Behavior Research Methods
2007, 39 (4), 731-734

Methodological and computational considerations for multiple correlation analysis

Gwowen Shieh and Chien-Feng Kung
National Chiao Tung University, Hsinchu, Taiwan

The squared multiple correlation coefficient has been widely employed to assess the goodness-of-fit of linear regression models in many applications. Although there are numerous published sources that present inferential issues and computing algorithms for multinormal correlation models, the statistical procedure for testing substantive significance by specifying the nonzero-effect null hypothesis has received little attention. This article emphasizes the importance of determining whether the squared multiple correlation coefficient is small or large in comparison with some prescribed standard and develops corresponding Excel worksheets that facilitate the implementation of various aspects of the suggested significance tests. In view of the extensive accessibility of Microsoft Excel software and the ultimate convenience of general-purpose statistical packages, the associated computer routines for interval estimation, power calculation, and sample size determination are also provided for completeness. The statistical methods and available programs of multiple correlation analysis described in this article purport to enhance pedagogical presentation in academic curricula and practical application in psychological research.

The study of correlation coefficients among variables is one of the most fundamental issues across a variety of disciplines including psychological research. In particular, the majority of the literature has been focused on the multiple correlation coefficient between a criterion variable and one set of predictor variables in the context of linear regression models. As the squared multiple correlation coefficient or the strength of association ρ^2 represents the fraction of reduction in the variance of criterion variable accounted for by the predictor variables and the overall usefulness of the regression model, extensive results have been derived that give various expressions, approximations, and computing algorithms for the theoretical properties of the sample squared multiple correlation coefficient or the coefficient of determination, R^2, when the criterion and predictor variables have a joint multivariate normal distribution. See Johnson, Kotz, and Balakrishnan (1995, chap. 32) and Stuart and Ord (1994, chap. 16) for comprehensive discussions and further details. A primary concern regarding regression analysis is the conception of the two distinct scenarios of fixed (conditional) and random (unconditional) modeling formulations that ultimately lead to different inferential procedures. One must have a clear understanding of the respective setups and how they can be utilized before the issues involved in the construction of an appropriate regression model can be fully explained. Notably, Sampson (1974) gave an excellent and thorough description of the two modeling formulations in which the random setting adopts the convenient assumption that all variables have a joint multivariate normal distribution. The procedures for power calculation, interval estimation, and sample size determination under the fixed regression models are well known (see Murphy & Myors, 2004; Smithson, 2003, and the references therein for further details). However, the corresponding methods are more complex under the random model. Specifically, we focus our attention here on the situation that the criterion and predictor variables have a joint multinormal distribution.

METHODOLOGICAL CONSIDERATION

Although the sample squared multiple correlation coefficient is routinely computed with all major commercial software packages, these packages do not offer a full range of inferential procedures for multiple correlation analysis. For the purpose of interval estimation, power calculation, and sample size determination for the squared multiple correlation coefficient, various methodological developments and computer programs are presented in Algina and Olejnik (2003), Dunlap, Xin, and Myers (2004), Mendoza and Stafford (2001), Shieh (2006), and Steiger and Fouladi (1992). The currently available computing algorithms presented in these articles are useful in important and distinctive ways, in that they implement different statistical methods under diverse user interfaces. In considering inference on multiple correlation coefficients, the R2 package developed by Steiger and Fouladi (1992) appears to be the most versatile numerical routine because it includes not only the percentile and cumulative distribution

G. Shieh, gwshieh@mail.nctu.edu.tw

function of R^2, exact confidence interval estimation for ρ^2, power calculation, and sample size determination for standard significance test H_0: $\rho^2 = 0$, but also the hypothesis testing procedure H_0: $\rho^2 = \rho_0^2$ ($\rho_0^2 > 0$) not generally available in the other aforementioned computer programs. However, there is room for some further extensions in two aspects. First, the calculation of minimum sample size required for the sample squared multiple correlation coefficient R^2 to fall into a prescribed interval with adequate accuracy was illustrated in Algina and Olejnik (2003) and Shieh (2006). See Kelley and Maxwell (2003) for a recent treatment of sample size planning for accuracy in parameter estimation within the multiple regression framework. Second, it is a natural generalization to incorporate both the power calculation and sample size determination for testing hypotheses involving nonzero target values for ρ^2. Note that Wilcox (1980) presented the exact sample sizes using the indifference zone approach to the problem of determining whether ρ^2 is above or below a known constant. Moreover, Fowler (1985), Murphy and Myors (1999), and Steiger (2004) repeatedly stressed the notion of testing substantive significance in the context of general linear models.

COMPUTATIONAL CONSIDERATION

From a purely computational perspective, the R2 package of Steiger and Fouladi (1992) for multiple correlation analysis is a stand-alone MSDOS program, whereas the existing algorithms of Algina and Olejnik (2003), Dunlap, Xin, and Myers (2004), Mendoza and Stafford (2001), and Shieh (2006) are associated with more advanced software systems, such as FORTRAN, Mathematica, SAS, and SPSS. Since these large commercial packages may not be generally available and have their own unique user interfaces, extra effort is required for performing the necessary analyses, thus incurring greater complexity. Conceivably, it should be of practical interest for researchers or students to conduct multiple correlation analysis on a readily accessible computing platform. In view of the widespread availability of personal computers operated with Microsoft Windows systems, this article aims to present a Microsoft Excel program that contains a full range of inference procedures for multiple correlation analysis. The program is referred to as RHO-SQUARE, and detailed information is described in the following section. The proposed package has the distinct features of being accessible, accurate, flexible, and free.

It should be emphasized that there has been some concern regarding the accuracy of statistical procedures in Microsoft Excel (see, for example, Knusel, 2005; McCullough & Wilson, 2005). In the RHO-SQUARE program, only limited statistical functions of Excel are employed in our developed routines, along with other standard mathematical functions. Specifically, the central F cumulative distribution function and its inverse are utilized for computing the cumulative probability and quantile of regular F distribution. In the vital part of numerical computation, the formulation and expansion of Lee (1972) are incorporated to evaluate the notoriously sophisticated cumulative distribution function of R^2. The computation is theoretically exact provided that the auxiliary functions can be evaluated exactly. In conjunction with basic computation techniques that require the standard numerical methods of one-dimensional integration and interval-halving algorithms, the suggested Excel program provides alternative routines for performing multiple correlation analysis. In order to verify the accuracy of the RHO-SQUARE program, four sets of comparisons were conducted. First, the lower bounds of the upper 95% confidence intervals presented in Table 5 of Mendoza and Stafford (2001) with $N = 50$ for given values of number of predictor variables P and observed value of R^2 were recalculated. It appears that the results are almost identical, and the discrepancy is obviously due to a different roundoff scheme. Second, the present program yielded exactly the same minimum sample sizes required for the prescribed interval $(0, \rho^2 + b)$ of squared multiple correlation coefficient with coverage probability at least 0.95 and $P = 5$ given in Table 4 of Shieh (2006). Third, the computed powers and those presented in Table 1 of Dunlap, Xin, and Myers (2004) are practically equal for the selected sample size N, number of predictor variables P, target value $\rho_0^2 = 0$ under null hypothesis, true value $\rho_1^2 = \rho^2$ under alternative hypothesis, and significance level $\alpha = 0.05$. The last comparison is performed for the minimum sample size N needed to test hypothesis in order to attain the specified power for the chosen number of predictor variables $P = 5$, target value $\rho_0^2 = 0$ under null hypothesis, true value $\rho_1^2 = \rho^2$ under alternative hypothesis, and significance level $\alpha = 0.05$. All the results generated by the RHO-SQUARE program coincided with those presented in Gatsonis and Sampson (1989) except for three cases that differed only by one unit. The differences seem to be negligible with respect to the magnitude of the resulting sample sizes. According to these comparisons, we conclude that there is an excellent agreement between the RHO-SQUARE program and the existing algorithms in all cases. Other Excel programs that attempt to address related but different aspects of normal correlation analysis can be found in Alf and Graf (2002) and Barnette (2005). Finally, it is noteworthy that the inferential procedures considered in the RHO-SQUARE program depend on the specific multinormal assumption of the criterion and predictor variables as in the other abovementioned software packages for multiple correlation analysis. When the underlying normality assumption is not present, it is questionable that the procedures will give accurate and reasonable results. In view of this significant limitation, therefore, it seems prudent to ensure that the properties of the associated variables are well understood before the standard inferential methods are adopted by researchers as general analytic procedures.

THE RHO-SQUARE PROGRAM

For the ultimate goal of presenting a full account of exact procedures for the analysis of squared multiple correlation coefficient, the RHO-SQUARE program includes three sets of algorithms for performing the calculations related to the basic distributional properties of R^2 and the

statistical methods of interval estimation and hypothesis testing for ρ^2. The program has four pages of worksheets. The first page contains a brief introduction, followed by three worksheets that are organized to present the following features.

Distributional Properties of R^2

The probability density function and cumulative density function of R^2 are plotted for given values of population squared multiple correlation coefficient ρ^2, sample size N, and number of predictor variables P. Moreover, the percentile and cumulative probability for the prescribed model configuration can be computed by specifying the selected cumulative probability and percentage points of R^2.

Interval Estimation for ρ^2

All the lower one-sided, upper one-sided, and two-sided confidence interval estimation problems are considered. For each formulation, the exact 100 $(1 - \alpha)$ percent confidence interval of ρ^2 can be calculated for given values of sample size N, number of predictor variables P, confidence level $1 - \alpha$, and observed value of R^2. Conversely, the exact coverage probability of a specified interval (R^2_L, R^2_U) is computed with respect to the underlying population squared multiple correlation coefficient ρ^2, sample size N, number of predictor variables P, and interval limits R^2_L and R^2_U. Additionally, in order to ensure adequate estimation accuracy with appropriate sample size, the smallest sample size N required for the sample squared multiple correlation coefficient to fall into the interval (R^2_L, R^2_U) with probability $1 - \alpha$ can be determined for the chosen values of population squared multiple correlation coefficient ρ^2, number of predictor variables P, interval limits R^2_L and R^2_U, and desired confidence coefficient $1 - \alpha$.

Hypothesis Testing for ρ^2

The subsequent one- and two-tail tests of hypotheses can be conducted: H_0: $\rho^2 \leq \rho^2_0$, H_0: $\rho^2 \geq \rho^2_0$, and H_0: $\rho^2 = \rho^2_0$. In each case, the critical values and p value are calculated under the given quantities of sample size N, number of predictor variables P, significance level α, target value ρ^2_0 under null hypothesis, and observed value of R^2. For the purpose of power calculation, the exact power is computed for the input values of sample size N, number of predictor variables P, significance level α, target value ρ^2_0 under null hypothesis, and true value ρ^2_1 under alternative hypothesis. The power approach to sample size determination can be performed as well. The program calculates the minimum sample size N needed to test hypotheses in order to attain the specified power for the chosen number of predictor variables P, significance level α, target value ρ^2_0 under null hypothesis, true value ρ^2_1 under alternative hypothesis, and desired power level.

EXAMPLES

In order to facilitate the application and illustrate the features of the RHO-SQUARE program, the following numerical examples are presented.

Example 1

Suppose a linear regression analysis is performed with $N = 50$, $P = 5$, and the sample squared multiple correlation coefficient $R^2 = .3$. Then the computed lower, upper, and two-sided 95% confidence intervals are (0, .4245), (.0589, 1), and (.0337, .4603), respectively. For research planning purposes, RHO-SQUARE can easily determine the precise minimum sample size for R^2 to fall into the interval (R^2_L, R^2_U) with a prescribed probability. Assume that $\rho^2 = .4$ and $P = 5$, the required sample sizes to ensure the desired accuracy with probability .95 for selected intervals $(R^2_L, R^2_U) = (0, .6)$, $(.3, 1)$, and $(.3, .6)$ are 58, 99, and 109, respectively.

Example 2

Consider the hypothesis testing problem for the squared multiple correlation coefficient ρ^2 with $N = 100$ and $P = 5$. The test for confirming a substantial level of strength of association in terms of H_0: $\rho^2 \leq .3$ versus H_1: $\rho^2 > .3$ can be readily conducted with RHO-SQUARE. The respective critical values for Type I error rates .05 and .01 are .4566 and .5068. Suppose the observed value of $R^2 = .5$, then the associated p value can be found as .0128. Hence, the rejection of null hypothesis at the .05 significance level demonstrates that the strength of association exceeds the chosen threshold of .3. On the other hand, one may be interested in determining whether the strength of association is trivial or minimum with H_0: $\rho^2 \geq .2$ versus H_1: $\rho^2 < .2$. In this situation, the critical values for $\alpha = .05$ and .01 are .1242 and .0865, respectively, if one observed $R^2 = .10$, so the corresponding p value is .0192 and the null hypothesis is rejected at the .05 significance level. Accordingly, the result suggests that the level of strength of association is not high enough to make a real difference. Moreover, the power approach to sample size determination can be performed as well. The minimum sample size $N = 153$ is required for testing the hypothesis H_0: $\rho^2 \geq .2$ with specified parameter values of $\rho^2_0 = .2$ and $\rho^2_1 = .05$, significance level .05, and nominal power .90.

CONCLUSION

Given the complex interrelationships that exist among multiple variables in psychology and other social science settings, it is important for researchers to become conversant with various analytic techniques for squared multiple correlation coefficient. Knowledge of corresponding inferential procedures is often critical for investigators to address scientific hypotheses and confirm credible effects. Furthermore, the a priori determination of a proper sample size necessary to achieve some specified power and accuracy is a salient problem encountered frequently in applied settings. More important, it is more efficient for researchers or students to be able to conduct multiple correlation analysis on a readily accessible computing platform with modern personal computers. The developed Excel program offers a wide range of potentially useful tools for multiple correlation analysis and concurrently accounts for some prominent statistical notions that were not found in the existing routines.

AUTHOR NOTE

This research was partially supported by National Science Council Grant NSC-95-2416-H-009-031-MY2. The authors thank the referees for several valuable comments. The complete set of numerical verifications for the accuracy of the RHO-SQUARE program with other published results is available from the first author at gwshieh@mail.nctu.edu.tw. The program is also available at no cost to interested researchers upon request. It is hoped that the proposed multiple correlation analysis software will facilitate pedagogical presentation in academic curriculum and practical application in psychological research. Correspondence concerning this article should be addressed to G. Shieh, Department of Management Science, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan 30050, Taiwan (e-mail: gwshieh@mail.nctu.edu.tw).

REFERENCES

ALF, E. F., JR., & GRAF, R. G. (2002). A new maximum likelihood estimator for the population squared multiple correlation. *Journal of Educational & Behavioral Statistics*, **27**, 223-235.

ALGINA, J., & OLEJNIK, S. (2003). Sample size tables for correlation analysis with applications in partial correlation and multiple regression analysis. *Multivariate Behavioral Research*, **38**, 309-323.

BARNETTE, J. J. (2005). ScoreRel CI: An Excel program for computing confidence intervals for commonly used score reliability coefficients. *Educational & Psychological Measurement*, **65**, 980-983.

DUNLAP, W. P., XIN, X., & MYERS, L. (2004). Computing aspects of power for multiple regression. *Behavior Research Methods, Instruments, & Computers*, **36**, 695-701.

FOWLER, R. L. (1985). Testing for substantive significance in applied research by specifying nonzero effect null hypotheses. *Journal of Applied Psychology*, **70**, 215-218.

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

JOHNSON, N. L., KOTZ, S., & BALAKRISHNAN, N. (1995). *Continuous univariate distributions* (2nd ed., Vol. 2). New York: Wiley.

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.

KNUSEL, L. (2005). On the accuracy of statistical distributions in Microsoft Excel 2003. *Computational Statistics & Data Analysis*, **48**, 445-449.

LEE, Y. S. (1972). Tables of upper percentage points of the multiple correlation coefficient. *Biometrika*, **59**, 175-189.

MCCULLOUGH, B. D., & WILSON, B. (2005). On the accuracy of statistical procedures in Microsoft Excel 2003. *Computational Statistics & Data Analysis*, **49**, 1244-1252.

MENDOZA, J. L., & STAFFORD, K. L. (2001). Confidence interval, power calculation, 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 & Psychological Measurement*, **61**, 650-667.

MURPHY, K. R., & MYORS, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology*, **84**, 234-248.

MURPHY, K. R., & MYORS, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

SHIEH, G. (2006). Exact interval estimation, power calculation and sample size determination in normal correlation analysis. *Psychometrika*, **71**, 529-540.

SMITHSON, M. (2003). Confidence intervals. *Quantitative Applications in the Social Sciences Series* (No. 140). Thousand Oaks, CA: Sage.

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

STEIGER, J. H., & FOULADI, R. T. (1992). R2: A computer program for interval estimation, power calculations, sample size estimation, and hypothesis testing in multiple regression. *Behavior Research Methods, Instruments, & Computers*, **24**, 581-582.

STUART, A., & ORD, J. K. (1994). *Kendall's advanced theory of statistics* (6th ed., Vol. 1). New York: Halsted.

WILCOX, R. R. (1980). Some exact sample sizes for comparing the squared multiple correlation coefficient to a standard. *Educational & Psychological Measurement*, **40**, 119-124.

(Manuscript received March 24, 2006;
revision accepted for publication September 21, 2006.)