation of the preceding paragraph and then note that for large N , λ^* is well approximated by $\lambda^{**} = [(N - p - 1)/(N - 1)]\lambda^*$, and hence the Poisson random variable K^* can be approximated by a Poisson random variable K^{**} with parameter λ^{**} . Also, for large N , the denominator χ^2 in Equation 7 can be approximated by its degrees of freedom and the $(1 - \alpha)$ th percentile of the $F_{p, N-1}$ distribution can be approximated by the corresponding critical value of χ_p^2 distribution. It follows now that conditional power and, hence, also the unconditional power can be approximated by Cohen's approximation, namely,

$$\text{Prob}(p^{-1}\chi_{p+2K^{**}}^2 \geq \chi_p^{2, 1-\alpha}), \tag{9}$$

where K^{**} has a Poisson distribution with parameter λ^{**} .

One advantage of this type of Approximation 9 is that for fixed α , the approximate power function depends only on p and λ^{**} (u and L , respectively, in Cohen's notation). Thus, Cohen can present power calculations in just a few tables. On the other hand, the exact power must be calculated in terms of α , p , N , and ρ_{YX}^2 (or equivalently, $\rho_{YX}^2/(1 - \rho_{YX}^2)$), and it necessarily requires either more extensive tabulation or computer programs.

We computed a table of sample sizes, similar to Table 1, using Cohen's approximation and tables. The comparison of our two tables showed that Cohen's approximation works quite well in many situations. In general, the results of the approximation are better for $\alpha = .05$ and .1 than for $\alpha = .01$. Specifically, for $\alpha = .05$ the following observations can be made:

Small power and small ρ . The approximate sample sizes are slightly smaller (by at most 5) than the exact ones, for values of u up to 10. Beyond $u = 10$ the trend is reversed, and the approximate sample sizes are actually larger than the exact ones. The discrepancy increases as u becomes large; for example, for $\rho = .15$, Power = .25 and $u = 20, 40, 60$, the exact and approximate sample sizes are (296,301), (418,431), (515,535), respectively.

Large power and large ρ . The approximate sample sizes are again smaller than the exact ones, and this pattern holds for most values of u . It is only for u above 50 that the approximate sample sizes exceed the exact ones.

Power and ρ in the middle of their respective ranges. The approximation works very well for most values of u . Relatively large discrepancies begin to occur when u exceeds 25.

Other Approximations and Tables

A number of other approximations to the distribution of $R_{YX}^2/(1 - R_{YX}^2)$ have been proposed in the literature. Anderson (1984, p. 148) noted several authors who suggested various technical approximations. A more recent approach is described by Kraemer and Thiemann (1987), who used a "master table" for determining sample sizes in various design settings.

(text continues on page 523)

Table 1
 Sample Sizes for .05-Level Tests

ρ	Power										
	0.25	0.50	0.60	0.67	0.70	0.75	0.80	0.85	0.90	0.95	0.99
$u = 1$											
0.10	165	382	487	569	614	690	780	892	1,044	1,291	1,826
0.15	73	169	215	251	271	304	344	394	460	569	805
0.20	41	94	120	140	151	169	191	219	256	317	448
0.25	26	60	76	88	95	107	121	138	162	200	282
0.30	18	41	52	60	65	73	82	94	110	136	193
0.35	13	29	37	43	47	52	59	68	79	98	138
0.40	10	22	28	32	35	39	44	51	59	73	103
0.45	8	17	21	25	27	30	34	39	45	56	79
0.50	7	13	17	20	21	24	27	30	35	44	62
0.55	5	11	13	16	17	19	21	24	28	35	49
0.60	5	9	11	13	13	15	17	19	22	28	39
0.65	4	7	9	10	11	12	14	16	18	22	31
0.70	3	6	7	8	9	10	11	13	15	18	25
0.75	3	5	6	7	7	8	9	10	12	14	20
0.80	2	4	5	5	6	6	7	8	9	11	16
0.85	2	3	4	4	5	5	6	6	7	9	12
0.90	2	3	3	3	4	4	4	5	6	7	9
$u = 2$											
0.10	225	493	617	713	765	853	957	1,085	1,257	1,534	2,125
0.15	99	217	272	314	337	376	422	478	554	676	937
0.20	56	121	151	175	188	209	234	266	308	376	521
0.25	35	76	95	110	118	132	148	167	194	237	328
0.30	24	52	65	75	81	90	101	114	132	161	223
0.35	18	37	47	54	58	64	72	82	95	116	160
0.40	13	28	35	40	43	48	54	61	71	86	119
0.45	10	22	27	31	33	37	41	47	54	66	91
0.50	8	17	21	24	26	29	32	36	42	51	71
0.55	7	13	17	19	20	23	25	29	33	40	56
0.60	6	11	13	15	16	18	20	23	26	32	45
0.65	5	9	11	12	13	15	16	18	21	26	36
0.70	4	7	9	10	11	12	13	15	17	21	29
0.75	3	6	7	8	9	9	10	12	14	16	23
0.80	3	5	6	6	7	7	8	9	11	13	18
0.85	2	4	5	5	5	6	6	7	8	10	14
0.90	2	3	4	4	4	4	5	5	6	7	10
$u = 3$											
0.10	269	572	710	816	873	970	1,083	1,221	1,407	1,705	2,336
0.15	119	252	313	359	385	427	477	538	620	751	1,029
0.20	66	140	174	200	214	237	265	299	344	417	572
0.25	42	88	109	126	134	149	167	188	217	263	360
0.30	29	60	75	85	91	102	113	128	147	179	245
0.35	21	43	53	61	66	73	81	92	106	128	176
0.40	16	32	40	46	49	54	60	68	78	95	131
0.45	12	25	30	35	37	41	46	52	60	73	100
0.50	10	19	24	27	29	32	36	40	47	56	78
0.55	8	15	19	21	23	25	28	32	37	44	61
0.60	6	12	15	17	18	20	22	25	29	35	49
0.65	5	10	12	14	15	16	18	20	23	28	39
0.70	4	8	10	11	12	13	14	16	19	22	31
0.75	4	6	8	9	9	10	11	13	15	18	24
0.80	3	5	6	7	7	8	9	10	12	14	19
0.85	3	4	5	5	6	6	7	8	9	11	15
0.90	2	3	4	4	4	5	5	6	6	8	11
$u = 4$											
0.10	306	637	786	900	961	1,064	1,185	1,332	1,529	1,844	2,506
0.15	135	281	346	396	423	469	522	587	673	812	1,104
0.20	75	156	192	220	235	260	289	326	374	451	613
0.25	47	98	121	138	148	164	182	205	235	284	386

Table 1 (continued)

ρ	Power										
	0.25	0.50	0.60	0.67	0.70	0.75	0.80	0.85	0.90	0.95	0.99
<i>u = 4 (continued)</i>											
0.30	32	67	82	94	100	111	124	139	160	193	262
0.35	23	48	59	67	72	79	89	100	114	138	188
0.40	17	35	44	50	53	59	66	74	85	102	140
0.45	13	27	33	38	41	45	50	56	65	78	106
0.50	11	21	26	29	31	35	39	44	50	61	83
0.55	9	17	20	23	25	27	30	34	39	48	65
0.60	7	13	16	18	20	22	24	27	31	38	52
0.65	6	11	13	15	16	17	19	22	25	30	41
0.70	5	9	10	12	12	14	15	17	20	24	33
0.75	4	7	8	9	10	11	12	14	16	19	26
0.80	3	6	7	7	8	9	9	11	12	15	20
0.85	3	4	5	6	6	7	7	8	9	11	15
0.90	2	3	4	4	5	5	5	6	7	8	11
<i>u = 5</i>											
0.10	338	694	852	972	1,037	1,146	1,273	1,428	1,635	1,963	2,653
0.15	149	305	375	428	456	504	560	628	719	864	1,168
0.20	83	169	208	237	253	280	311	349	399	480	649
0.25	52	106	131	149	159	176	195	219	251	302	408
0.30	35	72	89	101	108	119	133	149	170	205	277
0.35	25	52	63	72	77	85	95	106	122	146	198
0.40	19	38	47	54	57	63	70	79	90	109	147
0.45	15	29	36	41	43	48	53	60	69	83	112
0.50	11	23	28	32	34	37	41	46	53	64	87
0.55	9	18	22	25	26	29	32	36	42	50	68
0.60	7	14	17	20	21	23	26	29	33	40	54
0.65	6	11	14	16	17	18	20	23	26	31	43
0.70	5	9	11	12	13	14	16	18	21	25	34
0.75	4	7	9	10	10	11	13	14	16	20	27
0.80	3	6	7	8	8	9	10	11	13	15	21
0.85	3	5	5	6	6	7	8	8	10	11	16
0.90	2	4	4	4	5	5	6	6	7	8	11
<i>u = 6</i>											
0.10	367	744	911	1,037	1,105	1,219	1,351	1,513	1,728	2,070	2,784
0.15	161	327	401	456	486	536	595	666	760	911	1,226
0.20	89	181	222	253	269	297	330	369	422	505	680
0.25	56	114	139	159	169	187	207	232	265	318	428
0.30	38	77	94	108	115	127	140	157	180	216	290
0.35	27	55	67	77	82	90	100	112	128	154	208
0.40	20	41	50	57	61	67	74	83	95	114	154
0.45	16	31	38	43	46	51	56	63	72	87	117
0.50	12	24	29	33	35	39	43	49	56	67	91
0.55	10	19	23	26	28	31	34	38	44	53	71
0.60	8	15	18	21	22	24	27	30	34	41	56
0.65	6	12	14	16	17	19	21	24	27	33	45
0.70	5	9	11	13	14	15	17	19	21	26	35
0.75	4	8	9	10	11	12	13	15	17	20	28
0.80	4	6	7	8	8	9	10	11	13	16	21
0.85	3	5	5	6	6	7	8	9	10	12	16
0.90	2	4	4	5	5	5	6	6	7	8	11
<i>u = 7</i>											
0.10	393	790	965	1,096	1,167	1,286	1,423	1,591	1,813	2,166	2,903
0.15	173	347	424	482	513	565	626	700	798	953	1,277
0.20	96	192	235	267	284	313	347	388	442	529	709
0.25	60	121	147	167	178	196	218	243	278	332	445
0.30	41	82	100	113	121	133	147	165	188	225	302

(table continues)

Table 1 (continued)

ρ	Power										
	0.25	0.50	0.60	0.67	0.70	0.75	0.80	0.85	0.90	0.95	0.99
$u = 7$ (continued)											
0.35	29	58	71	81	86	95	105	118	134	161	216
0.40	21	43	52	60	64	70	78	87	99	119	160
0.45	16	33	40	45	48	53	59	66	75	90	122
0.50	13	25	31	35	37	41	45	51	58	70	94
0.55	10	20	24	27	29	32	35	40	45	55	74
0.60	8	15	19	21	23	25	28	31	36	43	58
0.65	7	12	15	17	18	20	22	25	28	34	46
0.70	5	10	12	13	14	16	17	19	22	27	36
0.75	4	8	9	10	11	12	14	15	17	21	28
0.80	4	6	7	8	9	9	10	12	13	16	22
0.85	3	5	6	6	7	7	8	9	10	12	16
0.90	2	4	4	5	5	5	6	6	7	8	11
$u = 8$											
0.10	417	833	1,014	1,151	1,224	1,347	1,489	1,663	1,893	2,256	3,012
0.15	183	366	446	506	538	592	655	731	832	992	1,325
0.20	101	202	247	280	298	328	363	405	461	550	735
0.25	63	127	155	175	187	206	227	254	289	345	462
0.30	43	86	105	119	126	139	154	172	196	234	313
0.35	31	61	74	85	90	99	110	123	140	167	224
0.40	23	45	55	62	66	73	81	91	103	124	166
0.45	17	34	41	47	50	55	61	69	78	94	126
0.50	13	26	32	36	39	43	47	53	60	72	97
0.55	10	20	25	28	30	33	37	41	47	56	76
0.60	8	16	20	22	24	26	29	32	37	44	60
0.65	7	13	15	17	19	20	23	25	29	35	47
0.70	5	10	12	14	15	16	18	20	23	27	37
0.75	4	8	10	11	11	13	14	16	18	21	29
0.80	4	6	7	8	9	10	11	12	14	16	22
0.85	3	5	6	6	7	7	8	9	10	12	16
0.90	2	4	4	5	5	5	6	6	7	8	11
$u = 9$											
0.10	440	873	1,061	1,202	1,278	1,405	1,551	1,730	1,966	2,339	3,115
0.15	193	383	466	528	561	617	682	760	864	1,029	1,370
0.20	107	212	258	292	311	342	377	421	479	570	760
0.25	67	133	161	183	195	214	237	264	300	358	477
0.30	45	90	109	124	132	145	160	179	203	242	324
0.35	32	64	78	88	94	103	114	127	145	173	231
0.40	24	47	57	65	69	76	84	94	107	128	171
0.45	18	35	43	49	52	57	64	71	81	97	130
0.50	14	27	33	38	40	44	49	55	62	75	100
0.55	11	21	26	29	31	34	38	43	49	58	78
0.60	9	17	20	23	24	27	30	33	38	46	62
0.65	7	13	16	18	19	21	23	26	30	36	49
0.70	6	10	12	14	15	17	18	20	23	28	38
0.75	5	8	10	11	12	13	14	16	18	22	30
0.80	4	6	8	8	9	10	11	12	14	17	23
0.85	3	5	6	6	7	7	8	9	10	12	17
0.90	2	4	4	5	5	5	6	6	7	9	12
$u = 10$											
0.10	462	910	1,105	1,250	1,328	1,458	1,609	1,793	2,036	2,418	3,210
0.15	202	399	485	549	583	641	707	788	895	1,063	1,412
0.20	112	221	268	304	323	354	391	436	495	589	783
0.25	70	138	168	190	202	222	245	273	311	369	492
0.30	47	93	113	128	136	150	166	185	210	250	333
0.35	33	66	80	91	97	107	118	132	150	178	238
0.40	24	49	59	67	71	79	87	97	110	132	176
0.45	18	37	45	51	54	59	66	73	83	100	133
0.50	14	28	34	39	41	45	50	56	64	77	103
0.55	11	22	27	30	32	35	39	44	50	60	80

Table 1 (continued)

ρ	Power										
	0.25	0.50	0.60	0.67	0.70	0.75	0.80	0.85	0.90	0.95	0.99
<i>u = 10 (continued)</i>											
0.60	9	17	21	24	25	28	31	34	39	47	63
0.65	7	13	16	18	20	22	24	27	31	37	50
0.70	6	11	13	14	15	17	19	21	24	29	39
0.75	5	8	10	11	12	13	14	16	18	22	30
0.80	4	6	8	9	9	10	11	12	14	17	23
0.85	3	5	6	7	7	7	8	9	10	12	17
0.90	2	4	4	5	5	5	6	6	7	9	12
<i>u = 15</i>											
0.10	556	1,074	1,295	1,459	1,547	1,693	1,862	2,067	2,336	2,758	3,627
0.15	243	470	567	640	678	743	817	907	1,026	1,212	1,594
0.20	133	259	313	353	374	410	451	501	567	670	883
0.25	83	161	195	220	234	256	282	314	355	420	553
0.30	55	108	131	148	157	173	190	212	240	284	374
0.35	39	77	93	105	111	122	135	150	170	202	267
0.40	28	56	68	77	82	90	99	110	125	148	197
0.45	21	42	51	58	61	67	74	83	94	112	149
0.50	16	32	39	44	47	51	57	63	72	86	114
0.55	12	24	30	34	36	40	44	49	56	66	89
0.60	10	19	23	26	28	31	34	38	43	52	69
0.65	8	15	18	20	22	24	26	29	34	40	54
0.70	6	11	14	16	17	18	20	23	26	31	42
0.75	5	9	11	12	13	14	16	17	20	24	32
0.80	4	7	8	9	10	11	12	13	15	18	24
0.85	3	5	6	7	7	8	9	10	11	13	18
0.90	2	4	4	5	5	6	6	7	7	9	12
<i>u = 20</i>											
0.10	635	1,211	1,454	1,634	1,730	1,890	2,073	2,296	2,586	3,042	3,973
0.15	276	529	636	715	758	828	909	1,006	1,135	1,335	1,745
0.20	151	291	350	394	417	456	501	555	627	738	965
0.25	93	181	218	245	260	285	313	347	391	461	604
0.30	62	121	146	165	175	191	210	233	264	311	408
0.35	43	85	103	116	123	135	149	165	187	221	290
0.40	31	62	75	85	90	99	109	121	137	162	213
0.45	23	46	56	63	67	74	81	90	103	122	161
0.50	17	35	42	48	51	56	62	69	78	93	123
0.55	13	26	32	37	39	43	47	53	60	71	95
0.60	10	20	25	28	30	33	36	41	46	55	74
0.65	8	16	19	22	23	25	28	31	36	43	57
0.70	6	12	15	17	18	19	21	24	27	33	44
0.75	5	9	11	13	13	15	16	18	21	25	34
0.80	4	7	8	9	10	11	12	14	15	18	25
0.85	3	5	6	7	7	8	9	10	11	13	18
0.90	2	4	5	5	5	6	6	7	8	9	12

Note. Required sample size = tabled value + $u + 1$.

A number of related tables concerning the multiple correlation coefficient have also been presented. Kramer (1963) gave for certain cases the exact upper 5 percentage points for the distribution of the multiple correlation coefficient. Lee (1972) gave a much more extensive range of upper percentage points than did Kraemer. Hager and Möller (1986) provided approximate sample size calculations using Cohen's (1977) approach for a broad range of values of α , particularly small ones. Hager and Möller's tables are intended for use in multiple testing set-

tings. Related work has been reported by Jannerone, Tarnowski, and Macera (1987).

Discussion

In practice, most studies are observational or quasi-experimental, and for such studies the values at the independent variables cannot be set by design. They have to be treated as outcomes, and their variability should be taken into account in the

analysis. Thus, power and sample size calculations need to be done using the unconditional model in these types of studies.

Our treatment of the unconditional model assumed a multivariate normal distribution for (Y, \mathbf{X}) . Transformations may be used on the original X 's to ensure normality, if these are nearly continuous random variables. When underlying normality is not present, theoretical calculations show that robustness of the distribution of $R^2_{Y\mathbf{X}}$ depends on multivariate notions of skewness and kurtosis. In particular, Muirhead (1982) showed that when the distribution of (Y, \mathbf{X}) is elliptically symmetric (e.g., Kariya & Sinha, 1989), the asymptotic distribution of $R^2_{Y\mathbf{X}}$ depends on what he termed the *kurtosis parameter* (Muirhead, 1982, pp. 41, 184). In this setting, when the kurtosis parameter is larger than what occurs under normality, correspondingly larger sample sizes are required to achieve the same power. In practical situations, assessing whether a nonnormal underlying distribution of (Y, \mathbf{X}) allows for the robustness of $R^2_{Y\mathbf{X}}$ is a difficult task. There are experimental studies, for example, in which several of the independent variables represent dichotomous outcomes, whereas others are continuous but not normally distributed. The power in such situations is invariably not a function of the underlying multiple correlation coefficient. Thus, specification of meaningful alternative hypotheses is also difficult. On the other hand, even if such alternative hypotheses could be specified, it is doubtful that the power calculations we present (or the approximation of Cohen, 1977) give results that are accurate and reasonable.

The comparison of the exact sample size values to the approximate sample sizes obtained using Cohen's (1977) tables shows that this approximation is quite good in many situations of practical interest. For the user of this approximation who does not want to use our exact tables, a conservative course of action would be to augment the approximate sample size by 5, especially if the number of independent variables is small (below 10).

The computation of the unconditional sample size and power requires only a modest amount of programming. Alternatively, one can either use our tables or obtain a copy of our interactive

program and use it on any computer that supports IMSL sub-routines. The program takes very little time to run and is more versatile than the use of the tables, because it can accommodate any combination of parameters. For this reason, the publication of more extensive tables—for example, for values of α other than .01, .05, and .1—becomes unnecessary.

References

- Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York: Wiley.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.
- Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Hager, W., & Möller, H. (1986). Tables for the determination of power and sample size in univariate and multivariate analyses of variance and regression. *Biometrical Journal*, 28, 647–663.
- Jannerone, R. S., Tarnowski, P. A., & Macera, C. A. (1987). *Establishing optimal treatment of nuisance parameters*. Unpublished manuscript.
- Johnson, N. L., & Kotz, S. (1969). *Discrete distributions*. New York: Wiley.
- Kariya, T., & Sinha, B. K. (1989). *Robustness of statistical tests*. Orlando, FL: Academic Press.
- Kraemer, H. C., & Thiemann, S. (1987). *How many subjects*. Beverly Hills, CA: Sage.
- Kramer, K. H. (1963). Tables for constructing confidence limits on the multiple correlation coefficient. *Journal of the American Statistical Association*, 58, 1082–1085.
- Lee, Y. S. (1972). Tables of the upper percentage points of the multiple correlation. *Biometrika*, 59, 175–189.
- McKinlay, S. M. (1975). The observational study: A review. *Journal of the American Statistical Association*, 70, 503–520.
- Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. New York: Wiley.
- Sampson, A. R. (1974). A tale of two regressions. *Journal of the American Statistical Association*, 69, 682–689.

Received July 28, 1988

Revision received October 12, 1988

Accepted February 16, 1989 ■