Chilean Journal of Statistics
Volume 13, Number 2 (December 2022), Pages 187–199
DOI: 10.32372/chjs.13-02-04

Statistical inference
Research Paper

Computing the noncentrality parameter to the distribution of the square of the sample multiple correlation coefficient

Izabela R. Cardoso de Oliveira[1,*] and Daniel Furtado Ferreira[1]

[1]Department of Statistics, Universidade Federal de Lavras, Lavras, Brazil

(Received: 11 August 2021 · Accepted in final form: 17 May 2022)

Abstract

The distribution of the square of the sample multiple correlation coefficient, R^2, can be expressed as a negative binomial mixture of the central incomplete beta function and used to test hypotheses about the population correlation coefficient. Efficient algorithms for obtaining the distribution function were proposed, but no report was found in literature on algorithms for obtaining the inverse of the distribution and for calculating the noncentrality parameter. In this study we propose an algorithm that combines the method proposed by Benton and Krishnamoorthy (2003) with the inversion of the distribution function with respect to the noncentrality parameter, using the Newton-Raphson method. Such method provides mechanisms for obtaining confidence intervals for the population multiple correlation coefficient. Furthermore, this algorithm can be used to calculate minimal detectable differences in tests of hypotheses with a given pre-specified power. The algorithm is proposed and successfully implemented in `R`. Applications to soil data collected through BiosBrasil project and to state.x77 `R` dataset are used to illustrate its use while obtaining confidence interval for the coefficient of determination in multiple regression models.

Keywords: Algorithm · Coefficient of determination · Incomplete beta function · Negative binomial.

Mathematics Subject Classification: Primary 62E05 · Secondary 62J05.

1. Introduction

In applied research regression models are widely used to describe the relationship between one or more predictor variables and the response variable. The multiple correlation coefficient generalizes the correlation coefficient and is used in multiple regression analysis to assess the goodness of fit of the model. For a set of variables following a multivariate normal distribution, this coefficient represents the maximum correlation between a response variable, Y, and a linear combination of predictor variables. The square of the sample multiple correlation coefficient, that is, the coefficient of determination, denoted R^2, measures the proportion of total variation explained by the regression. Its distribution is useful, for example, in hypothesis testing on the multiple correlation coefficient of the population, ρ.

[*] Corresponding author. Email: izabela.oliveira@ufla.br

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

The distribution of the square of the sample multiple correlation coefficient (Muirhead, 1982; Benton and Krishnamoorthy, 2003) is placed within the framework of noncentral distributions. These distributions are generalizations of the corresponding central distributions, through the addition of the noncentrality parameter.

The gamma and beta noncentral distributions as well as those derived from them (noncentral t, F, chisquare and the distribution of the square of the sample multiple correlation coefficient) play an important role in statistics. All of them can be expressed as discrete mixtures of continuous distributions given by the general formula

$$P(X \leq x) = \sum_{i=0}^{\infty} P(Y = i|\delta) F_Z(x;\ \boldsymbol{\theta}_i), \tag{1.1}$$

where X is a continuous random variable, Y is a discrete random variable, δ is the noncentrality parameter related to the random variable Y, F_Z is the cumulative distribution function of the continuous variable Z, which dependents on the parameter vector $\boldsymbol{\theta}_i$ and $P(Y = i|\delta)$ is the probability mass function of Y, given the ith value and the noncentrality parameter δ.

As pointed out by Baharev and Kemény (2008) and Benton and Krishnamoorthy (2003), when the noncentrality parameter is large many of the existing algorithms for noncentral distributions produce incorrect results. In the calculation of the distribution function they are based on computing recursively the probabilities of the mixture from $i = 0$, which may lead to several problems: i) if the mean of the discrete random variable Y is very large, $P(Y = 0|\delta)$ will be very small and underflow errors can occur; ii) the processing time of this algorithm increases excessively as the mean of Y increases; iii) the algorithm can be inefficient to be used as auxiliary algorithm while calculating percentiles, confidence intervals and noncentrality parameters.

A way to overcome the aforementioned problems is to start the calculation at the point $k = \lceil E(Y) \rceil$, where the operator $\lceil x \rceil$ is the ceiling function that returns the smallest integer not less than x Benton and Krishnamoorthy (2003). In general, the dominant series in Equation (1.1) is the probability $P(Y = i|\delta)$ of the discrete random variable Y, which has its maximum close to $E(Y)$. Then the probabilities for the other terms of the sum are computed recursively from that point. A second alternative, proposed by Baharev and Kemény (2008), computes the distribution function recursively from the interval limited by two integers k_1 and k_2, obtained from the discrete part of the mixture. These authors used the Newton-Raphson method for obtaining the noncentrality parameter δ of the noncentral beta distribution. Combining these ideas, Oliveira and Ferreira (2012) proposed algorithms for the noncentral gamma, the generalization of the noncentral chisquared distribution. Their method was implemented as a `R` package denoted `ncg` (Ferreira et al., 2012).

Computational methods to obtain the distribution of the square of the sample multiple correlation coefficient were presented by Benton and Krishnamoorthy (2003). However, no report was found on methods for calculating the noncentrality parameter of this distribution. Such a method could be used while obtaining confidence intervals for the coefficient of determination, which is widely used in applied research. The problem of constructing confidence interval for the squared multiple correlation coefficient (ρ^2) is known. In 1972, Lee (1972) provided percentage points for the distribution of squared sample multiple correlation coefficient, r^2, which allows exact confidence intervals for ρ^2 using the pivoting the distribution function approach. Confidence intervals can be obtained through the calculator that accompanies the book by Krishnamoorthy (2006). Furthermore, Krishnamoorthy and Xia (2008) have provided a method computing sample size to obtain the exact confidence interval within a given precision.

In this paper, we propose an algorithm that uses both the Benton and Krishnamoorthy

(2003) and Newton-Raphson methods for the inversion of the noncentral cumulative distribution function of the square of the sample multiple correlation coefficient with respect to the noncentrality parameter.

The paper is organized as follows. A review of the distribution of the square of the multiple correlation coefficient is the subject of the following section. Section 3 presents the recursive methods used in the proposed algorithm, which is presented in Appendix A. In Section 4 we use this algorithm to calculate confidence intervals for the coefficient of determination in multiple regression models. . Soil data sampled in Amazonas state through BiosBrasil project and the state.x77 `R` dataset are then used. A Monte Carlo study is performed in Section 5 to evaluate the coverage probability of the confidence intervals. The `R` implementation is given in Appendix B.

2. The distribution of the square of the sample multiple correlation coefficient

Let $\boldsymbol{X} = [X_1, X_2, \cdots, X_p, X_{p+1}]^\top$ be a random vector with variance-covariance matrix $\boldsymbol{\Sigma}$ (positive-definite) and the following partitions

$$\boldsymbol{X} = \left[\begin{array}{c} X_1 \\ \hline \boldsymbol{X}_2 \end{array}\right] \qquad \text{and} \qquad \boldsymbol{\Sigma} = \left[\begin{array}{c|c} \sigma_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right],$$

where $\boldsymbol{X}_2 = [X_2, X_3, \cdots, X_p, X_{p+1}]^\top$ and $\boldsymbol{\Sigma}_{22}$ is $p \times p$, so that $\text{Var}(X_1) = \sigma_{11}$, $\text{Cov}(\boldsymbol{X}_2) = \boldsymbol{\Sigma}_{22}$, $\text{Cov}(X_1, \boldsymbol{X}_2) = \boldsymbol{\Sigma}_{12}$ $(1 \times p)$ and $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^\top$.

The population multiple correlation coefficient between X_1 and $\boldsymbol{X}_2$ $(p \times 1)$, denoted by ρ, is the maximum correlation between X_1 and the linear function $\boldsymbol{a}^\top \boldsymbol{X}_2$ of $\boldsymbol{X}_2$. Therefore

$$\rho = \max_{\boldsymbol{a}} \frac{\text{Cov}(X_1, \boldsymbol{a}^\top \boldsymbol{X}_2)}{\sqrt{\text{Var}(X_1)\text{Var}(\boldsymbol{a}^\top \boldsymbol{X}_2)}} = \max_{\boldsymbol{a}} \frac{\boldsymbol{a}^\top \boldsymbol{\Sigma}_{21}}{\sqrt{\sigma_{11} \boldsymbol{a}^\top \boldsymbol{\Sigma}_{22} \boldsymbol{a}}}.$$

The maximum is achieved when $\boldsymbol{a} = \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}$ and is given by

$$\rho = \left(\frac{\boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}}{\sigma_{11}} \right)^{1/2},$$

where ρ is the absolute value of the ordinary coefficient of correlation (Muirhead, 1982).

Considering $\boldsymbol{X} \sim N_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and that $\boldsymbol{\mu}$ is partitioned as $\boldsymbol{X}$, then the conditional distribution of X_1 given $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is normal with mean and variance given by

$$\begin{aligned} E(X_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) &= \mu_1 + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \\ \text{Var}(X_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) &= \sigma_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}, \end{aligned}$$

respectively.

Since $E(X_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2)$ can be viewed as the regression function of X_1 on $\boldsymbol{x}_2$, the amount of variability of X_1 that can be reduced by conditioning on $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is

$$\sigma^2_{X_1 . \boldsymbol{x}_2} = \sigma_{11} - \text{Var}(X_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) = \sigma_{11} - \left(\sigma_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \right) = \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}.$$

Thus the square of the multiple correlation coefficient is defined by

$$\rho^2 = \frac{\sigma^2_{X_1 \cdot \boldsymbol{x}_2}}{\sigma_{11}} = \frac{\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}}{\sigma_{11}}. \tag{2.2}$$

Now consider a random sample $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, ..., $\boldsymbol{X}_n$ of size n from a $(p+1)$-variate normal distribution and set

$$\boldsymbol{W} = (n-1)\boldsymbol{S} = \sum_{j=1}^{n}\left(\boldsymbol{X}_j - \bar{\boldsymbol{X}}_.\right)\left(\boldsymbol{X}_j - \bar{\boldsymbol{X}}_.\right)^\top,$$

where $\boldsymbol{S}$ is the sample variance-covariance matrix and $\boldsymbol{W}$ is the sum of square and cross products matrix. Consider partitions of $\boldsymbol{W}$ and $\boldsymbol{S}$ similar to those applied to $\boldsymbol{X}$ and $\boldsymbol{\Sigma}$, thus

$$\boldsymbol{W} = \left[\begin{array}{c|c} W_{11} & \boldsymbol{W}_{12} \\ \hline \boldsymbol{W}_{21} & \boldsymbol{W}_{22}\end{array}\right] \qquad \text{and} \qquad \boldsymbol{S} = \left[\begin{array}{c|c} S_{11} & \boldsymbol{S}_{12} \\ \hline \boldsymbol{S}_{21} & \boldsymbol{S}_{22}\end{array}\right].$$

The square of the sample multiple correlation coefficient is defined as

$$R^2 = \frac{\boldsymbol{W}_{12}\boldsymbol{W}_{22}^{-1}\boldsymbol{W}_{21}}{W_{11}} = \frac{\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21}}{S_{11}}.$$

If the multivariate normal is the parental distribution, R^2 will be the maximum likelihood estimator of ρ^2. Here the interest lies on the distribution of R^2 in a sample from the multivariate normal distribution, where the population square of the multiple correlation coefficient, ρ^2, is different from zero. The distribution function of R^2 (Muirhead, 1982) is

$$\begin{aligned} F_{R^2}(x|p,\, n, \rho^2) = P(R^2 \le x|p,\, n, \rho^2) = &\sum_{i=0}^{\infty} \frac{\Gamma[(n-1)/2+i]}{\Gamma(i+1)\Gamma[(n-1)/2]}\left(\rho^2\right)^i\left(1-\rho^2\right)^{(n-1)/2} \\ &\times I_x(p/2+i,\; (n-1-p)/2), \end{aligned} \tag{2.3}$$

where the first factor in Equation (2.3) is the negative binomial probability function with success probability $1-\rho^2$, number of success $(n-1)/2$ and number of fails i, ρ^2 is the noncentrality parameter corresponding to the population square of the multiple correlation coefficient, $0 < x < 1$ is the observed value of R^2, $I_x(p/2+i,\; (n-1-p)/2)$ is the central incomplete beta function with parameters $\alpha = p/2$ and $\beta = (n-1-p)/2$, n is the sample size and p is the number of predictor variables. Note that this distribution is a special case of Equation (1.1).

3. Theoretical aspects of the recursive computation

Since Benton and Krishnamoorthy (2003) have not present a specific algorithm for the inverse of the R^2 distribution with respect to the noncentrality parameter, we will do this as follows. The maximum k of $P(Y = i)$ occurs around the mean of the negative binomial distribution and it is the smallest integer not less than the mean, that is

$$k = \left\lceil \frac{(n-1)\rho^2}{2(1-\rho^2)} \right\rceil = \left\lceil \frac{\nu\rho^2}{2(1-\rho^2)} \right\rceil,$$

where $\lceil x \rceil$ is the smallest integer not less than x and $\nu = n - 1$. The sum in Equation (2.3) is calculated recursively starting from the kth term. The following results can be used while calculating the negative binomial probability function recursively from the ith point:

$$P(Y = i+1) = \frac{\nu/2 + i}{i+1}\rho^2 P(Y = i), \quad i = 0,\, 1,\, 2, \cdots$$
$$P(Y = i-1) = \frac{i}{\nu/2 + i - 1}\rho^{-2} P(Y = i), \quad i = 1,\, 2,\, 3, \cdots$$

At point k of the infinite sum in Equation (2.3), we define the following functions

$$g(x|\alpha + k, \beta) = \frac{\Gamma(\alpha+\beta+k-1)}{\Gamma(\alpha+k)\Gamma(\beta)} x^{\alpha+k-1}(1-x)^{\beta},$$
$$g(x|\alpha + k + 1, \beta) = \frac{\Gamma(\alpha+\beta+k)}{\Gamma(\alpha+k+1)\Gamma(\beta)} x^{\alpha+k}(1-x)^{\beta},$$

and calculate the next term of the central incomplete beta function using forward calculations, that is,

$$I_x(\alpha+k+1,\beta) = I_x(\alpha+k,\beta) - g(x|\alpha+k,\beta) \times \frac{\alpha+\beta+k-1}{\alpha+k} x.$$

In the same way, backward computation of the central incomplete beta function can be done by

$$I_x(\alpha+k-1,\beta) = I_x(\alpha+k,\beta) + g(x|\alpha+k+1,\beta) \times \frac{\alpha+k}{(\alpha+\beta+k-1)x}.$$

For the Newton-Raphson method, we obtain the first derivative of Equation (2.3) with respect to ρ^2, that is

$$\frac{dF_{R^2}(x|p,\, n, \rho^2)}{d\rho^2} = \sum_{i=0}^{\infty} \frac{\Gamma(\nu/2+i)\left(\rho^2\right)^i \left(1-\rho^2\right)^{\nu/2}}{\Gamma(i+1)\Gamma(\nu/2)} I_x\left(\frac{p}{2}+i,\ \frac{\nu-p}{2}\right)\left[\frac{2i(1-\rho^2)-\nu\rho^2}{2\rho^2(1-\rho^2)}\right]. \tag{3.4}$$

Observe that the ith term of the infinite series from Equation (3.4) is the value of the ith term of the distribution function multiplied by t_i, where

$$t_i = \frac{2i(1-\rho^2)-\nu\rho^2}{2\rho^2(1-\rho^2)}.$$

The first derivative can be easily computed using forward and backward calculations, which are, respectively,

$$t_{i+1} = t_i + \frac{1}{\rho^2} \quad \text{and} \quad t_{i-1} = t_i - \frac{1}{\rho^2}.$$

The proposed algorithm to obtain the noncentrality parameter, δ, is presented in Appendix A. In the following section we illustrate its use while obtaining cconfidence intervals for the coefficient of determination in multiple regression models.

4. Application

In this section we consider soil data from BiosBrasil project (http://www.biosbrasil.ufla.br/), which study site is located in Benjamin Constant, Amazonas State, Brazil. Data refer to 30 observations of soil chemical variables from the land use system (LUS) young secondary forest (Table 1).

Table 1. Data on soil chemical variables from the land use system (LUS) young secondary forest. Source: BiosBrasil Project.

obs	pH	Ca	Mg	BS	obs	pH	Ca	Mg	BS
1	4.4	3.1	2.1	5.3	16	5.0	5.1	2.2	7.4
2	4.6	4.1	2.4	6.8	17	5.1	7.3	1.7	9.2
3	5.2	8.3	2.6	11.1	18	5.1	8.5	3.6	12.3
4	4.4	5.2	2.3	7.7	19	4.7	5.2	1.9	7.3
5	5.0	8.5	4.1	12.8	20	4.6	5.0	2.5	7.7
6	4.9	4.2	2.5	6.9	21	4.8	7.8	2.2	10.3
7	4.9	10.5	4.8	15.5	22	4.6	4.4	2.1	6.6
8	4.8	6.9	3.1	10.4	23	5.2	7.9	3.1	11.2
9	4.3	3.4	1.6	5.3	24	5.4	7.0	2.3	9.6
10	4.4	2.6	1.8	4.5	25	4.8	5.0	2.5	7.6
11	5.3	7.6	2.1	10.0	26	4.7	2.9	1.2	4.4
12	4.6	5.4	3.6	9.2	27	5.0	6.4	2.6	9.1
13	4.7	4.9	2.1	7.2	28	4.9	5.7	1.2	7.0
14	5.8	10.4	4.8	15.5	29	5.1	7.3	2.2	9.6
15	5.4	7.1	2.2	9.5	30	4.8	6.2	1.8	8.2

Example 4.1 The response variable (Y) is pH and the predictors variables are Ca, Mg and base saturation. We assumed the linear model $Y = \beta_0 + \beta_1\mathrm{Ca} + \beta_2\mathrm{Mg} + \beta_3\mathrm{BS} + \epsilon$.

The fitted model is $\hat{Y} = 4.176061 + 0.146359\mathrm{Ca} - 0.109784\mathrm{Mg} + 0.009647\mathrm{BS}$ and the square of the multiple correlation coefficient, R^2, is 0.6106. The proposed algorithm was implemented in R with code reported in Appendix B. We set $x = 0.6106$, $\nu = 26$, $p = 3$, $prob = 0.975$, for the lower limit, and $prob = 0.025$ for the upper limit of the confidence interval. The estimated 95% confidence interval for ρ^2 is

$$\mathrm{IC}_{0.95}(\rho^2) : [0.257367, 0.777491].$$

For illustration purposes, we also apply our algorithm to high and low R^2 values.

Example 4.2 Taking base saturation (BS) as response (Y), Ca and pH as predictors, the fitted model is $\hat{Y} = 4.23622 + 1.41713\mathrm{Ca} - 0.83616\mathrm{pH}$ and R^2 equals 0.9499. We set $x = 0.9499$, $\nu = 27$, $p = 2$, $prob = 0.975$, for the lower limit, and $prob = 0.025$ for the upper limit of the confidence interval. The estimated 95% confidence interval is

$$\mathrm{IC}_{0.95}(\rho^2) : [0.8845673, 0.9748879].$$

Example 4.3 Finally, an application that results in a low R^2 value is obtained with the state.x77 data available in the `datasets` R package. Data refer to statistics (eight columns) for the 50 states (rows) of the United States of America. We consider per capita income (1974) as response (Y), Life expectancy in years ($1969 - 71$), and murder and non-negligent manslaughter rate per 100,000 population (1976) as predictors. The fitted model is $\hat{Y} =$

$-9027.01 + 188.36$Life Exp $+ 15.19$Murder and R^2 equals 0.119. For this application, we set $x = 0.119$, $\nu = 47$, $p = 2$, $prob = 0.975$, for the lower limit, and $prob = 0.025$ for the upper limit of the confidence interval. The estimated 95% confidence interval is

$$\mathrm{IC}_{0.95}(\rho^2) : [0.001772758, 0.3047753].$$

5. Monte Carlo simulations

Monte Carlo simulations were performed to verify the accuracy of confidence intervals for ρ^2. Random samples of $(p+1)$-dimensional multivariate normal distributions were generated considering the sample sizes $30, 50, 100$ and 200, and $\rho^2 = 0.1$, 0.5 and 0.9. The random vectors $[Y|\boldsymbol{X}^\top]^\top$ were simulated from a multivariate normal distribution, without loss of generality, with vector mean $\mathbf{0}$ $((p+1)\times 1)$ and covariance matrix $\boldsymbol{\Sigma} = (1-\varphi)\boldsymbol{I} + \varphi\boldsymbol{J}$, where $0 < \varphi < 1$, $\boldsymbol{I}$ is an identity matrix of order $p+1$ and $\boldsymbol{J}$ is a unitary matrix $(p+1)\times(p+1)$. The φ parameter was chosen after set the p value, to achieve a desired ρ^2 by trial and error using expression given in Equation (2.2). In each case, 1.000 samples were simulated and the coverage probability of the 95% confidence intervals was computed. A 99% exact binomial confidence interval of the coverage probability was computed to verify if the estimated level differs from the nominal level of 95%. This interval was computed using the pseudo random observation of 950 success obtained in 1.000 Monte Carlo simulations and was [0.9295, 0.9661]. Observed values that deviate from 0.95 but still are in this interval were considered a Monte Carlo error.

Results from the simulation study are reported in Table 2. Note that in almost all cases, the coverage probability is close to the confidence value and does not differ significantly $(P \leq 0.01)$ from the nominal confidence level of 95%. Only in one case, the coverage probability was significantly smaller and in 3 cases, they were greater than the nominal level of 95%, what is just a minor issue (Table 2). Then, for different sample sizes and strength of relationship between the response and predictor variables, assuming $(p+1)$-dimensional normality, the exact interval based on the proposed algorithm can be used while estimating ρ^2. A detected issue was due to the initial value of the noncentrality parameter, leading to convergence problems in the simulation study. To mitigate this issue, we choose the sample estimate to be the initial value in the algorithm. For example, for the case where $\rho^2 = 0.1$, $n = 30$ and $p = 10$, the coverage probability was 84.8% (Table 2). However, if we choose a fixed value of 0.01 to the initial value for each simulation, the coverage probability was 93.7%, that can be considered exact. Therefore, additional studies are needed to further understand the sensitivity in the choice of initial values.

Table 2. Coverage probability for 95% confidence intervals obtained through 1.000 Monte Carlo simulations of multivariate normal distributions, considering different values of ρ^2, p and n.

		n			
ρ^2	p	30	50	100	200
0.1	5	0.950	0.934	0.957	0.962
	10	0.848	0.938	0.951	0.950
0.5	5	0.958	0.972	0.945	0.951
	10	0.926	0.974	0.970	0.961
0.9	5	0.969	0.944	0.952	0.950
	10	0.957	0.968	0.958	0.944

6. Conclusions, limitations and future research

The algorithm for obtaining the noncentrality parameter of the distribution of the square of the sample multiple correlation coefficient was successfully proposed. Using real datasets, we apply it to obtain confidence intervals for the coefficient of determination of multiple regression models. For this, we use the `R` implementation, which is made available as Appendix B. Finally, a simulation study showed the good accuracy of the intervals obtained under different scenarios, allowing its use in many real applications.

In the simulation study we detected an issue related to the initial value of the noncentrality parameter. Possibly it is related to the use of the built-in gamma function in `R`. Our hypothesis is that a more precise implementation of this function would be enough to solve this issue. We started to investigate this by implementing our algorithm in `Java` and no such problem was detected. However, additional studies are needed.

Author contributions Conceptualization, D.F.; methodology, I.O., D.F.; software, I.O., D.F.; validation, I.O., D.F.; formal analysis, I.O.; investigation, I.O., D.F.; data curation, I.O.; writing-original draft preparation, I.O.,D.F.; writing-review and editing, I.O.; visualization, I.O., D.F.; supervision, I.O., D.F. All authors have read and agreed the published version of the manuscript.

Acknowledgements The authors thanks to the editors, associate editor, and the anonymous referees for their valuable comments to improve this manuscript.

Funding The authors thanks Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the financial support.

Conflicts of interest The authors declare no conflict of interest.

References

Baharev, A., and Kemény, S., 2008. On the computation of the noncentral F and noncentral beta distribution. Statistics and Computing 18(3), 333-340.

Benton, D., and Krishnamoorthy, K., 2003. Computing discrete mixtures of continuous distributions: noncentral chisquare, noncentral t and the distribution of the square of the sample multiple correlation coefficient. Computational Statistics & Data Analysis 43(2), 249-267.

Ferreira, D.F., Oliveira, I. R.C., and Toledo, F.H., 2012. ncg: Computes the noncentral gamma function. R package version 0.1.1.
http://CRAN.R-project.org/package=ncg

Krishnamoorthy, K., 2006. Handbook of Statistical Distributions with Applications. Chapman & Hall/CRC.

Krishnamoorthy, K., and Xia, Y., 2008. Sample size calculation for estimating or testing a nonzero multiple correlation coefficient. Multivariate Behavioral Research 43, 382-410.

Lee, Y.S., 1972. Tables of upper percentage points of the multiple correlation coefficient. Biometrika 59, 175-189.

Muirhead, R.J., 1982. Aspects of multivariate statistical theory. Wiley, New York.

Oliveira, I. R. C., and Ferreira, D. F., 2012. Computing the noncentral gamma distribution, its inverse and the noncentrality parameter. Computational Statistics (Zeitschrift) 28(4), 1663–1680.

Appendix A: The algorithm

Using the expressions presented in Section 3 we propose the following algorithm to obtain the noncentrality parameter, δ, of the R^2 distribution.

(1) set $0 \leq x \leq 1$, $\nu > 0$, $p > 0$ and $0 \leq prob \leq 1$;

(2) obtain $\alpha = p/2$ and $\beta = (\nu - p)/2$;

(3) set $errtol = 1 \times 10^{-12}$ for the maximum error and $maxitr$ for the maximum number of interactions;

(4) get an initial value of ρ^2, given by $d_{\text{new}} = d_0 (= x)$;

(5) set $it = 1$ and repeat steps (6) to (18) N_{max} times;

(6) $d = d_{\text{new}}$ and $k = \lceil \nu d / [2(1-d)] \rceil$;

(7) $a = \alpha + k$ and $b = \beta$;

(8) $betac = I_x(a, b)$ and $betad = betac$;

(9) $gxc = \dfrac{\Gamma(a+b-1)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^b$ and $gxd = gxc \times (a+b-1) \times x/a$;

(10) $tic = [2k(1-d) - \nu d]/[2d(1-d)]$ and $tid = tic$;

(11) calculate the kth term to the negative binomial probability and represent it by "pbnec" and "pbned" for forward and backward computations as follow:
$pbnec = \exp\{\ln\Gamma(\nu/2 + k) - \ln\Gamma(k+1) - \ln\Gamma(\nu/2) + k\ln(d) + (\nu/2)\ln(1-d)\}$ and $pbned = pbnec$;

(12) $remain = 1 - pbnec$;

(13) $cdf = pbnec \times betac$;

(14) calculate the first derivative at the kth term of the infinity series: $g = pbnec \times betac \times tic$;

(15) set $i = 1$;

(16) repeat steps (16)a to (16)t until convergence;

- a) starting forward sum:
  $gxc = gxc \times (a+b+i-2) \times x/(a+i-1)$;
- b) $betac = betac - gxc$;
- c) $tic = tic + 1/d$;
- d) $pbnec = pbnec \times d(\nu/2 + k + i - 1)/(k+i)$;
- e) $cdf = cdf + pbnec \times betac$;
- f) $g = g + pbnec \times betac \times tic$;
- g) $error = remain \times betac$;
- h) $remain = remain - pbnec$;
- i) getting backward sum, if there are still remaining terms. Thus,
  if $(i > k)$ then do:
  if $(error \leq errortol)$ or $(i > maxitr)$ go to step (17);
  do $i = i + 1$ and go to step (16);
  end of if $(i > k)$;
- j) else $(i \leq k)$. There are still remaining terms for the backwards computations and the steps (16).(16)k to (16).(16)t should be evaluate;
- k) $gxd = gxd \times (a - i + 1)/(x \times (a + b - i))$;
- l) $betad = betad + gxd$;
- m) $tid = tid - 1/d$;
- n) $pbned = pbned \times (k - i + 1)/(d(\nu/2 + k - i)$;
- o) $cdf = cdf + pbned \times betad$;
- p) $g = g + pbned \times betad \times tid$;
- q) $remain = remain - pbned$;
- r) if $(remain \leq errortol)$ or $(i > maxitr)$ go to (17);
- s) $i = i + 1$;
- t) go to step (16);

(17) if $d - (cdf - prob)/g \leq 0$, then $d_{\text{new}} = d/2$; else if $d - (cdf - prob)/g \geq 1$, then $d_{\text{new}} = d + (1-d)/2$; else $d_{\text{new}} = d - (cdf - prob)/g$ (Newton-Raphson's step);

(18) if $|d_{\text{new}} - d| \leq d \times tol$, where $tol = 1 \times 10^{-12}$, then return d_{new} and exit; else go to step (6), updating the interactions counter $(it = it + 1)$ before. Check if the maximum iterations N_{max} has been exceeded. If true, go to (19);

(19) print the error message: "iterative process did not converge in N_{max} steps."

Note that the first two conditions in step (17) refers to a protection for the Newton-Raphson step size. If the step size leads to a value to be used in the next step which is less than zero or greater than 1, the method ignores the Newton-Raphson formula and uses the expressions displayed in this step. Specifically, if the value for the new step is less than zero, the current value of the noncentrality parameter is too distant and far greater than the true value, which is the solution of the equation. Then, the value of the noncentrality parameter is reduced by half. On the other hand, if the new value of the noncentrality parameter exceeds 1 in the Newton-Raphson step, the value to be updated will be half the previous value plus 1/2.

Appendix B: R functions

This appendix contains the R functions used in the paper. First we implement the function lnFunGamaLanczos that should be used within the algorithm to obtain the noncentrality parameter (inrho). The use of the proposed algorithm to obtain a confidence interval for the coefficient of determination is illustrated. Finally, the R function to evaluate the coverage probability of confidence intervals is also made available.

ln of gamma fuction

```
# lngamma function
lnFunGamaLanczos <- function(z)
{
    Lanczos <- function(z)
    {
        lc <- c(5716.400188274341379136, -14815.30426768413909044,
                14291.49277657478554025, -6348.160217641458813289,
                1301.608286058321874105, -108.1767053514369634679,
                2.605696505611755827729, -0.7423452510201416151527e-2,
                0.5384136432509564062961e-7, -0.4023533141268236372067e-8)
        if (z < 0.5) return(log(pi) - log(sin(pi * z)) - Lanczos(1.0 - z)) else
        {
            g <- 9.0
            y <- z - 1.0
            lnfg <- 1.000000000000000174663 #rho0
            for (i in 1:10)
            {
                y <- y + 1.0
                lnfg <- lnfg + lc[i] / y
            }
            t <- z + g - 0.5;
            lnfg <- log(sqrt(2 * pi)) + (z - 0.5) *
                log(t) - t + log(lnfg);
            return(lnfg)
        }
    }
    if (z <= 0) # reflection formula
    {
        lnfg <- log(pi) - log(abs(sin(pi * z))) - Lanczos(1.0 - z);
    } else lnfg <- Lanczos(z);
    return(lnfg);
}
```

R function to obtain the noncentrality parameter of the distribution of the square of the sample coefficient of correlation (R^2)

```
inrho <- function(x,p,ni,prob)
{
    if ((x <= 0) | (x >= 1)) stop("x should be between 0 and 1!")
    if (ni < 0) stop("nu should be equal or greater than 0!")
    if ((p <= 0)) stop("p should be greater than 0!")
    alpha <- p / 2.0
    beta <- (ni - p) / 2.0
    b <- beta
    errortol <- 1e-12
    maxitr <- 5000
    dn <- 0.237
    it <- 1
    convnewton <- FALSE
    while (!convnewton)
    {
        d <- dn
        k <- ceiling(ni * d/(2 * (1 - d)))
        a <- alpha + k
        betac <- pbeta(x, a, b)
        betad <- betac
        gxc <- lnFunGamaLanczos(a + b - 1) - lnFunGamaLanczos(a) - lnFunGamaLanczos(b)
               + (a - 1) * log(x) + b * log(1 - x)
        gxc <- exp(gxc)
        gxd <- gxc * (a + b - 1) * x / a
        pbnec <- lnFunGamaLanczos(ni / 2 + k) - lnFunGamaLanczos(k + 1) -
                lnFunGamaLanczos(ni / 2) + k * log(d) + (ni / 2) * log(1 - d)
        pbnec <- exp(pbnec)
        pbned <- pbnec
        remain <- 1 - pbnec
        cdf <- pbnec * betac
        tic <- (2 * k *(1 - d)- ni * d) / (2 * d * (1 - d))
        tid <- tic
        g <- pbnec * betac * tic
        i <- 1
        convergiu <- FALSE
        while (!convergiu)
        {
            gxc <- gxc * (a + b + i - 2) * x / (a + i - 1)
            betac <- betac - gxc
            tic <- tic + 1.0 / d
            pbnec <- pbnec * d * (ni / 2 + k + i - 1) / (k + i)
            cdf <- cdf + pbnec * betac
            g <- g + pbnec * betac * tic
            error <- remain * betac
            remain <- remain - pbnec
            if (i > k)
            {
                if ((error <= errortol) | (i > maxitr)) convergiu <- TRUE
                i <- i + 1
            } else
            {
                gxd <- gxd * (a - i + 1) / (x * (a + b - i))
                betad <- betad + gxd
                tid <- tid - 1.0 / d
                pbned <- pbned * (k - i + 1) / (d * (ni / 2 + k - i))
                cdf <- cdf + pbned * betad
                g <- g + pbned * betad * tid
                remain <- remain - pbned
                if ((remain <= errortol) | (i > maxitr)) convergiu <- TRUE
                i <- i + 1
            }
        }#convergiu
        if ((d - (cdf - prob) / g) <= 0) dn <- d / 2 else
            if ((d - (cdf - prob) / g) >= 1) dn <- d + (1 - d) / 2.0 else
                dn <- d - (cdf - prob) / g
        if ((abs(dn - d) <= d * errortol) | (it > maxitr)) convnewton <- TRUE
        it <- it + 1
    }#convnewton
    return(dn)
}#general
```

Confidence interval for R^2: application to soil data

```
inrho(x=0.6106,p=3,ni=26,prob=0.975)

inrho(x=0.6106,p=3,ni=26,prob=0.025)
```

Coverage probability for confidence intervals

```
library(MASS)

pc <- function(N, n, p, rho)
{
    S <- (1-rho)*diag(p+1) + rho*(matrix(1,p+1,p+1))
    delta2 <- as.numeric(t(S[1,2:(p+1)])%*% solve(S[(2):(p+1),(2):(p+1)])%*%
        S[1,2:(p+1)]/S[1,1])
    cont <- 0
    for (i in 1:N){
    X <- mvrnorm(n,rep(c(0), times=p+1), S)
    reg <- lm(X[,1]~X[,2:(p+1)])
    R2 <- summary(reg)$r.squared
    nu <- anova(reg)[2,1]
    if (delta2 >= inrho(R2,p,nu,0.975)
        & delta2 <= inrho(R2,p,nu,0.025)) {
        y <- 1} else {y <- 0}
        if (y==1) cont <- cont + 1/N
    }
    return(cont)
}

N <- 1000
```

Coverage probability when $\mathrm{R}^2 = 0.5$, $p = 5$ and $n = 30$

```
p <- 5
rho <- 0.5741657565656567
n <- 30
pc_05_5_30 <- pc(N, n, p, rho)
```