EFFECT OF TWO-STAGE SAMPLING ON THE R-SQUARE STATISTIC

Govinda J. Weerakkody, Sumalee Givaruangsawat, Mississsippi State University *
Govinda J. Weerakkody, Mathematics and Statistics, Mississippi State, MS 39762

Keywords: Two-stage Sampling, R^2-statistic, intracluster correlation.

1 Introduction

R^2 is a popular and very widely used statistic in regression analysis to quantify the linear association between a response variable Y and p predictor variables, $X_1, \cdots, X_p$. Let ρ^2 represent the squared multiple correlation coefficient between Y and $X_1, \cdots, X_p$. Numerous articles are available in statistical literature on the properties of R^2 as an estimator of ρ^2 when the observations are uncorrelated. However, relatively little is known about the behavior of R^2 when the available observations are correlated such as the data that result from complex sampling schemes. For an example, two-stage cluster sampling results in correlated data within clusters, with positive intracluster correlation ρ^*. In addition, correlated observations also arise in split-plot type industrial experiments where some measurements are taken on a larger-sized experimental unit and some other measurements are taken on smaller-sized experimental units within a larger-sized experimental unit. In this case, measurements taken within the same larger sized experimental unit tend to be correlated. In this paper, we study the quality of R^2 as an estimator of ρ^2 in the presence of such correlated data.

Let $\underline{z}'_{ij} = (y_{ij}, \underline{x}'_{ij})$, with $\underline{x}'_{ij} = (x_{ij1}, \cdots, x_{ijp})$, for $j = 1, \ldots, n$ and $i = 1, \ldots, k$ be a vector of responses containing a y-measurement and p x-measurements on the j th subject within the i th cluster. Data set contain information on kn subjects from k clusters, each with n subjects. It is assumed that $\underline{z}_{ij}$ follows a $(1+p)$-variate normal distribution with a mean vector $\underline{\mu}' = (\mu_y, \underline{\mu}'_x)$ and a $(1+p)\times(1+p)$ covariance matrix Σ having the structure

$$\Sigma = \begin{pmatrix} \sigma_{yy} & \underline{\sigma}'_{yx} \\ \underline{\sigma}_{yx} & \Sigma_{xx} \end{pmatrix}.$$

It is also assumed that $cov(\underline{z}_{ij}, \underline{z}_{i'j'}) =$

$$\begin{cases} \Sigma \otimes \begin{pmatrix} 1 & \rho^* \\ \rho^* & 1 \end{pmatrix}, & \text{for } i = i', j \neq j' \\ \Sigma \otimes I_2, & \text{for } i \neq i' \end{cases}$$

with $\rho^* > 0$ where $\otimes$ represents the Kroneker product of matrices. Let $\underline{z}'_i = (\underline{z}'_{i1}, \underline{z}'_{i2}, \ldots, \underline{z}'_{in})$, for $i = 1, 2, \ldots, k$, $\underline{z}' = (\underline{z}'_1, \underline{z}'_2, \ldots, \underline{z}'_k)$ and

$$\Sigma_{\rho^*} = (1-\rho^*)(I_n - \frac{1}{n}J_n) + (1+(n-1)\rho^*)\frac{1}{n}J_n. \quad (1)$$

Thus,

$$\underline{z} \sim N(\underline{\mu} \otimes \underline{j}_n \otimes \underline{j}_k, \quad \Sigma \otimes \Sigma_{\rho^*} \otimes I_k). \quad (2)$$

That is, the observations within a cluster are equally correlated with positive intraclass correlation coefficient ρ^* and observations from different clusters are uncorrelated. It should be pointed out that the above model is somewhat limited in that it assumes the intracluster correlation of each variable to be equal to ρ^*.

Let $\underline{Y}'_i = (y_{i1}, y_{i2}, \ldots, y_{in})$, $X'_i = (\underline{x}_{i1}, \cdots, \underline{x}_{in})$, $\underline{Y}' = (\underline{Y}'_1, \underline{Y}'_2, \ldots, \underline{Y}'_n)$, $X' = (X'_1, \cdots, X'_k)$, $\beta_0 = \mu_y - \underline{\sigma}'_{xy}\Sigma^{-1}_{xx}\underline{\mu}_x$, and $\underline{\beta} = \underline{\sigma}'_{xy}\Sigma^{-1}_{xx}$. Then the matrix form of the regression model that describes the linear relationship between y_{ij} and $\underline{x}'_{ij}$ can be expressed as :

$$\underline{Y} = j_{kn}\beta_0 + X\underline{\beta} + \underline{\epsilon}, \underline{\epsilon} \sim (0, \sigma^2\Sigma_{\rho^*} \otimes I_k) \quad (3)$$

where $\sigma^2 = \sigma_{yy} - \underline{\sigma}'_{xy}\Sigma^{-1}_{xx}\underline{\sigma}_{xy}$ and Σ_{ρ^*} as given in (1). Lack of independency among the observation within clusters invalidates usual statistical inference based on the ordinary least squares method. The different aspects of the effect of the correlated structure as described in equation (2) on statistical inference are

*Authors wish to thank Professor Dallas E. Johnson for suggesting this problem

given in Campbell (1977), Scott and Holt (1982), Christensen (1984), Thomas and Rao (1987), Wu, Holt, and Holms (1988), Weerakkody and Johnson (1992), Rao, Sutradhar, and Yue (1993), and Weerakkody and Givaruangsawat (1995). Weerakkody and Johnson (1992) discussed the estimation of $\underline{\beta}$ in model (3), Wu, Holt, and Holms (1988) studied the effect of two-stage sampling on the testing hypotheses about $\underline{\beta}$ and Rao, Sutradhar, and Yue (1993) proposed a transformation which makes the resulting transformed data uncorrelated which then was used to develop a statistical test for testing hypotheses about β. Our focus is the estimation of ρ^2. The squared multiple correlation between $\underline{Y}$ and the p x-variables is given by

$$\rho^2 = \frac{\underline{\sigma}_{xy}\Sigma_{xx}^{-1}\underline{\sigma}_{xy}}{\sigma_{yy}}.$$

It is well known that, if the data were uncorrelated, ρ^2 is estimated using

$$R^2 = \frac{\underline{Y}'X(X'X)^{-1}X'\underline{Y}}{\underline{Y}'(I_{nk} - (1/n)J_n \otimes (1/k)J_k)\underline{Y}} \qquad (4)$$

and for large sample sizes R^2 will have the following properties:

$$\begin{aligned} Var(R^2) &= \frac{4(kn-p-1)^2\rho^2(1-\rho^2)^2}{(kn-1)(kn+1)(kn+3)} \\ &+ o\left((kn)^{-2}\right) \\ Bias(R^2) &= \frac{p}{kn-1}(1-\rho^2) - \\ &\quad \frac{2(kn-p-1)}{(kn-1)(kn+1)}\rho^2(1-\rho^2) \\ &+ o\left((kn)^{-2}\right). \qquad (5) \end{aligned}$$

Due to the correlation structure given in (2), the usual R^2 given in equation (4) does not have the same properties that it has in the presence of uncorrelated data. In particular, large sample properties of the usual R^2 given in equations (5) are not valid for two-stage sampling data. As such we derived approximate variance and bias formulas of the usual R^2 for data with correlation structure described in equation (2) and are given in the section below.

2 The Behavior of the usual R^2-statistic

Now we discuss the properties of the usual R^2 as an estimator of ρ^2 in the presence of correlated data as described in (2). If ρ^* is small, one would expect little or no impact on the quality of R^2 as an estimate for ρ^2. On the other hand, if ρ^* is large indicating strong dependency among the observations within clusters, one expects R^2 to perform poorly. This is expected because the effective dimensionality of the data would be much less than kn when ρ^* is large. In order to study the behavior of R^2 in the presence of correlated data, we present approximate expresssions for the variance and the bias of R^2 in the following theorem.

Theorem: Let kn data points be correlated as described in (2) and R^2 be as defined in (4). Then for large values of k and n

$$\begin{aligned} Var(R^2) &= \frac{4(m-p)^2}{m(m+2)(m+4)}\rho^2(1-\rho^2)^2 \\ &+ o\left((kn)^{-2}\right) \\ Bias(R^2) &= \frac{p}{m}(1-\rho^2) - \frac{2(m-p)}{m(m+2)}\rho^2(1-\rho^2) \\ &+ o\left((kn)^{-2}\right) \end{aligned}$$

where $m = \frac{(k(n-1)+R(\rho^*)(k-1))^2}{k(n-1)+[R(\rho^*)]^2(k-1)}$ and $R(\rho^*) = \frac{1+(n-1)\rho^*}{1-\rho^*}$.

Before proving the theorem above, we establish the following results.

Lemma 1: Let

$$\begin{aligned} A_W &= \sum_{i=1}^{k}\sum_{j=1}^{n}(\underline{z}_{ij} - \bar{\underline{z}}_i)(\underline{z}_{ij} - \bar{\underline{z}}_i)' \\ A_B &= n\sum_{i=1}^{k}(\bar{\underline{z}}_i - \bar{\underline{z}})(\bar{\underline{z}}_i - \bar{\underline{z}})' \qquad (6) \end{aligned}$$

where $\bar{\underline{z}}_i = \frac{1}{n}\sum_{j=1}^{n}\underline{z}_{ij}$ and $\bar{\underline{z}} = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n}\underline{z}_{ij}$. Then

$$\begin{aligned} A_W &\sim W_{p+1}\left[(1-\rho^*)\Sigma, k(n-1)\right], \\ A_B &\sim W_{p+1}\left[(1+(n-1)\rho^*)\Sigma, (k-1)\right], \qquad (7) \end{aligned}$$

and A_W and A_B are independently distributed where $W_{p+1}(\Sigma, m)$ represents the central Wishart distribution with covariance-variance matrix Σ and degrees of freedom m.

Proof: Let $P_{n\times(n-1)}$ be the matrix such that $PP' = (I_n - \frac{1}{n}J_n)$ and $P'P = I_{n-1}$. Such P exists since $(I_n - \frac{1}{n}J_n)$ is an idempotent matrix of rank $n-1$. Note that equation (1) implies that the (n-1) columns of matrix P are the eigenvectors of Σ_{ρ^*} corresponding to the eigenvalue $(1-\rho^*)$ and the remaining eigenvector of Σ_{ρ^*} is $(1/n)j_n$ corresponding to the eigenvalue $1+(n-1)\rho^*$. Now, let

$[\underline{z}_{i1}(w), \cdots, \underline{z}_{i(n-1)}(w)] = [\underline{z}_{i1}, \cdots, \underline{z}_{in}]P$ and $\bar{\underline{z}}_i = [\underline{z}_{i1}, \cdots, \underline{z}_{in}](1/n)j_n$ for $i = 1, \cdots, k$. Then by equation (1), $\underline{z}_{ij}(w)$'s are a sample of k(n-1) independent variates from $N(0, (1-\rho^*)\Sigma)$. Further, since

$$\begin{aligned} A_W &= \sum_{i=1}^{k}\sum_{j=1}^{n}(\underline{z}_{ij} - \bar{\underline{z}}_i)(\underline{z}_{ij} - \bar{\underline{z}}_i)' \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n-1}\underline{z}_{ij}(w)\underline{z}'_{ij}(w), \end{aligned}$$

$A_W \sim W_{p+1}((1-\rho^*)\Sigma, k(n-1))$. In addition, equations (10 and (2) also imply that the $\bar{\underline{z}}_i$'s are a sample of k independent variates from $N(\mu, \frac{1+(n-1)\rho^*}{n}\Sigma)$ and hence

$$\begin{aligned} A_B &= n\sum_{i=1}^{k}(\bar{\underline{z}}_i - \bar{\underline{z}})(\bar{\underline{z}}_i - \bar{\underline{z}})' \\ &\sim W_{p+1}((1+(n-1)\rho^*)\Sigma, k-1). \end{aligned}$$

Finally, since $\underline{z}_{ij}(w)$'s and $\bar{\underline{z}}_i$'s are independent, so are A_W and A_B. This completes the proof of the Lemma 1.

Next, we establish the following result which approximates the distribution of the weighted average of two independent central Wishart matrices. This approximation is similar to that given in Satterthwaite (1946) for the Chi-Square variates.

Lemma 2: Suppose $W_1 \sim W_{p+1}(\Sigma, m_1)$, $W_2 \sim W_{p+1}(\Sigma, m_2)$, and that W_1 and W_2 are independent. Then for large m_1 and m_2,

$$\begin{aligned} W &= \frac{1}{m_1+m_2}[a_1W_1 + a_2W_2] \\ &\sim \frac{a}{m_1+m_2}W_{p+1}(\Sigma, m). \end{aligned} \quad (8)$$

where $a = \frac{a_1^2m_1+a_2^2m_2}{a_1m_1+a_2m_2}$ and $m = \frac{(a_1m_1+a_2m_2)^2}{a_1^2m_1+a_2^2m_2}$.

Proof: Since W_1 and W_2 are independently distributed, the characteristic function of W is given by

$$\begin{aligned} \phi_W(T) &= E\left\{exp\left(\frac{-itrace(a_1W_1T + a_2W_2T)}{m_1+m_2}\right)\right\} \\ &= \phi_{W_1}(\frac{a_1}{m_1+m_2}T)\phi_{W_2}(\frac{a_2}{m_1+m_2}T) \\ &= |I_{p+1} - \frac{2ia_1}{m_1+m_2}T\Sigma|^{-m_1/2} \\ &\quad |I_{p+1} - \frac{2ia_2}{m_1+m_2}T\Sigma|^{-m_2/2}. \end{aligned}$$

Let λ_j for $j = 1, \cdots, 1+p$ be eigenvalues of $T\Sigma$. Then $-2ln(\phi_W(T)) =$

$$\begin{aligned} & m_1\sum_{j=1}^{p+1}ln(1 - 2i\tfrac{a_1}{m_1+m_2}\lambda_j) \\ & +m_2\sum_{j=1}^{p+1}ln(1 - 2i\tfrac{a_2}{m_1+m_2}\lambda_j) \\ \approx\ & m_1\sum_{j=1}^{p+1}\left(-2i\tfrac{a_1}{m_1+m_2}\lambda_j\right) + \left(-4\tfrac{a_1^2}{(m_1+m_2)^2}\lambda_j^2\right) \\ & +m_2\sum_{j=1}^{p+1}\left(-2i\tfrac{a_2}{m_1+m_2}\lambda_j\right)\left(-4\tfrac{a_2^2}{(m_1+m_2)^2}\lambda_j^2\right) \\ =\ & m\sum_{j=1}^{p+1}\left(-2i\tfrac{a}{m_1+m_2}\lambda_j\right) + \left(-4\tfrac{a^2}{(m_1+m_2)^2}\lambda_j^2\right) \\ \approx\ & m\sum_{j=1}^{p+1}ln(1 - 2i\tfrac{a}{m_1+m_2}\lambda_j) \\ =\ & -2ln\left(|I_{p+1} - 2i\tfrac{a}{m1+m_2}T\Sigma|^{-m/2}\right). \end{aligned}$$

Therefore,
for large samples, $W \dot{\sim} \frac{a}{m_1+m_2}W_{p+1}(\Sigma, m)$. This completes the proof of the lemma 2. Now we the prove Theorem.

Proof of Theorem: Using the result given in lemma 1 and substituting $m_1 = k(n-1)$, $m_2 = k-1$, $a_1 = (1-\rho^*)$, and $a_2 = (1+(n-1)\rho^*)$ in the lemma 2 above,

$$\frac{1}{kn-1}(A_W + A_B) \dot{\sim} \frac{a}{kn-1}W_{p+1}(\Sigma, m)$$

where $a = \frac{k(n-1)+[R(\rho^*)]^2(k-1)}{k(n-1)+R(\rho^*)(k-1)}$ and m is as given in the theorem. The variance and the bias results of R^2 in the theorem directly follows from the above since R^2 is derived from $A = (A_W + A_B)$ using the invariance principle. of maximum likelihood estimators. This completes the proof of the theorem.

When $\rho^* = 0$, $R(\rho^*) = 1$, $m = kn-1$, and hence the results given in the theorem coincides with those in equations (5). Note that the equations (5) contain $var(R^2)$ and $Bias(R^2)$ in the presence of kn uncorrelated data. Let R_W^2 and R_B^2 be the usual estimators of ρ^2 based on within-cluster information and between-cluster information, respectively. When $\rho^* \to 1$ (i.e., when the observations within a cluster are perfectly correlated), $m \to (k-1)$ and hence the variance and the bias of R^2 is similar to those of R_B^2. As such, the usual R^2 estimator is not quite acceptable as a estimator of ρ^2, particularly when the intracluster correlation is large. Based on results of a simulation study that we conducted, we recommend the use of R_W^2 to estimate ρ^2 when k is small. For large k, we recommend $wR_W^2 + (1-w)R_B^2$ where $w = \frac{k(n-1)}{kn-1}$ to estimate ρ^2.

References

[1] Christensen, R. (1984), *A Note on Ordinary Least Squares Methods for Two-Stage Sampling,* Journal of the American Statistical Association, 79, 720-721.

[2] Holt, D., and Scott, A.J. (1981), *Regression Analysis Using Survey Data,* The Statistician, 30, 169-178.

[3] Rao, J.N.K., Sutradahar, B.C., and Yue, K. (1993), *Generalized Least Squares F-test in Regression Analysis with Two-Stage Cluster Sampling,* Journal of the American Statistical Association, 88, 1388-1391.

[4] Satterhwaite, F.E. (1946), *An Approximate Distribution of Estimates of Variance Component,* Biometrics, 2, 110-114.

[5] Scott, A.J. and Holt, D., (1982), *The Effect of Two-Stage Sampling on Ordinary Least Squares Methods,* Journal of the American Statistical Association, 77, 848-854.

[6] Thomas, D.R., and Rao, J.N.K. (1987), *Small Sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics Under Cluster Sampling,* Journal of the American Statistical Association, 82, 630-636.

[7] Weerakkody, G.J. and Johnson, D.E. (1992), *Estimation of Within Model Parameters in Regression Models With a Nested Error Structure,* Journal of the American Statistical Association, 87, 708-713.

[8] Weerakkody, G.J. and Givaruangswat, S. (1995), *Estimation of the correlation coefficient in the presence of correlated observations from bivariate a normal population,* Comm. Stat. (Theory and Methods) 24(7), 1705-1719.

[9] Wu, C.F. J., Holt, D., and Holms, D.J., (1988) *The Effect of Two-Stage Sampling on the F Statistic,* Journal of the American Statistical Association, 83, 150-159.