Biometrika (2014), **101**, 3, *pp.* 748–754
Printed in Great Britain

doi: 10.1093/biomet/asu023
Advance Access publication 14 July 2014

Inference on multiple correlation coefficients with moderately high dimensional data

By SHURONG ZHENG

School of Mathematics & Statistics, Northeast Normal University, Renmin Street No. 5268, Changchun 130024, Jilin Province, China

zhengsr@nenu.edu.cn

DANDAN JIANG

School of Mathematics, Jilin University, Qianjing Street No. 2699, Changchun 130012, Jilin Province, China

jiangdandan@jlu.edu.cn

ZHIDONG BAI

School of Mathematics & Statistics, Northeast Normal University, Renmin Street 5268, Changchun 130024, Jilin Province, China

baizd@nenu.edu.cn

AND XUMING HE

Department of Statistics, University of Michigan, 1085 South University, Ann Arbor, Michigan 48109-1107, U.S.A.

xmhe@umich.edu

Summary

When the multiple correlation coefficient is used to measure how strongly a given variable can be linearly associated with a set of covariates, it suffers from an upward bias that cannot be ignored in the presence of a moderately high dimensional covariate. Under an independent component model, we derive an asymptotic approximation to the distribution of the squared multiple correlation coefficient that depends on a simple correction factor. We show that this approximation enables us to construct reliable confidence intervals on the population coefficient even when the ratio of the dimension to the sample size is close to unity and the variables are non-Gaussian.

Some key words: Independent component model; Multiple correlation; Testing.

1. Introduction

The multiple correlation coefficient is important in multivariate analysis and measures how strongly a given variable can be correlated with a linear combination of other variables. In this article, we use ρ for the population version of the multiple correlation between a one-dimensional random variable y and a $(p-1)$-dimensional random vector x, and R for the sample multiple correlation coefficient from a sample of size n. In linear regression of y on x, the coefficient of determination R^2 is routinely used to measure the strength of the linear relationship.

Multiple correlation has a long history and has been widely used. The distribution of R^2 was first investigated by Fisher (1928) under the assumption of multivariate normality. Various methods of deriving its distribution can be found in Wilks (1932), Garding (1941), Moran (1950), Williams (1978), Gurland & Asiribo (1991), Johnson et al. (1995), Nandi & Choudhury (2005), and Anderson (2003). It is well recognized that R^2 is a biased estimate of ρ^2, and an adjusted estimate of ρ^2 can be found in Anderson (2003, p. 153). Recent uses of multiple correlation can be found in Pena & Rodruguez (2002) for constructing a portmanteau test of fit and in Dutilleul et al. (2008) for assessing correlation in spatial processes.

Although the exact density of R^2 under normality is known, it is given by an infinite series when $\rho^2 > 0$ and thus is inconvenient to use. More importantly, the convergence of the series is slow, especially when n is large. For non-Gaussian variables, the density of R^2 is unavailable. Efforts have been made to allow inference on multiple correlation without normality. For example, Muirhead (1982) obtains the exact null distribution of R^2 under spherical symmetry, Ali & Nagar (2002) give the null distribution of multiple correlation for normal mixture models, and Ogasawara (2006) obtains approximations to the R^2 distribution by using the Edgeworth expansion. Existing work, however, does not provide effective inferences on multiple correlation when p is moderately large.

In this article, we provide an approximation for inference on multiple correlation coefficients under an independent component model. Our approximation is shown to work well for a wide range of dimensions $p < n$. The analysis of multiple correlation in high or ultra-high dimensions with $p \geqslant n$ is not attempted.

2. A quick review of multiple correlation

Let $Y_1, \dots, Y_n$ be a random sample from a p-dimensional population with covariance matrix Σ. Let $\bar{Y}$ and $\hat{\Sigma}$ be the usual sample mean and the sample covariance matrix. We consider the case $2 \leqslant p < n$. The covariance matrices are partitioned into

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{(1)}^{\mathrm{T}} \\ \sigma_{(1)} & \Sigma_{22} \end{pmatrix}, \quad \hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_{11} & a_{(1)}^{\mathrm{T}} \\ a_{(1)} & A \end{pmatrix},$$

where σ_{11} and $\hat{\sigma}_{11}$ are scalars. We further assume that Σ_{22} and A are nonsingular. Then the squared multiple correlation coefficients between the first component of Y and the other components, both for the population and for the sample, are

$$\rho^2 = \frac{\sigma_{(1)}^{\mathrm{T}} \Sigma_{22}^{-1} \sigma_{(1)}}{\sigma_{11}}, \quad R^2 = \frac{a_{(1)}^{\mathrm{T}} A^{-1} a_{(1)}}{\hat{\sigma}_{11}}.$$

If $Y_i \sim N(\mu, \Sigma)$, then for the hypothesis testing problem

$$H_0 : \rho = 0, \quad H_1 : \rho \neq 0,$$

the rejection region at significance level α is

$$\frac{R^2}{1 - R^2} \frac{n - p}{p - 1} > F_{p-1,n-p}(1 - \alpha)$$

where $F_{p-1,n-p}(1 - \alpha)$ is the $1 - \alpha$ quantile of the F distribution with degrees of freedom $p - 1$ and $n - p$. Under the alternative hypothesis H_1, the density $g(t; \rho^2)$ of R^2 (Anderson, 2003) is

$$\begin{aligned} g(t; \rho^2) &= \frac{(1 - t)^{(n-p-2)/2}(1 - \rho^2)^{(n-1)/2}}{\Gamma\{(n - p)/2\}\Gamma\{(n - 1)/2\}} \sum_{\mu=0}^{+\infty} \frac{\rho^{2\mu}\, t^{(p-1)/2+\mu-1}\Gamma^2\{(n - 1)/2 + \mu\}}{\mu!\Gamma\{(p - 1)/2 + \mu\}} \\ &= \frac{(1 - \rho^2)^{(n-1)/2}\, t^{(p-3)/2}(1 - t)^{(n-p)/2}}{B\{(n - p)/2, (p - 1)/2\}} F\{(n - 1)/2, (n - 1)/2, (p - 1)/2, t\rho^2\} \end{aligned} \quad (1)$$

where $F(\cdot,\cdot,\cdot,\cdot)$ is the hypergeometric function, $B(\cdot,\cdot)$ is the beta function and $\Gamma(\cdot)$ is the Gamma function. So a confidence interval for ρ^2 at the level of $(1-\alpha)100\%$ has lower and upper limits

$$L_A = \min_{\rho^2}\{\alpha/2 \leqslant G(R^2;\rho^2) \leqslant 1-\alpha/2\},$$
$$U_A = \max_{\rho^2}\{\alpha/2 \leqslant G(R^2;\rho^2) \leqslant 1-\alpha/2\}, \tag{2}$$

where $G(x;\rho^2)=\int_0^x g(t;\rho^2)\,\mathrm{d}t$ is the distribution function of R^2. Even under the Gaussian model, R^2 has an upward bias for ρ^2 when the true correlation is small.

THEOREM 1. *Under the normality assumption, $E(R^r) > \rho^r$ for any $r \geqslant 2$, when $(p-1)/(n-1) \geqslant \rho^2$ and $n > p$.*

Motivated by Gaussian models, Anderson (2003) (p. 153) considered an adjusted multiple correlation. By truncating the adjusted multiple correlation at zero, we use

$$R^{*2} = \max\left\{R^2 - \frac{p-1}{n-p}(1-R^2),\ \ 0\right\}$$

as the adjusted multiple correlation in this paper. We shall show that R^{*2} is consistent under a broader class of distributions and even when p grows proportionally to n. More importantly, we shall obtain a limiting distribution that enables us to construct approximate confidence intervals for ρ^2 for a wide range of p and n with $n > p$.

3. A NEW APPROXIMATION FRAMEWORK

We now assume the following independent component model for Y, that is, there exists a representation

$$Y_i = \begin{pmatrix} d_1 \\ D_2 \end{pmatrix} X_i \quad (i=1,\ldots,n), \tag{3}$$

where d_1 is a $1\times m$ row vector, D_2 is a $(p-1)\times m$ matrix of rank $p-1$ with $m \geqslant p$, and X_i is a random sample of m-random vectors with independently and identically distributed components.

Because the multiple correlation coefficient is invariant under location changes, any positive scale change on the first component of Y_i and any nonsingular affine transformation on the rest of the components, we assume without loss of generality that $E(X_i)=0$, $\mathrm{cov}(X_i)=I_m$, $\|d_1\|=1$, $D_2D_2^{\mathrm{T}}=I_{p-1}$, and $d_1D_2^{\mathrm{T}}=(\rho,0,\ldots,0)$. We study the asymptotic properties of R^2 under the setting $n\to\infty$ and $q_n = p/n \to q \in [0,1)$. To begin, let

$$\sigma^2(t) = 2\{q+(1-q)t\}^2 - 2\{-2(1-q)t^2+4(1-q)t+2q\}\{q+(1-q)t-1/2\}. \tag{4}$$

We now state our main results, followed by a discussion on the assumptions.

THEOREM 2. *Under the independent component model (3), and if X_i has finite fourth moment and the L_∞-norm $\|d_1\|_\infty = o(1)$, then $R^{*2}\to\rho^2$ almost surely, and*

$$n^{1/2}\{R^2 - q_n - (1-q_n)\rho^2\} \to N\{0,\sigma^2(\rho^2)\}$$

in distribution as $n\to\infty$ and $q_n = p/n \to q < 1$. As a consequence, we have, if $\rho^2 > 0$, then

$$n^{1/2}(R^{*2}-\rho^2) \to N\{0,\sigma^2(\rho^2)/(1-q)^2\}, \tag{5}$$
$$n^{1/2}\{g(R^{*2})-g(\rho^2)\} \to N(0,1) \tag{6}$$

in distribution, where $g(x)=\int_0^x (1-q)/\sigma(t)\,\mathrm{d}t$ for $x\geqslant 0$. If $\rho^2=0$, we have $n^{1/2}(R^2-q_n)\to N\{0,2q(1-q)\}$, and

$$(2n)^{1/2}\{\arccos(R) - \arccos(q_n^{1/2})\} \to N(0,1)\,. \tag{7}$$

The variance $\sigma^2(\rho^2)$ in (5) can be estimated by (4) with $t = R^{*2}$. Since $g(x)$ is monotone and can be computed numerically, Theorem 2 allows the construction of two asymptotic confidence intervals for ρ^2, one based on (5) and the other based on the variance-stabilizing transformation (6). For example, an $(1-\alpha)100\%$ asymptotic confidence interval for ρ^2 based on (6) is (L, U), where

$$L = \min\left\{0 \leqslant x \leqslant 1 : g(x) \geqslant g(R^{*2}) - n^{-1/2} z_{\alpha/2}\right\}$$
$$U = \max\left\{0 \leqslant x \leqslant 1 : g(x) \leqslant g(R^{*2}) + n^{-1/2} z_{\alpha/2}\right\}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal. For testing the null hypothesis that $\rho^2 = 0$, the asymptotic distribution of (7) may be used.

The independent component model generalizes the multivariate Gaussian model, but the representation (3) is not required to be unique. The latent variable X_i could be normal, but need not be. The assumption that $\|d_1\|_\infty = o(1)$ in Theorem 2 requires that the first component of Y_i is not dominated by a small number of components in X_i as p increases, but this condition can be removed if each component of X_i has fourth moment equal to 3 as in the Gaussian case.

Our main result is obtained under the assumption that $q_n = p/n \to q < 1$. A natural question is why we are interested in this limit, but not others, such as $p \propto n^r$ for some positive r. In a given application where both n and p are fixed, we cannot distinguish one value of r from another. What matters is which limiting distribution provides a good approximation. We find that the framework of $p \propto n$ has the advantage of an asymptotic normal approximation that works well for a wide range of (n, p) values with $n > p$. Similar observations have been made in Bai & Saranadasa (1996), Bai et al. (2009), Chen & Qin (2010), and Chen et al. (2010), and we shall confirm this through simulation in § 4.

4. Empirical evidence

In this section, we evaluate the performance of the asymptotic confidence intervals for ρ^2 based on (5) and (6), and compare them to confidence intervals based on the Edgeworth expansion method of Ogasawara (2006), the classical method of Anderson (2003) in (2), and ABC bootstrap confidence intervals as described in Davison & Hinkley (1997 § 5).

Bootstrap methods are ill-suited for datasets with p/n close to unity, because many bootstrap samples will not have enough distinct observations, and when the number of distinct observations is p or less, the value of R^2 cannot be computed. For example, if $(n, p) = (25, 20)$, only 1 out of 100 bootstrapped samples is expected to have more than p distinct observations. In this paper, we adopt the strategy of adding a small random perturbation to each observation in the bootstrapped samples before the R^{*2} are calculated.

The data are generated under the independent component model (3), with the following specifications, with $k = p(1+\rho)/2$ for the given values of p and ρ considered in our study. In our study, k is an integer between 2 and $p-1$.

(i) The vector d_1 has all its elements equal to $p^{-1/2}$.
(ii) The first row of D_2 has its first k elements equal to $p^{-1/2}$ and the rest equal to $-p^{-1/2}$.
(iii) For $l = 2, \ldots, k$, the lth row of D_2 has the first $l-1$ elements equal to $l^{-1/2}(l-1)^{-1/2}$, the lth element equal to $-l^{-1/2}(l-1)^{1/2}$ and the rest equal to 0.
(iv) For $l = k+1, \ldots, p$, the lth row of D_2 has the first k elements equal to 0, the next $l-k-1$ elements equal to $(l-k+1)^{-1/2}(l-k)^{-1/2}$, the lth element equal to $-(l-k+1)^{-1/2}(l-k)^{1/2}$, and the rest equal to 0.

We consider two scenarios of $q_n = p/n = 0{\cdot}2$ and $0{\cdot}8$, three values of $\rho = 0{\cdot}2, 0{\cdot}6, 0{\cdot}8$, and three values of moderate to high dimensions $p = 20, 30, 60$ at $q = 0{\cdot}2$ and $p = 20, 60, 200$ at $q = 0{\cdot}8$. With the above construction of d_1 and D_2, the desired ρ values are achieved. Table 1 reports the average coverage probability for nominal 95% intervals and interval length of ρ^2 obtained from 10 000 Monte Carlo samples

Table 1. *Coverage (%) and average length (in parentheses) of 95% confidence intervals for ρ^2 when the X_{ij} are uniformly distributed*

p		$\rho = 0{\cdot}2$	$\rho = 0{\cdot}6$	$\rho = 0{\cdot}8$
		$q = 0{\cdot}2$		
20	M1	95·9 (0·19)	94·9 (0·33)	94·9 (0·26)
	M2	98·7 (0·21)	95·3 (0·34)	96·6 (0·25)
	M3	71·6 (0·39)	90·9 (0·39)	93·7 (0·24)
	A1	90·5 (0·06)	0·0 (0·07)	0·0 (0·07)
	B1	43·6 (0·05)	67·8 (0·30)	76·9 (0·24)
30	M1	95·2 (0·17)	94·7 (0·27)	95·2 (0·21)
	M2	98·6 (0·18)	95·1 (0·28)	96·3 (0·20)
	M3	47·7 (0·34)	85·1 (0·34)	89·0 (0·22)
	A1	69·5 (0·04)	0·0 (0·05)	0·0 (0·05)
	B1	28·1 (0·02)	58·5 (0·26)	67·7 (0·20)
60	M1	95·4 (0·13)	95·3 (0·20)	95·0 (0·15)
	M2	98·4 (0·13)	94·8 (0·20)	95·1 (0·14)
	M3	5·6 (0·26)	62·8 (0·27)	78·4 (0·18)
	A1	0·0 (0·02)	0·0 (0·02)	0·0 (0·02)
	B1	7·4 (0·01)	32·8 (0·19)	45·6 (0·15)
		$q = 0{\cdot}8$		
p		$\rho = 0{\cdot}2$	$\rho = 0{\cdot}6$	$\rho = 0{\cdot}8$
20	M1	93·2 (0·52)	92·4 (0·68)	93·7 (0·83)
	M2	95·2 (0·76)	94·3 (0·78)	94·4 (0·75)
	M3	0·7 (0·62)	3·1 (0·44)	4·5 (0·26)
	A1	95·6 (0·20)	0·0 (0·23)	0·0 (0·25)
	B1	25·6 (0·12)	26·2 (0·18)	26·6 (0·20)
60	M1	95·1 (0·47)	94·6 (0·60)	94·6 (0·69)
	M2	95·7 (0·56)	95·7 (0·63)	94·4 (0·55)
	M3	0·0 (0·45)	0·0 (0·32)	0·0 (0·18)
	A1	90·5 (0·07)	0·0 (0·09)	0·0 (0·09)
	B1	9·7 (0·02)	10·9 (0·07)	13·2 (0·10)
200	M1	95·0 (0·29)	95·0 (0·46)	94·6 (0·33)
	M2	96·5 (0·37)	95·0 (0·46)	94·5 (0·30)
	M3	0·0 (0·30)	0·0 (0·22)	0·0 (0·13)
	A1	0·0 (0·03)	0·0 (0·03)	0·0 (0·03)
	B1	1·4 (0·01)	1·0 (0·02)	1·5 (0·04)

M1 and M2, asymptotic confidence intervals based on Theorem 2; M3, Edgeworth expansion method of Ogasawara (2006); A1, the classical method of Anderson (2003) in (2); B1, ABC bootstrap method.

when each component of X_i is uniformly distributed. A similar table is given in the Supplementary Material when X_i is normally distributed. It is clear from the tables that our proposed methods attain decent coverage probabilities in all cases. The competing methods fail badly for different reasons. More specifically, the method based on Edgeworth expansion of Ogasawara (2006) uses R^2 as the estimate of ρ^2 with an upward bias approximately $q(1-\rho^2)$. At $q = 0{\cdot}9$ and $\rho^2 = 0{\cdot}1$, the bias is more than 0·8, which leads to nearly zero coverage. The ABC method is biased because the bootstrap distribution of R^{*2} is unduly skewed towards the value 1 as we explained earlier, leading very low coverage probabilities.

In the Supplementary Material, we include the power curve for testing the null hypothesis that $\rho^2 = 0$ versus $\rho^2 > 0$ based on the distribution (7) at $p = 60$ and $q_n = 0{\cdot}2$, 0·6 and 0·8. The significance levels of the tests are close to the nominal level of 0·05, but the power of the test decreases quickly with q_n, reflecting the challenges of statistical inference in high dimensions. In the Supplementary Material, we also provide additional simulation results for two other models with different choices of d_1 and D_2.

Finally, we demonstrate the merit of the proposed method on the diffuse large-B-cell lymphoma microarray data of Rosenwald et al. (2002). The dataset contains the gene expression measurements of over 7399 genes as well as the survival times of 240 patients. We first select 120 genes that have the highest marginal correlation with the survival time, and then estimate the multiple correlation between survival time and the expressions of those genes. In this case, we have $q_n \approx 0{\cdot}5$, and the 95% confidence intervals are (0·15, 0·46) and (0·16, 0·47) from our proposed methods. The ABC bootstrap method would yield a confidence interval (0·12, 0·16), which appears narrow but biased. When we performed the same analysis by taking 120 randomly selected genes, the 95% confidence interval based on our proposed method (5) was [0, 0·29), with $R^2 = 0{\cdot}57$ and $R^{*2} = 0{\cdot}14$. The bootstrap distribution of R^{*2} has nearly all of its probability mass above 0·40, indicating again the problem of the bootstrap in analysing data of high dimensions. Our proposed method is able to produce a trustworthy confidence interval with minimal computational cost.

When the dimension of variables p is moderately large relative to the sample size n, classical methods of inference on the multiple correlation coefficients fail. In this paper, we propose a simple framework for approximating the distribution of the sample multiple correlation. The asymptotic approximation is easy to use but remains effective for a wide range of scenarios in terms of (n, p) as long as p/n is bounded away from 1. The problem of even higher dimensions, $p \geqslant n$, is not considered in this paper.

Acknowledgement

The authors thank the reviewers for their helpful comments and suggestions that led to an improvement of this article. This research was partially supported by the National Science Foundation, U.S.A., the National Natural Science Foundation of China, the Program for New Century Excellent Talents in University and the Program for Changjiang Scholars and Innovative Research Teams in University.

Supplementary material

Supplementary material available at *Biometrika* online includes the proof of Theorem 2 and additional simulation results, including the distributions of the p-values from various tests on ρ^2.

Appendix

Proof of Theorem 1

Our first step is to show

$$E(R^2) > \rho^2. \tag{A1}$$

Following (1), we have

$$E(R^2) = \frac{(1-\rho^2)^{(n-1)/2}}{\Gamma\{(n-1)/2\}} \sum_{\mu=0}^{\infty} \frac{\rho^{2\mu}\Gamma\{(n-1)/2+\mu\}}{\mu!} \frac{p-1+2\mu}{n-1+2\mu}.$$

Moreover, we can write

$$\rho^2 = \frac{(1-\rho^2)^{(n-1)/2}}{\Gamma\{(n-1)/2\}} \sum_{\mu=0}^{\infty} \frac{\rho^{2\mu}\Gamma\{(n-1)/2+\mu\}}{\mu!} \rho^2.$$

If $(p-1)/(n-1) \geqslant \rho^2$, then (A1) is obviously true because $(p-1+2\mu)/(n-1+2\mu) > (p-1)/(n-1) \geqslant \rho^2$ for all $\mu \geqslant 1$. By Jensen's inequality we then have $E(R^r) > \rho^r$ for any $r \geqslant 2$.

References

Ali, H. & Nagar, D. K. (2002). Null distribution of multiple correlation coefficient under mixture normal model. *Int. J. Math. Math. Sci.* **30**, 249–55.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis.* New York: Wiley.

Bai, Z. D. & Saranadasa, H. (1996) Effect of high dimension comparison of significance tests for a high dimensional two sample problem. *Statist. Sinica* **6**, 311–29.

Bai, Z. D., Jiang, D. D., Yao, J. F. & Zheng, S. R. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.* **37**, 3822–40.

Chen, S. X. & Qin, Y. L. (2010). A two sample test for high dimensional data with application to gene-set testing. *Ann. Statist.* **38**, 808–35.

Chen, S. X., Zhang, L. X. & Zhong, P. S. (2010) Testing high-dimensional covariance matrices. *J. Am. Statist. Assoc.* **105**, 810–9.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. New York: Cambridge University Press.

Dutilleul, P., Pelletier, B. & Alpargu, G. (2008). Modified F tests for assessing the multiple correlation between one spatial process and several others. *J. Statist. Plan. Infer.* **138**, 1402–15.

Fisher, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. *Proc. R. Soc. Lond.* **121**, 654–73.

Garding, L. (1941). The distributions of the first and second order moments, the partial correlation coefficients and the multiple correlation coefficient in samples from a normal multivariate population. *Skand. Aktuar.* **24**, 185–202.

Gurland, J. & Asiribo, O. (1991). On the distribution of the multiple correlation coefficient and the kinked chi-square random variable. *Statist. Sinica* **1**, 493–502.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995). *Continuous Univariate Distributions*. New York: Wiley.

Moran, P. A. P. (1950). The distribution of the multiple correlation coefficient. *Proc. Camb. Phil. Soc.* **46**, 521–2.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New York: Wiley.

Nandi, S. B. & Choudhury, S. (2005) Series representation of non-null distribution of the square of sample multiple correlation coefficient by use of the Mellin integral transform. *Commun. Statist.* A **34**, 679–85.

Ogasawara, H. (2006). Asymptotic expansion and conditional robustness for the sample multiple correlation coefficient under nonnormality. *Commun. Statist.* B **35**, 177–99.

Pena, D. & Rodruguez, J. (2002). A powerful portmanteau test of lack of fit for time series. *J. Am. Statist. Assoc.* **97**, 601–10.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L. et al. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New Engl. J. Med.* **346**, 1937–47.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika* **24**, 471–94.

Williams, E. J. (1978). A simple derivation of the distribution of the multiple correlation coefficient. *Commun. Statist.* A **7**, 1413–20.

[*Received January* 2013. *Revised April* 2014]