Conference on Applied Statistics in Agriculture 1999 - 11th Annual Conference Proceedings

AN IMPROVED ESTIMATOR FOR ASSESSING THE MEASURE OF AGREEMENT WITH A GOLD STANDARD

Brent D. Burch

Ian R. Harris

Roy T. St. Laurent

AN IMPROVED ESTIMATOR FOR ASSESSING THE MEASURE OF AGREEMENT WITH A GOLD STANDARD

Brent D. Burch, Ian R. Harris, and Roy T. St. Laurent

Department of Mathematics and Statistics, Northern Arizona University,
Flagstaff, Arizona 86011, U.S.A.

ABSTRACT

St. Laurent (1998, *Biometrics* **54**, 537–545) developed a measure of agreement for method comparison studies in which an approximate method of measurement is compared to a gold standard method of measurement. The measure of agreement proposed was shown to be related to a population intraclass correlation coefficient. This paper develops a family of estimators for the measure of agreement based on pivotal quantities. A blend of two particular members of the family is suggested as an estimator itself. In general, this estimator outperforms the maximum likelihood estimator in terms of bias and mean-squared error.

1 Introduction

St. Laurent (1998) proposed an estimator that can be used in method comparison studies where the aim is to assess the degree of agreement between a precise standard of measurement (the gold standard) and an approximate measurement. The gold standard method is often expensive and time consuming to apply, and may be invasive or destructive of the object being measured. The approximate method, on the other hand, is often inexpensive, quicker, and noninvasive. It is assumed that the gold standard and approximate measurements are on the same scale and no calibration is desired.

For example, Prigent et al. (1991) investigate induced myocardial infarcts in twelve dogs. For each dog they examined the percentage of heart muscle affected by the infarctions. An approximate measure is an image analysis from single photon emission computed tomography (SPECT). The gold standard of measurement is the percentage determined by pathologic examination. In another application, the loin eye area of beef cattle determined by an imaging procedure prior to slaughter may be compared to the loin eye area obtained by examining the carcass. In competitions involving judging, it may be of interest to compare the opinion of a novice judge (the approximate method) to that of an expert judge (the gold standard method). Additional examples of method comparison studies can be found in St. Laurent (1998).

St. Laurent (1998) suggests the use of a modified random effects model, given by

$$X_i = G_i + \epsilon_i, \tag{1}$$

to assess the agreement between an approximate method of measurement with a gold standard. X_i is the approximate measurement on the i^{th} unit; G_i, the gold standard measurement on the i^{th} unit is a random variable with mean μ and variance σ_G^2; and ϵ_i is a measurement error associated with the approximate measurement, independent of G_i, with mean 0 and variance σ^2. The model given in (1) assumes that the approximate and gold standard methods have the same mean and that the approximate measurements have larger variation $(\sigma_G^2 + \sigma^2)$ than the gold standard measurements (σ_G^2).

Using this model, the correlation between X_i and G_i is $\rho = \sigma_G/(\sigma_G^2 + \sigma^2)^{1/2}$ so ρ may be used to measure the agreement between the approximate and gold standard. ρ^2 is the proportion of variability in X due to G and is identical in form to an intraclass correlation coefficient. Note that $0 \leq \rho \leq 1$ and large values of ρ are of primary interest as they indicate the approximate and gold standard measurements are in relative agreement. In these situations, the approximate method may be deemed adequate and the often invasive or expensive gold standard method avoided.

St. Laurent (1998) investigates several estimators of ρ, and suggests the use of

$$r_g = \left(\frac{Y}{Y+1}\right)^{1/2}, \tag{2}$$

where $Y = S_{GG}/S_{DD}$, $S_{DD} = \sum(X_i - G_i)^2 = \sum \epsilon_i^2$, and $S_{GG} = \sum(G_i - \overline{G})^2$. An interesting feature of this problem is that the ϵ_i are observed and hence S_{DD} is calculable. If the ϵ_i and G_i are normally distributed, then r_g is the maximum likelihood estimator (MLE) of ρ.

In this paper the authors develop a family of estimators for ρ, which includes r_g. It will be shown that for various values of ρ certain members of this family are preferable to r_g in terms of bias and mean-squared error. Furthermore, a particular blend of two of these estimators has very little bias, with mean-squared error similar to that of r_g. When ρ is large, the blended estimator is superior to r_g in terms of bias and mean-squared error.

2 A Family of Estimators

Burch and Harris (1998) suggest a method of deriving estimators that is equivalent to shrinking confidence intervals to a point by reducing the confidence coverage to zero. A particular case of this method is to equate a pivotal quantity for the parameter of interest to a value of the pivoting distribution. It is this method that is used in the present paper. For general discussion of the procedure see Burch and Harris (1998).

Assuming that the ϵ_i are from a scale family of distributions with scale parameter σ, and that the G_i are from a location-scale family with parameters μ and σ_G, then $nY(\rho^{-2} - 1)/(n-1)$ is a pivotal quantity. That is, the distribution of $nY(\rho^{-2} - 1)/(n-1)$ does not depend on the value of the parameter ρ. The method of estimation requires that one solve the pivoting equation

$$nY(\widehat{\rho}_F{}^{-2} - 1)/(n-1) = F, \tag{3}$$

where F is a value from the support of the distribution of the pivotal quantity. Solving (3) for $\widehat{\rho}_F$ leads to the family of estimators indexed by F

$$\widehat{\rho}_F = \left(\frac{nY}{nY + (n-1)F} \right)^{1/2}. \tag{4}$$

This family contains r_g by selecting $F = n/(n-1)$. Another estimator of ρ, the ANOVA estimator proposed by St. Laurent (1998), corresponds to $F = 1$.

3 Some Properties of the Estimators

Properties of members of this family of estimators are presented in the following discussion. The values of $\widehat{\rho}_F$ are confined to the interval $[0, 1]$. Note that the estimator $\widehat{\rho}_F$ is not truncated as is often the case when estimating functions of variance components. As F increases, $\widehat{\rho}_F$ decreases and as F decreases, $\widehat{\rho}_F$ increases. For each value of F, as ρ goes to zero, $\widehat{\rho}_F$ goes to zero, and as ρ approaches 1, $\widehat{\rho}_F$ approaches 1. Thus for any value of F, the bias and mean-squared error of $\widehat{\rho}_F$ go to zero as ρ approaches the endpoints of the parameter space.

One immediate question is whether there is a member of the family that is better than any other member of the family in minimizing mean-squared error. Harris and Burch (1999) investigate a similar problem involving estimators of the intraclass correlation coefficient in a balanced one-way random effects model. The following result for the gold standard problem parallels results of Harris and Burch (1999).

Result: *Consider the family of estimators of ρ of the form $\widehat{\rho}_F$. For each F in the interval $[F_-, F_+]$, the estimator $\widehat{\rho}_F$ cannot be beaten everywhere in minimizing mean-squared error by any other estimator within the family of estimators of the form $\widehat{\rho}_F$.*

Note that $\widehat{\rho}_F = (1 + FU(\rho^{-2} - 1))^{-1/2}$, where $U = (n-1)/(nY(\rho^{-2} - 1))$. Let $\epsilon = 1 - \rho$, then near $\rho = 1$ one can write $\widehat{\rho}_F = 1 - \epsilon FU + o(\epsilon)$. It follows that $\text{Bias}(\widehat{\rho}_F) = \epsilon(1 - FE(U)) + o(\epsilon)$ and $\text{MSE}(\widehat{\rho}_F) = \epsilon^2[1 - 2FE(U) + F^2E(U^2)] + o(\epsilon^2)$. Note that $\text{MSE}(\widehat{\rho}_F)$ is an increasing function of F when ρ is close to 1 as long as $F > F_- = E(U)/E(U^2)$.

For ρ near 0, $\widehat{\rho}_F = \rho(FU)^{-1/2} + o(\rho^2)$, and hence $\text{Bias}(\widehat{\rho}_F) = \rho(F^{-1/2}E(U^{-1/2}) - 1) + o(\rho^2)$ and $\text{MSE}(\widehat{\rho}_F) = \rho^2(F^{-1}E(U^{-1}) - 2F^{-1/2}E(U^{-1/2}) + 1) + o(\rho^3)$. In this case, the $\text{MSE}(\widehat{\rho}_F)$ decreases for F-values less than $F_+ = \left(E(U^{-1})/E(U^{-1/2})\right)^2$, whereupon the mean-squared error starts to increase. Combining the results for ρ near the ends of the parameter space, one may conclude that $\widehat{\rho}_F$ is admissible within the family in terms of mean-squared error if F is contained in $[F_-, F_+]$.

The result above indicates that within this family of estimators of ρ there exists a collection of admissible estimators of ρ. From the collection of admissible estimators one may identify estimators that have desirable attributes. For instance, there are two members of the family that perform well in terms of bias in either the lower or upper region of the parameter space: $\widehat{\rho}_{F_0}$, where $F_0 = E(U^{-1/2})^2$, has negligible bias when ρ is close to 0; and

$\widehat{\rho}_{F_1}$, where $F_1 = 1/E(U)$, has negligible bias when ρ is close to 1. Using Jensen's inequality it may be shown that $F_- \leq F_1 \leq F_0 \leq F_+$ and hence both $\widehat{\rho}_{F_0}$ and $\widehat{\rho}_{F_1}$ are admissible within the family.

4 A Blended Estimator

Srivastava (1993) suggests the use of a composite or blended estimator for the intraclass correlation coefficient in unbalanced designs. The procedure is to blend together two estimators, one of which performs well in one section of the parameter space, and one of which performs well in another section. This method can be adapted to the gold standard problem by creating an estimator which is a blend of $\widehat{\rho}_{F_0}$ and $\widehat{\rho}_{F_1}$. One can blend these estimators on the ρ scale or the ρ^2 scale. As the derivation of $\widehat{\rho}_F$ relied on a pivotal quantity that involved ρ^2, blending is performed on the ρ^2 scale.

The blended estimator of ρ, denoted by $\widehat{\rho}_b$, is obtained from

$$\widehat{\rho}_b^2 = (1-\omega)\widehat{\rho}_{F_0}^2 + \omega\widehat{\rho}_{F_1}^2, \tag{5}$$

where ω is a weight parameter ($0 \leq \omega \leq 1$). When ρ^2 is small we would like the estimator to place more weight on $\widehat{\rho}_{F_0}^2$ which implies ω should be small. Similarly, when ρ^2 is large we would like the estimator to place more weight on $\widehat{\rho}_{F_1}^2$ which implies ω should be large. One can see that values of ω and ρ^2 are related as ω tends to copy ρ^2. Following Srivastava (1993), $\widehat{\rho}_b^2$ is substituted for ω in (5) and solving for $\widehat{\rho}_b$ yields

$$\widehat{\rho}_b = \left(\frac{\widehat{\rho}_{F_0}^2}{1+\widehat{\rho}_{F_0}^2-\widehat{\rho}_{F_1}^2}\right)^{1/2}. \tag{6}$$

As outlined in the next section, the performance of this estimator is excellent, particularly in terms of bias.

5 Performance of the Estimators

If ϵ_i and G_i are normally distributed, $U \sim F_{n,n-1}$ in which case the F-values that result in admissible estimators range from $F_- = n(n-5)/((n-1)(n+2))$ to $F_+ = n(n-1)/(n-2)^2$. In addition, $F_0 = n/(n-1)$, $F_1 = (n-3)/(n-1)$, and the maximum likelihood estimator of ρ under the normal assumption, r_g, is equal to $\widehat{\rho}_{F_0}$. If $\widehat{\rho}_{b_N}$ is defined to be the blended estimator under the normality assumptions, then

$$\widehat{\rho}_{b_N} = \left(\frac{nY^2+(n-3)Y}{nY^2+2(n-3)Y+(n-3)}\right)^{1/2}. \tag{7}$$

As previously mentioned, estimating ρ when it is large is of primary importance, hence estimators of the form (4) corresponding to small values of F are preferable. In this

regard, $\widehat{\rho}_{F_-}$, $\widehat{\rho}_{F_1}$, as well as the blended estimator $\widehat{\rho}_{b_N}$ are considered. These estimators are compared to the MLE, namely $r_g = \widehat{\rho}_{F_0}$, in terms of relative mean-squared error and relative absolute bias.

As ρ approaches 1 (or 0), it is possible to analytically find the relative mean-squared error and relative absolute bias by considering the expansion of these quantities in terms of ρ. The measures of performance of $\widehat{\rho}_{F_-}$, $\widehat{\rho}_{F_1}$, and $\widehat{\rho}_{b_N}$ relative to r_g are displayed in Table 1. When $\rho = 1$, $\widehat{\rho}_{b_N}$ and $\widehat{\rho}_{F_1}$ exhibit superior performance in terms of mean-squared error *and* bias. Furthermore, $\widehat{\rho}_{b_N}$ has the best performance in terms of mean-squared error and bias when $\rho = 0$. In other words, the blended estimator outperforms the competing estimators at *both* ends of the parameter space.

To investigate the behavior of the estimators for intermediate values of ρ, one has to resort to numerical calculations for specific cases. Figure 1 presents a comparison of $\widehat{\rho}_{b_N}$ and r_g in terms of bias. Results for $n = 10$, 20, and 50 are displayed. Figure 2 presents a comparison $\widehat{\rho}_{b_N}$ and r_g in terms of relative mean-squared error, defined as $\mathrm{MSE}(\widehat{\rho}_{b_N})/\mathrm{MSE}(r_g)$. As in Figure 1, the results in Figure 2 are displayed for $n = 10$, 20, and 50.

For a given sample size, the bias of $\widehat{\rho}_{b_N}$ is close to zero whereas r_g exhibits a conspicuously negative bias. In fact, Figure 1 indicates that for samples sizes greater than 50 the bias of $\widehat{\rho}_{b_N}$ is negligible across the entire parameter space. For large values of ρ, $\widehat{\rho}_{b_N}$ outperforms r_g in terms of bias and mean-squared error. In other words, r_g understates the degree of agreement when the agreement is very good, a problem corrected by the use of the blended estimator. Figures 1 and 2 suggest there is much to gain when using the estimator $\widehat{\rho}_{b_N}$ if ρ is large and little to lose if ρ is small.

Although comparisons between $\widehat{\rho}_{b_N}$ and r_g were made using normal distribution assumptions, the properties of the estimators $\widehat{\rho}_{b_N}$ and r_g when in reality ϵ_i and G_i are not normally distributed are also of interest. The authors have performed simulation studies with non-normally distributed random variables which result in conclusions similar to those reached above.

It can be shown that the asymptotic standard error of $\widehat{\rho}_{b_N}$ is given by

$$A.s.e.(\widehat{\rho}_{b_N}) = \left(\frac{(n-3)(2n-3)\rho^2(1-\rho^2)^2(n-3\rho^4)^2}{2n(n-5)(n-6\rho^2+3\rho^4)^3} \right)^{1/2} \tag{8}$$

which for large n is approximately $\rho(1-\rho^2)/n^{1/2}$. A 95% asymptotic confidence interval for ρ may be constructed as $\widehat{\rho}_{b_N} \pm 1.96 A.s.e.(\widehat{\rho}_{b_N})$. However, the endpoints of this interval may lie outside the range [0, 1]. For this reason, we recommend that large sample confidence intervals for ρ be constructed by inverting confidence intervals for $\ln(1/\rho^2 - 1)$ based on $\ln(1/\widehat{\rho}^2_{b_N} - 1)$, a monotonic transformation of $\widehat{\rho}_{b_N}$. The asymptotic standard error of $\ln(1/\widehat{\rho}^2_{b_N} - 1)$ is

$$A.s.e. \left(\ln \left[\frac{1}{\widehat{\rho}^2_{b_N}} - 1 \right] \right) = \left(\frac{2(2n-3)(n-3\rho^4)^2}{n(n-5)(n-3\rho^2)^2} \right)^{1/2} \tag{9}$$

which for large n is approximately $2/n^{1/2}$. Note that for large n, (9) is less dependent upon the value of ρ than (8).

6 Example

The results of this paper are applied to the myocardial infarction data from Prigent et al. (1991). Under consideration is the percentage of heart muscle affected by induced myocardial infarctions for $n = 12$ dogs as measured by pathology (the gold standard method) and SPECT (the approximate method). Table 2 displays the data which were reconstructed from Figure 2 in Prigent et al. (1991). From the reconstructed data one can show that $Y = 2.88$ and it follows that $\widehat{\rho}_{F_0} = r_g = 0.86$. In addition, $\widehat{\rho}_{F_1} = 0.89$, $\widehat{\rho}_{F\_} = 0.92$, and $\widehat{\rho}_{b_N} = 0.88$. Although these estimates may not seem to be very different from one another, most likely r_g understates the agreement of the SPECT measure of percentage heart muscle affected by myocardial infarction with the percentage as determined by pathologic examination, and $\widehat{\rho}_{F\_}$ overstates the agreement. A 95% asymptotic confidence interval for ρ using (9) is (0.67, 0.97). This interval is presented for illustrative purposes only as the actual coverage probability may not be close to 0.95 for $n = 12$.

7 Discussion and Conclusions

In method comparison studies, estimating the degree of agreement between different measurement techniques is often of primary importance. In this paper the authors consider a scenario in which an approximate measurement is compared to a precise standard of measurement (the gold standard). A family of admissible estimators for the measure of agreement which includes the familiar normal distribution based MLE is considered. An estimator which is a particular blend of two members of the family has excellent bias properties across the entire parameter space. Furthermore, the mean-squared error of the blended estimator is substantially smaller than the mean-squared error of r_g for large values of ρ. The authors favor the blended estimator even in those cases where the underlying distributions are not normal.

Acknowledgements

The research of Brent D. Burch was supported by the Northern Arizona University Organized Research Program. The authors thank the anonymous referee for valuable suggestions that improved the quality of this paper.

References

Burch, B. D. and Harris, I. R. (1998). Using confidence intervals to obtain a family of estimators of the intraclass correlation coefficient (or heritability). *Proceedings of the Kansas State University Conference on Applied Statistics in Agriculture*, 109–122.

Harris, I. R. and Burch, B. D. (1999). Pivotal estimation with applications for the intraclass correlation coefficient in the balanced one-way random effects model. *Journal of Statistical Planning and Inference* **83**, 257–276.

Prigent, F. M., Maddahi, J., Van Train, K. F. and Berman, D. S. (1991). Comparison of thallium-201 SPECT and planar imaging methods for quantification of experimental myocardial infract size. *American Heart Journal* **122**, 972–979.

St. Laurent, R. T. (1998). Evaluating agreement with a gold standard in method comparison studies. *Biometrics* **54**, 537–545.

Srivastava, M. S. (1993). Estimation of the intraclass correlation coefficient. *Annals of Human Genetics* **57**, 159–165.

Table 1: Relative MSE and Absolute Bias of Estimators as Compared to the MLE

	$\rho = 1$	$\rho = 1$	$\rho = 0$	$\rho = 0$
$\widehat{\rho}$	$\frac{\text{MSE}(\widehat{\rho})}{\text{MSE}(r_g)}$	$\frac{\lvert\text{Bias}(\widehat{\rho})\rvert}{\lvert\text{Bias}(r_g)\rvert}$	$\frac{\text{MSE}(\widehat{\rho})}{\text{MSE}(r_g)}$	$\frac{\lvert\text{Bias}(\widehat{\rho})\rvert}{\lvert\text{Bias}(r_g)\rvert}$
$\widehat{\rho}_{F-}$	$\frac{2(n-5)(2n-3)}{(n+2)(4n+15)}$	$\frac{2(2n-3)}{3(n+2)}$	$\frac{(n-1)(n+2)}{n-5} - 2\frac{(n+2)^{1/2}(n-2)}{(n-5)^{1/2}} + n - 2$	∞
$\widehat{\rho}_{F_1}$	$\frac{2(n-3)(2n-3)}{n(4n+15)}$	0	$\frac{n(n-1)}{n-3} - 2\frac{n^{1/2}(n-2)}{(n-3)^{1/2}} + n - 2$	∞
$\widehat{\rho}_{b_N}$	$\frac{2(n-3)(2n-3)}{n(4n+15)}$	0	1	$\frac{3}{n-3}$

Table 2: Percentage of Heart Muscle Affected by Induced Myocardial Infarct as Measured by Pathology (Gold Standard Method) and SPECT (Approximate Method) for 12 Dogs

i	G_i	X_i
1	9.1	5.1
2	7.7	7.1
3	21.4	13.1
4	18.5	16.9
5	28.7	34.4
6	12.9	13.0
7	13.2	17.1
8	20.3	19.4
9	26.2	23.2
10	30.0	24.2
11	31.2	23.8
12	24.0	28.3

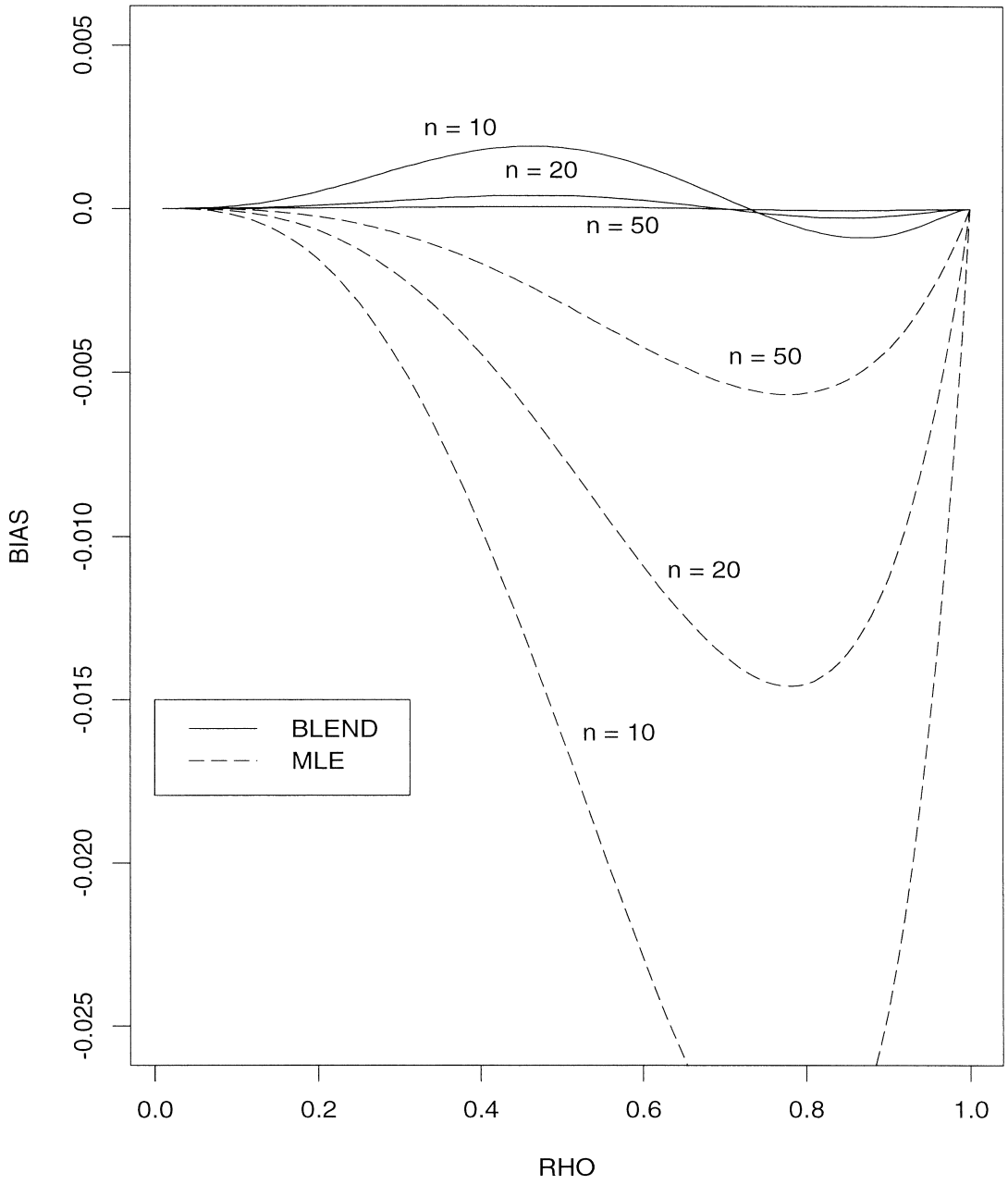

Figure 1: Comparison of $\widehat{\rho}_{b_N}$ and r_g in terms of Bias

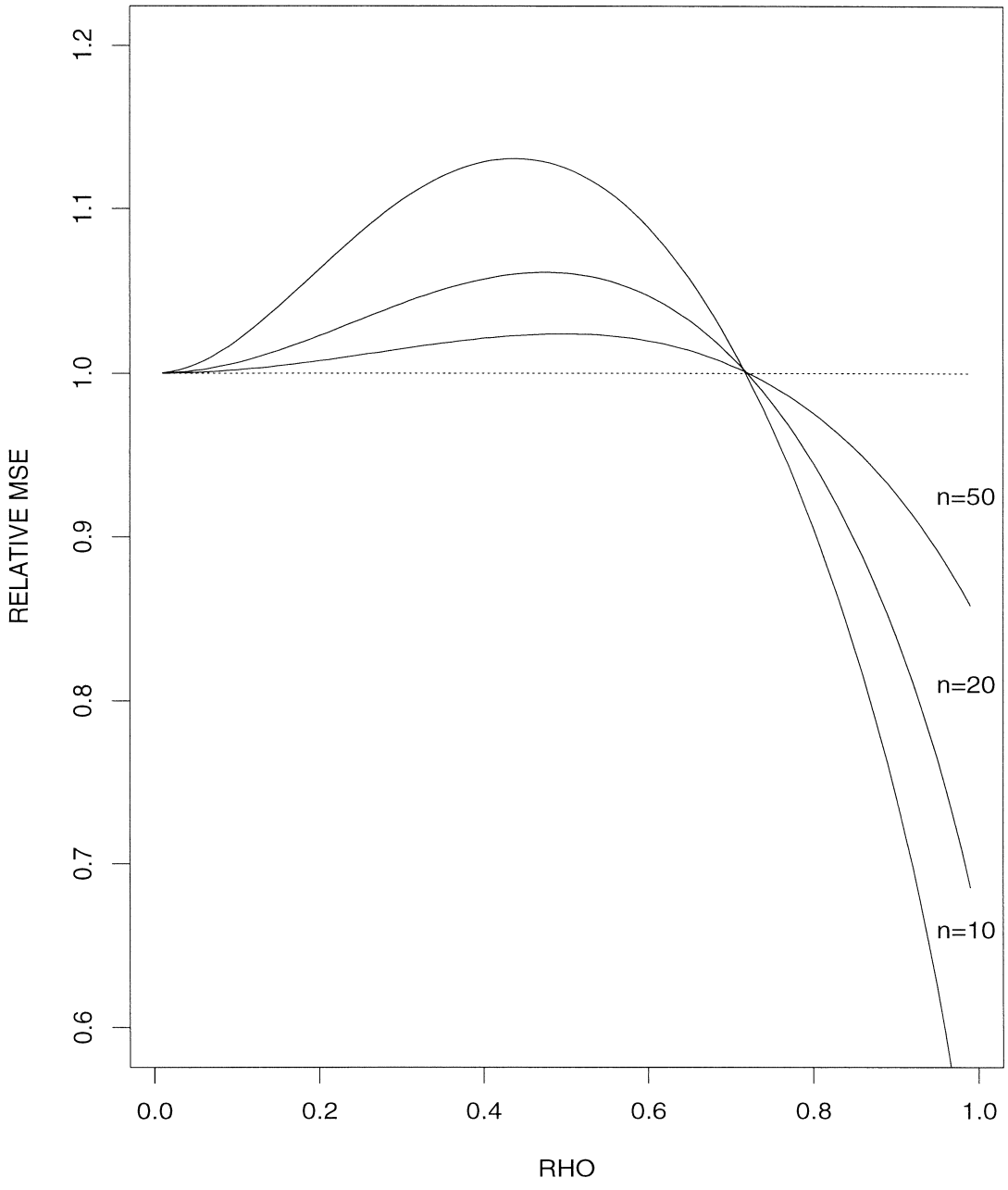

Figure 2: Comparison of $\widehat{\rho}_{b_N}$ and r_g in terms of Relative Mean-Squared Error, $\mathrm{MSE}(\widehat{\rho}_{b_N})/\mathrm{MSE}(r_g)$