April 1970

Misidentification in Estimating the Paternal Sib Correlation

L. Dale Van Vleck
University of Nebraska-Lincoln, dvan-vleck1@unl.edu

Misidentification in Estimating the Paternal Sib Correlation

L. D. VAN VLECK
Department of Animal Science
Cornell University, Ithaca, New York 14850

Abstract

Misidentification of sires of cows in paternal half-sib analyses of variance biases the estimate of heritability downward. Four models of misidentification yield approximately the same reductions in estimated heritability. The reduced estimates are approximately p^2 of the actual heritability where p is the fraction of cows whose sire is correctly identified. The reduction in heritability estimates from daughter-dam regression would be to a fraction, p, of the actual heritability where p is the fraction of dams correctly identified. Empirical paternal sib analyses of New York Dairy Herd Improvement Holstein records suggest that misidentification is more common among grade cows with registered dams and grade cows having grade dams which are not identified than among registered cows with registered dams.

Introduction

There is some doubt whether all dairy cows on Dairy Herd Improvement (DHI) testing programs are correctly identified by sire and dam. A recent survey by Everett and Wadell (2) suggests that one-third to one-half of DHI cows in New York have no sire or dam identification or are misidentified. This is a severe problem which reduces the amount of possible genetic progress by selection because of the reduction in the average correlation between true genetic value and estimated genetic value.

Misidentification of sire and dam among the remaining cows remains a possibility. Such misidentification may account for part of the discrepancy between estimates of heritability from the regression of daughter record on dam record and from intrasire correlation (1, 5). The effect of such misidentification on the estimates of heritability from the intrasire correlation is the subject of this paper.

Received for publication April 27, 1970.

Theoretical Results

The following model described a record of the j^{th} daughter of the i^{th} sire:

$$y_{ij} = \mu + s_i + e_{ij} \text{ where}$$

μ is a constant,
s_i is a random effect common to records of daughters of the i^{th} sire, and
e_{ij} is a random effect common to a record of the j^{th} daughter of the i^{th} sire.

The s_i are uncorrelated with mean, zero, and common variance, σ^2_s. The e_{ij} are uncorrelated with mean, zero, and common variance, σ^2_e. The s_i and e_{ij} are uncorrelated for all values of i and j.

In the problem of misidentification of the sire of a cow, several modifications of the model were used.

1) Within a supposed sire group where each misidentified daughter is from a different sire which has no other daughters in the population, the model for the sum of the i^{th} sire group with n_i "daughters" is

$$y_{i.} = n_i \mu + n_{il} s_i + \sum_{k=n_{il}+1}^{n_i} s_{ik} + e_{i.}$$

where n_{il} of the cows are daughters of the i^{th} sire and $n_i - n_{il}$ of the cows are daughters of that many other sires which have no other daughters in any other sire group. The usual dot notation indicates summation over that subscript. The expectations of the usual three quadratics (total sum of squares, sire sum of squares, and correction factor) are in Table 1.

2) Within a supposed sire group where all misidentified daughters are from the same sire (the ik^{th} sire) which has no other daughters in the population, the model for the sum of the i^{th} sire group becomes

$$y_{i.} = n_i \mu + n_{il} s_i + (n_i - n_{il})\, s_{ik} + e_{i.}$$

The expectations of the sums of squares are also in Table 1.

3) The same as 1) except that the total number of daughters for each sire is actually n_i

TABLE 1. Expectations of sums of squares for misidentified data.[a]

Pattern of misidentified records in a group	$E[\text{Sire SS}] = E[\sum_i y^2_{i.}/n_i]$	$E[C F] = E[y^2_{..}/n_.]$
1) All by different sires, no repeats in other groups	$n_.\mu^2 + [\sum_i \frac{n_{ii}(n_{ii}-1)}{n_i} + S]\sigma^2_s + S\sigma^2_e$	$n_.\mu^2 + 1/n_. [\sum_i n^2_{ii} + \sum_i (n_i - n_{ii})]\sigma^2_s + \sigma^2_e$
2) All by same sire in a group, no repeats in other groups	$n_.\mu^2 + [n_. + 2\sum_i \frac{n_{ii}(n_{ii}-n_i)}{n_i}]\sigma^2_s + S\sigma^2_e$	$n_.\mu^2 + 1/n_. [\sum_i n^2_i + 2\sum_i n_{ii}(n_{ii} - n_i)]\sigma^2_s + \sigma^2_e$
3) Same as 1) except misidentified really belong in other groups	$n_.\mu^2 + [\sum_i \frac{n_{ii}(n_{ii}-1)}{n_i} + S]\sigma^2_s + S\sigma^2_e$	$n_.\mu^2 + 1/n_. \sum_i n^2_i \sigma^2_s + \sigma^2_e$
4) Same as 2) except misidentified really belong in another group	$n_.\mu^2 + [n_. + 2\sum_i \frac{n_{ii}(n_{ii}-n_i)}{n_i}]\sigma^2_s + \sigma^2_e$	$n_.\mu^2 + 1/n_. \sum_i n^2_i \sigma^2_s + \sigma^2_e$

[a] Expectation of total sum of squares $= E[\sum_i \sum_j y^2_{ij}]$ is $n_.(\mu^2 + \sigma^2_s + \sigma^2_e)$ for all four patterns.

of which only n_{ii} are correctly identified, i.e., the other $n_i - n_{ii}$ are included singly in other groups. The expectations are in Table 1 and differ from 1) only in that the expectation of the correction factor is the usual expectation for all records correctly identified.

4) The same as 2) except that the total number of daughters for each sire is actually n_i of which only n_{ii} are correctly identified, i.e., the other $n_i - n_{ii}$ are included in another group, always making up the total number of misidentified cows in that group. The expectations are shown in Table 1 and only differ from 2) in that the expectation of the correction factor is the usual expectation for all records correctly identified.

Actual data do not allow detection of the correctly identified daughters, n_{ii}, and the misidentified cows, $n_i - n_{ii}$, in each group. Therefore, to compare the resulting biases in estimating σ^2_s and σ^2_e or rather the intraclass correlation, $t = \sigma^2_s/(\sigma^2_s + \sigma^2_e)$, if the misidentification is one of the four situations described above, the model was simplified so that each sire group contained the same number of cows, i.e., $n_i = n$ for all i. The fraction identified correctly was p for all sires so that $n_{ii} = pn$ for all i. These conditions lead to the expectations (under the conditions for 1, 2, 3, and 4) of σ^2_e and σ^2_s estimated in the usual way from the three sums of squares which are in Table 2. Expectations of estimates are indicated by $E(\hat{\sigma}^2_s)$ and $E(\hat{\sigma}^2_e)$.

The usual estimates are biased if data are misidentified by any of the four conditions. The expectations of the estimates of σ^2_e all contain a portion of σ^2_s, whereas the expectations of the estimates of σ^2_s do not contain a portion of σ^2_e. The coefficients of σ^2_s in the expectations of $\hat{\sigma}^2_s$ always include p and n for all four conditions but contain the total number of cows or equivalently the total number of sires only in Cases 3 and 4. In all cases when $p = 1$ (no misidentification) the coefficients reduce to the usual—$E(\hat{\sigma}^2_s) = \sigma^2_s$ and $E(\hat{\sigma}^2_e) = \sigma^2_e$.

The expected bias in estimating heritability or equivalently the intraclass correlation, $t = \sigma^2_s/(\sigma^2_s + \sigma^2_e)$, can be compared for the balanced case from the foregoing equations.

The bias for the four cases which probably encompass most situations was computed for all combinations of $t = .01, .02, \ldots, .20$, $n = 2, 3, \ldots, 100$, and $p = .05, .10, \ldots, .95$ to examine any pattern in the biases. For the last two cases the number of groups, S, was 500. Table 3 is an extract of those results for $t = .04$ and $.08$; $n = 5, 10, 50$, and 100; and $p =$

TABLE 2. Expectations of usual estimates of sire and residual components of variance from misidentified data assuming equal numbers per group and equal proportions of correctly identified records, p.

Pattern of misidentification	$E(\hat{\sigma}^2_s)$	$E(\hat{\sigma}^2_e)$
1)	$[\frac{z-1}{n-1}]\,\sigma^2_s$	$[\frac{n-z}{n-1}]\,\sigma^2_s + \sigma^2_e$
2)	$[x - \frac{1-x}{n-1}]\,\sigma^2_s$	$\frac{n(1-x)}{n-1}\,\sigma^2_s + \sigma^2_e$
3)	$[\frac{Sz-n}{n(S-1)} - \frac{n-z}{n(n-1)}]\,\sigma^2_s$	$[\frac{n-z}{n-1}]\,\sigma^2_s + \sigma^2_e$
4)	$[\frac{Sx-1}{S-1} - \frac{1-x}{n-1}]\,\sigma^2_s$	$[\frac{n(1-x)}{n-1}]\,\sigma^2_s + \sigma^2_e$

[a] $z = p^2n - p + 1$.
[b] $x = 1 + 2p(p-1)$.
[c] n = number of records per sire group.
[d] S = number of sire groups.

.80, .85, .90, and .95. The expectations are remarkably similar for all four conditions for different sizes of $n \geqq 5$, and appear proportional for all true values of t. The bias due to changing p is mostly proportional to p^2. For example, the coefficients of σ^2_s in $E(\hat{\sigma}^2_s)$ and $E(\hat{\sigma}^2_e)$ add to unity for Patterns 1 and 2 so that the expectations by parts of t are for Pattern 1, $t(p^2n - p)/(n-1)$, and for Pattern 2, $tn(1 + 2p^2 - 2p)/(n-1)$.

Misidentification of daughters and dams may produce a similar bias in estimating heritability

TABLE 3. Expected values of the intraclass correlation for different fractions of misidentification (1 − p), number of animals per group (n), four patterns of misidentification (see Table 1), and true intraclass correlation (t).

		t = .04 Pattern				t = .08 Pattern			
p	n	1	2	3	4	1	2	3	4
.80	5	.024	.024	.024	.024	.048	.048	.048	.048
	10	.025	.026	.025	.026	.050	.052	.050	.052
	20	.025	.027	.025	.027	.051	.053	.050	.053
	100	.026	.027	.026	.027	.051	.054	.051	.054
.85	5	.028	.027	.028	.027	.055	.055	.055	.054
	10	.028	.029	.028	.029	.057	.057	.057	.057
	20	.029	.029	.029	.029	.058	.059	.058	.059
	100	.029	.030	.029	.030	.058	.059	.058	.059
.90	5	.031	.031	.031	.031	.063	.062	.063	.062
	10	.032	.032	.032	.032	.064	.064	.064	.064
	20	.032	.032	.032	.032	.064	.065	.064	.065
	100	.032	.033	.032	.033	.065	.065	.065	.065
.95	5	.036	.035	.036	.035	.071	.070	.071	.070
	10	.036	.036	.036	.036	.072	.072	.072	.072
	20	.036	.036	.036	.036	.072	.072	.072	.072
	100	.036	.036	.036	.036	.072	.072	.072	.072

TABLE 4. Number of cows (N) and sire groups (S) represented in various cow and dam identification groups and years of first freshening.

Cow: Dam:	Registered Registered		Grade Registered		Grade Identified grade		Grade No identification	
Year of freshening	N	S	N	S	N	S	N	S
Official records								
1964	18,162	366	910	175	9,278	313	1,565	225
1965	18,608	381	793	171	9,383	322	1,551	224
1966	18,895	371	870	174	10,153	325	1,669	206
1967	20,252	430	876	161	10,042	363	1,679	222
1968	20,678	515	823	167	10,831	444	1,453	232
Average	19,319	413	854	170	9,937	353	1,583	222
Unofficial records								
1964	1,708	174	114	42	1,999	166	618	124
1965	2,206	191	164	43	2,328	183	773	118
1966	2,289	182	173	55	2,348	156	759	86
1967	2,550	198	161	50	2,503	161	675	85
1968	2,465	207	134	44	2,578	188	655	97
Average	2,244	190	149	47	2,351	171	696	102

from daughter-dam regression or correlation. Such bias would be directly related to p, the fraction of misidentified pairs of data. The expectation of the covariance between daughters and dams would be p Cov(daughter, dam) if only a fraction 1 — p of the pairs are really not related.

If, for example, true heritability were .32 and .05 of cows were wrongly identified as to sire and .05 as to dam, expectations of the estimates of heritability are .288 from a paternal sib analysis and .304 from a daughter-dam regression. But if the fraction misidentified were .10, the corresponding expectations are .260 and .288. However, there would appear a reasonable probability that misidentification of sires would be more frequent than misidentification of dams because of time as well as additional chances for transcription errors.

Obviously, the fraction of misidentified sires and dams cannot be determined from production records. Only blood typing or similar procedures could establish lower limits of misidentification. Published reports of such studies are scarce although rumors are more numerous. Johansson and Rendel (4) report limited studies in Sweden, Norway, and Denmark where the stated parentage was incorrect in about 4% of the animals.

Empirical Results

Some pattern may, however, emerge from analyses of records of cows classified according to whether they are registered or nonregistered and whether the dams are registered, nonregistered but with identification, or nonregistered with no identification. Such analyses were done for first lactation records of artificially sired Holstein cows freshening in 1964 through 1968. The records were further classified as official (DHIA, DHIR) or unofficial (Owner Sampler). The numbers of sires and numbers of daughters included in each analysis among and within sire groups are shown in Table 4. All records were adjusted to a mature equivalent age, 2 times milked per day, and 305-day lactation length, and were expressed as deviations from adjusted herdmate averages (3) to minimize the effects of herds and year-seasons. Test deviations were differences in the lactation fat to lactation milk ratio of the cow and the ratio of her herdmate fat average to her herdmate milk average. The heritability estimates from four times the intrasire correlation are in Table 5.

The data from registered cows with registered dams would be expected to be most correctly identified since the most care should be taken

TABLE 5. Estimates of heritability for various cow and dam identification groups and years of first freshening.

	Official records				Unofficial records			
Cow:	Registered	Grade	Grade	Grade	Registered	Grade	Grade	Grade
Dam:	Registered	Registered	Ident. grade	No ident. grade	Registered	Registered	Ident. grade	No ident. grade
Year				Milk deviations				
1964	.292	.245	.220	.257	.161	−.109	.258	.046
1965	.262	.104	.238	.178	.215	−.088	.271	.358
1966	.276	.094	.207	.190	.335	.231	.126	.170
1967	.279	−.008	.181	.069	.236	.154	.271	−.018
1968	.256	.218	.183	.129	.279	.869	.184	.221
Average	.273	.131	.206	.165	.245	.211	.222	.155
SE[a]	.010	.051	.012	.033	.025	.151	.024	.051
				Fat deviations				
1964	.267	.249	.178	.200	.117	−.509	.234	.016
1965	.225	.207	.205	.251	.165	.106	.217	.325
1966	.234	.136	.192	.193	.253	.301	.138	.069
1967	.231	−.002	.192	.136	.173	−.133	.309	−.001
1968	.246	.265	.169	.149	.265	.891	.236	.194
Average	.241	.171	.187	.186	.195	.131	.227	.121
SE[a]	.010	.051	.012	.033	.025	.151	.024	.051
				Test deviations				
1964	.437	.406	.386	.374	.389	−.117	.484	.580
1965	.432	.082	.429	.327	.641	.226	.383	.600
1966	.471	.353	.452	.196	.460	.803	.506	.798
1967	.432	.131	.482	.240	.512	−.379	.546	.604
1968	.417	.256	.403	.265	.376	.747	.484	.380
Average	.438	.246	.430	.280	.476	.256	.481	.592
SE[b]	.014	.055	.016	.037	.031	.155	.030	.057

[a] Approximate standard error of average heritability estimate assuming heritability equal to .24.
[b] Assuming heritability equal to .40.

in identification. The results generally agree with that hypothesis. The highest estimates of heritability of milk deviations were both for official and unofficial records of cow and dam registered. The same pattern holds for official records for fat yield and milk fat percentage although the average heritability for test for nonregistered cow, nonregistered, identified dam was nearly as high as for registered cows having registered dams. The pattern is not so clear for unofficial records. The estimates for registered cows, registered dams and nonregistered cows, nonregistered, identified dams were nearly the same for both fat yield and fat percentage. The estimate for fat percentage for nonregistered cows with nonidentified dams was especially high but based on an average of only 700 cows per year. The lowest estimates were generally associated with nonregistered cows with registered dams. An explanation for the low estimates is that the cows were not registered because the owner had reason to doubt the identification of either the sire or dam—more probably the sire.

Conclusions

Misidentification of the sire can lead to substantial underestimation of heritability from the intrasire correlation. The reduced estimate appears proportional to the square of the frac-

tion of cows whose sire is correctly identified. This reduction probably is greater than the reduction in the daughter on dam regression caused by the dam being misidentified. The difference in reduction of heritability estimates by misidentification of sires and dams may explain at least part of the difference in heritability estimates from daughter on dam regression and paternal half-sib correlation (1, 5).

References

(1) Bradford, G. E., and L. D. Van Vleck. 1964. Heritability in relation to selection differential in cattle. Genetics, 49: 819.

(2) Everett, R. W., and L. H. Wadell. 1969. Better identification needed. Dairy Herd Management, 6: 41.

(3) Heidhues, T., L. D. Van Vleck, and C. R. Henderson. 1961. Actual and expected accuracy of sire proofs under the New York system of sampling bulls. Z. Tierzücht. Züchtungsbiol., 75: 323.

(4) Johansson, I., and J. Rendel. 1968. Genetics and Animal Breeding. W. H. Freeman and Co., San Francisco. p. 204.

(5) Van Vleck, L. D., and G. E. Bradford. 1965. Comparison of heritability estimates from daughter-dam regression and paternal half-sib correlation. J. Dairy Sci., 48: 1372.