Theor Appl Genet (1987) 74:549–553

THEORETICAL AND APPLIED GENETICS

A note on the effect of observations with missing data on genetic correlation estimates

J. I. Weller and M. Ron

ARO, the Volcani Center, P.O. Box 6, Bet Dagan 50250, Israel

Received November 13, 1986; Accepted May 8, 1987
Communicated by E. J. Eisen

Summary. Various studies have estimated covariance components as half the difference between the variance component of the sum of the variable values, for each observation, and the sum of the corresponding variable variance components. Although the variance components for the separate variables can be computed using all available data, the variance components of the sum can be computed only from those observations with records for both variables. Previous studies have suggested eliminating observations with missing data, because of possible selection bias. The effect of missing data on estimates of covariance components and genetic correlations was tested on sample beef cattle data and simulated data by randomly deleting differing proportions of records of one variable for each pair of variables analyzed. Estimates of genetic correlations computed with observations with missing data eliminated, were more accurate than estimates computed using all available data. Furthermore, when observations with missing data were included, estimates of genetic correlation far outside the parameter space were common. Therefore, this method should be used only if observations with missing data have been eliminated.

Key words: Covariance components – Genetic correlations – Unbalanced data

Introduction

Several methods have been developed to estimate covariance components from unbalanced data (Anderson 1984; Harvey 1970; Henderson 1953, Henderson 1984b; Schaeffer et al. 1978; Searle and Rounsaville 1974; Thompson 1973). Accurate estimation is most difficult when some individuals have both variables recorded, and some have only one variable recorded. Although methods have been presented to deal with this type of data structure (Anderson 1984; Henderson 1984 b; Schaeffer et al. 1978; Thompson 1973), most studies have chosen to eliminate observations with missing data (Henderson 1984 b). If both variables are recorded on all observations, covariance components can be computed relatively easily by equating sums of cross products to the expectations, similar to the estimation of variance components where sums of squares are equated to their expectations (Harvey 1970). Many studies (Agyemang et al. 1985; David et al. 1983; Iloeje et al. 1981; Manfredi et al. 1984; Mondardes and Hayes 1985; Rothschild et al. 1979) have used the following equation, derived by Searle and Rounsaville (1974):

$$\mathrm{Cov}_{xy} = (\mathrm{Var}_{x+y} - \mathrm{Var}_x - \mathrm{Var}_y)/2 \qquad (1)$$

where Cov_{xy} is the estimate of the covariance component of a given factor for variables x and y; Var_x, Var_y, and Var_{x+y} are the variance components of the same factor for x, y, and the sum of x and y; respectively.

Searle and Rounsaville (1974) proved that if both variables are recorded on all individuals, estimates of covariance components derived from equation (1) with Henderson method III (Henderson 1953) will be equal to those derived by equating sums of cross products to their expectations (Harvey 1970). This will be true for other methods commonly used to estimate variance and covariance components from unbalanced data. Anderson (1984) suggested that equation (1) be used only if both variables are recorded on all individuals, since missing data on one variable may be due to selection on the other variable, such as the case of first and second parity milk yield in dairy cattle. However, if the probability of a missing record on one variable is independent of the value recorded for the other variable, intuitively it would seem reasonable to include these observations in the calculation of Var_x and Var_y, but not Var_{x+y}, as they should not bias the estimates, and should have reduced estimation error variances. Alternatively, observations with only one variable can be entirely deleted from the analysis. Henderson (1984 a) has shown that even if estimates of variance components are unbiased, functions of these estimates may be biased. Thus even if all the right-hand terms in equation (1) are estimated by unbiased estimates, but from different data sets, the estimate of the covariance component may be biased. Henderson (1973) concluded that unbiased estimates should be preferred over biased estimates with smaller prediction error variances. However, it would seem that if, under actual conditions, bias is minimal, and the accuracy of the biased estimator is significantly higher, biased estimates may be preferred. An example may be breeding values based on later parity lactation records.

Most reports that used equation (1) do not state whether observations with missing data were deleted, but estimates of genetic correlations outside the parameter space are common (Henderson 1984a). In general, this has been ascribed to relatively small data sets, and the general problems of estimating variance components from unbalanced data.

The goal of this study was to test the effect of missing data for one variable on estimates of covariance components derived from equation (1), both on field and computer simulated data.

Materials and methods

Field data were 1,475 male Israeli Holstein calves slaughtered at one slaughter house during 1984. Four variables of economic importance were studied:

1) Meat, the weight of meat produced from the carcass in kg.
2) Growth rate (slaughter weight – birth weight)/slaughter age. Birth weight was estimated as 35 kg, and slaughter age ranged between 240 and 450 days.
3) Percent meat, 100 * Meat/carcass weight.
4) Meat gain (Meat – (Percent meat *.35))/slaughter age.

Basic statistics of these variables are listed in Table 1 and phenotypic correlations in Table 2. Calves were progeny of 22 sires, from 33 herds. The data set was unbalanced.

Ten data sets of two variables each were generated by Monte Carlo simulation. Each record was simulated using the following formula:

$$Y_{ijkl} = S_{ij} + SC_j + H_{ik} + e_{ijkl} + ec_{jkl} . \quad (2)$$

Where Y_{ijkl} is the record of the 1^{th} progeny of the j^{th} sire from the k^{th} herd for the i^{th} variable; S_{ij} is the effect of the j^{th} sire specific to the i^{th} variable; SC_j is the effect of the j^{th} sire common to both variables; H_{ik} is the effect of the k^{th} herd for the i^{th} variable; e_{ijkl} is the residual of the l^{th} progeny of the j^{th} sire in the k^{th} herd specific to the i^{th} variable; and ec_{jkl} is the residual common to both variables. Twenty sires and twenty herds were generated for each simulation.

Table 1. Basic statistics of the variables studied[a]

Variable	Mean	SD
Meat (kg)	229.00	25.400
Growth rate (kg/day)	1.09	0.118
Percent meat	53.90	2.200
Meat gain (kg/day)	0.59	0.063

[a] Data were from 1,475 calves. Variables are defined in the text

Table 2. Phenotypic correlations between the variables studied[a]

	Growth rate	Percent meat	Meat gain
Meat	0.58	0.22	0.67
Growth rate	–	– 0.22	0.92
Percent meat		–	0.16

[a] Data were from 14,75 calves. Variables are defined in the text

The effects were simulated by random sampling from normal distributions with standard deviations of 177 for S_{ij} and SC_j, 500 for H_{ij}, and 685 for e_{ijkl} and ec_{jkl}. Thus the expectation was that the sire component of variance would be 6.25% of the total variance ($h^2 = 0.25$) and both the genetic and environmental correlations would be 0.5. The number of progeny per sire was computed as 5 times a value sampled from a chi-squared distribution with 10 degrees of freedom rounded to the closest integer. This approximated the unbalanced design of progeny per sire generally found in field data. This procedure resulted in about 1000 progeny per data set, but the number varied slightly among data sets due to random sampling. The following algorithm was applied to the progeny of each sire in order to generate an unbalanced distribution of sires' progeny across herds:

1. The progeny was assigned into one of the twenty herds by random sampling from a uniform distribution over the range of 1 to 20, $n_c = 0.9$.
2. A random number, n, was sampled from a uniform distribution over the range of 0 to 1.0.
3. If $n < n_c$ then the next progeny was assigned to the same herd as the previous progeny, $n_c = n_c - 0.1$, and the procedure was continued from step 2.
4. If $n \geqq n_c$, the procedure was continued from step 1.

Sire and error components of variance were computed by Henderson's method III (Henderson 1953) for the four field data variables and the two simulated variables of each data set. The analysis model was:

$$Y_{jkl} = S_j + H_k + e_{jkl} \quad (3)$$

where Y_{jkl} is the variable value of the l^{th} calf, son of the j^{th} sire from the k^{th} herd; S_j is the random effect of the j^{th} sire, H_k is the fixed effect of the k^{th} herd; and e_{jkl} is the residual associated with each record. The sire covariance components were computed by equation (1) between the two variables of each simulated data set and the pairs of variables: meat and growth rate, meat and meat gain, and percent meat and meat gain. These combinations were chosen because the variances were radically different for the two variables included in each pair.

The field data were analyzed with no missing data and with 1, 10, 20, 33, 50, 67, and 75% of the records of the variable with the smaller variance randomly deleted. Data were deleted only from the variable with the smaller variance, in order to enhance the effect of missing data on the covariance component estimates. By deleting these records, the variance component of the variable with the larger variance could be computed using observations not included in the other variance component estimate. The simulated data was analyzed with no missing data and with 1 and 10% of the records randomly deleted for the second variable by sampling from a uniform distribution. The variance and covariance components were also computed with the variable with missing data divided by 10 and 100 to obtain ratios between the variables similar to those in the field data. Genetic correlations were estimated as;

$$r = \mathrm{Cov}_{xy}/(\mathrm{Var}_x\,\mathrm{Var}_y)^{1/2} \tag{4}$$

where r is the estimated genetic correlation, Cov_{xy} is the sire covariance component estimate for variables x and y, and Var_x and Var_y are the sire variance component estimates for the two variables. Genetic correlations were estimated from variance component estimates derived i) using all available data, in variable units, $r_{(x,y)}$; ii) from the restricted data set (observations with missing data deleted) in variable units, $r\,\mathrm{res}_{(x,y)}$; iii) for the field data in standardized units, $r_{(xs,ys)}$ (variable values divided by each variable's standard deviation); and iv) for the simulated data with the variable with missing data divided by 10, $r_{(x10,y10)}$, and divided by 100, $r_{(x100,y100)}$. It can be readily shown that for the restricted data sets, division of the variables by a constant will not affect the estimates of covariance component and therefore were not computed.

Accuracy of the genetic correlation estimates for the simulated data sets was determined by the estimation error variance computed as follows:

$$\mathrm{EEV}_{mn} = \left[\sum_{p=1}^{10} (r_{mnp} - R_p)^2\right] / 10 \tag{5}$$

where EEV_{mn} is the estimation error variance for the mth method of analysis [$r_{(x,y)}$, $r\,\mathrm{res}_{(x,y)}$, $r_{(x10,y10)}$, and $r_{(x100,y100)}$] with the nth fraction of missing data; r_{mnp} is the estimated genetic correlation for the mth analysis of the pth data set with the nth fraction of missing data; and R_p is the true genetic correlation for the pth data set, computed as the correlation between the true sire effects, $S_{ij} + SC_j$, for the two variables. Since the estimates were deviated from R_p rather than the mean of the estimates, the sum of squares was divided by 10, the number of R_p values generated.

For the field data the true genetic correlation is unknown, and only three pairs of variables were analyzed. Therefore accuracy was estimated by the following approximation of EEV.

$$\mathrm{EEV}'_m = \left[\sum_{q=1}^{3}\sum_{n=2}^{8} (r_{mnq} - r_{m1q})^2\right] / 21 \tag{6}$$

where EEV'_m is an estimate of EEV for the mth method of analysis [$r_{(x,y)}$, $r\,\mathrm{res}_{(x,y)}$, and $r_{(xs,ys)}$]; r_{mnq} is the estimate of r computed with the nth level of missing data for the qth pair of variables, in the mth analysis. r_{m1q} is the estimate of r for the qth variable pair in the mth analysis with no missing data. Equation (6) is based on the assumption that the r_{m1q} estimates derived from the complete data with no missing values are nearly equal to the expectations. As in equation (5), no degrees of freedom were lost and the sum of squares was divided by 21. Similar results were obtained for the within variable-pair genetic correlation means squares estimated by the following equation, and are therefore not presented.

$$\mathrm{MS}_m = \left[\sum_{q=1}^{3}\sum_{n=1}^{8} (r_{mnq} - \bar{r}_{m\cdot q})^2\right] / 21 \tag{7}$$

where $\bar{r}_{m\cdot q}$ is the mean of the eight correlation estimates computed for the qth pair of variables in the mth analysis. Heterogeneity of EEV and EEV′ values were tested by the F statistic.

Results and discussion

Variance components and genetic correlations for the three pairs of field data variables are listed in Table 3. Cov_{xy} for $r\,\mathrm{res}_{(x,y)}$ estimates was computed from variance component estimates derived with missing data deleted for both variables, while $r_{(x,y)}$ estimates were derived using the estimated sire component of variance with no missing data for the first variable. Although the change in the variance component of the first variable is in most cases small in proportion to its own value, the change is quite large in proportion to the magnitude of the variance component of the second variable. All the genetic correlation estimates computed without observations with missing data, $r\,\mathrm{res}_{(x,y)}$, were within the parameter space, despite that up to 75% of the records were deleted, and estimates outside the parameter space have been commonly found (Henderson 1984a). The maximum deviation between $r\,\mathrm{res}_{(x,y)}$ computed on the complete data set, and between those computed on the restricted data sets, was 0.3. Conversely, the genetic correlations obtained when observations with missing data on one variable were used to compute the variance component for the other variable, were nearly all outside the parameter space of -1 to 1, and in most cases by large margins, even if only 1% of the observations had missing data. Only one of the estimates derived from standardized data was outside the parameter space, but in nearly all cases the estimates derived after elimination of observations with missing data were closer to the estimates obtained with the complete data set. EEV′ values were 623.6 for $r_{(x,y)}$; 0.067 for $r_{(xs,ys)}$; and 0.012 for $r\,\mathrm{res}_{(x,y)}$. These estimates differ significantly from each other at $P < 0.05$.

The EEV values for the genetic correlations computed from the simulated data sets are presented in

Table 3. Estimates of variance components and genetic correlations with varying fractions of the records deleted for one variable[a]

Variables	Percent missing	Sire variance components[b]			Genetic correlations[c]		
		x	y	x+y	r $\text{res}_{(x,y)}$	$r_{(x,y)}$	$r_{(xs,ys)}$
Meat × growth rate	0	290,078	6.77	291,516	0.51	0.51	0.51
	1	299,089	6.69	300,587	0.53	3.77	0.55
	10	324,333	6.32	325,818	0.52	13.20	0.60
	20	327,063	6.07	328,535	0.52	14.45	0.61
	33	271,948	6.54	273,177	0.46	-6.13	0.41
	50	208,224	3.88	208,607	0.21	-38.40	0.00
	67	228,776	6.26	230,206	0.59	-22.22	0.42
	75	191,842	12.45	193,772	0.62	-25.34	0.32
Meat × meat gain	0	290,078	2.23	291,220	0.70	0.70	0.70
	1	299,089	2.23	300,261	0.70	6.23	0.73
	10	324,337	2.07	325,505	0.71	22.88	0.81
	20	327,063	1.98	328,243	0,73	25.15	0.84
	33	271,948	1.97	272,862	0.62	-11.38	0.57
	50	208,224	1.18	208,754	0.53	-69.36	0.28
	67	228,776	2.28	230,062	0.89	-36.91	0.69
	75	191,842	3.99	193,122	0.74	-45.06	0.46
Percent meat × meat gain	0	1,937.7	2.28	1,984.5	0.33	0.33	0.33
	1	1,805.7	2.30	1,854.6	0.36	-0.64	0.32
	10	2,122.2	2.07	2,164.1	0.30	1.77	0.36
	20	2,250.4	1.98	2,292.2	0.30	2.84	0.39
	33	1,309.7	1.97	1,323.7	0.12	-4.98	0.00
	50	1,538.1	1.18	1,557.3	0.21	-3.98	0.01
	67	3,994.3	2.28	4,074.7	0.41	16.07	1.06
	75	1,867.6	3.99	1,914.4	0.25	-0.15	0.23

[a] Data were from 1,475 calves. Variables are defined in the text

[b] The variable denoted by "x" is the first variable listed for each pair. Variable units are listed in Table 1. Sire variance components are listed in 10,000 (variable units)^2, for convenience in presentation

[c] r $\text{res}_{(x,y)}$, observations with missing data deleted; $r_{(x,y)}$, observations in variable units with missing data included; $r_{(xs,ys)}$, observations in standardized units with missing data included

Table 4. Estimation error variances for estimates of genetic correlations from simulated data

Type of estimate[a]	Percent missing[b]	Estimation error variance[c]	No. of estimates outside the parameter space[d]
r $\text{res}_{(x,y)}$	0	0.025	0
	1	0.024	0
	10	0.041	0
$r_{(x,y)}$	1	0.024	0
	10	0.046	0
$r_{(x10,y10)}$	1	0.066	0
	10	0.867	3
$r_{(x100,y100)}$	1	4.616	8
	10	81.676	9

[a] r $\text{res}_{(x,y)}$, observations with missing data deleted; $r_{(x,y)}$, observations with missing data included; $r_{(x10,y10)}$, observations with missing data included with the variable with missing data divided by 10; $r_{(x100,y100)}$, observations with missing data included with the variable with missing data divided by 100

[b] Results with no missing data are equal for all four types of estimates and are therefore presented only for r $\text{res}_{(x,y)}$

[c] Calculation of estimation error variance is described in the text

[d] Ten data sets were generated

Table 4. These results are similar to the results on field data, except that with only 1% of the data deleted, the r $\text{res}_{(x,y)}$ estimates were no more accurate than the $r_{(x,y)}$ estimates. With 10% of the data deleted, the EEV of the r $\text{res}_{(x,y)}$ estimates is 12% lower than the EEV of the $r_{(x,y)}$ estimate. The EEV of both $r_{(x\,10,\,y\,10)}$ estimates with missing data are significantly greater than the EEV of the corresponding r $\text{res}_{(x,y)}$ and $r_{(x,y)}$ estimates, $P < 0.001$. The EEV of the $r_{(x\,100,\,y\,1000)}$ estimates are so large that these estimates are virtually meaningless. Similar trends are evident for the number of estimates outside the parameter space.

The results presented here demonstrate that even if a small fraction of the observations have missing data, genetic correlations computed from equation (1) may be meaningless, apparently due to bias. This problem is somewhat alleviated if the variances of the variables are similar, but transforming the records to equal variances does not result in more accurate estimates than those derived by elimination of missing data. The fact that similar results were obtained both on field and simulated data is an indication that these results should have general applicability.

Acknowledgements. We wish to thank Dr. D. Gianola for useful correspondence and Dr. A. Genizi for his helpful suggestions. This research was supported by the US-Israel Binational Agricultural Research and Development Fund, Project No. US-334-81. This manuscript is contribution no. 1497-E, 1985 series, from the Agricultural Research Organization, The Volcani Center, Bet Dagan, Israel.

References

Agyemang K, Clapp EC, Van Vleck LD (1985) Variance-covariance components associated with trimester yields of milk and fat and multiple trait sire evaluation for trimester yields. J Dairy Sci 68:1233–1240

Anderson RD (1984) Variance components. In: Blup school handbook. Animal Genetics and Breeding Unit, University of New England, NSW, Australia

David PJ, Johnson, RK, Socha TE (1983) Genetic and phenotypic parameters estimated from Nebraska specific-pathogen-free swine field records. J Anim Sci 57:1117–1123

Harvey WR (1970) Estimation of variance and covariance components in the mixed model. Biometrics 26:485–504

Henderson CR (1953) Estimation of variance and covariance components. Biometrics 9:226–234

Henderson CR (1973) Sire evaluation and genetic trends. In: Proc Anim Breed Genet Symp in Honor of Dr JL Lush. American Society of Animal Science and American Dairy Science Association, Champaign IL, p 10

Henderson CR (1984a) Applications of linear models in animal breeding. University of Guelph, Guelph, Canada

Henderson CR (1984b) Estimation of variances and covariances under multiple trait models. J Dairy Sci 67:1581–1589

Iloeje MU, Van Vleck LD, Wiggans GR (1981) Components of variance for milk and fat yields in dairy goats. J Dairy Sci 64:2290–2293

Manfredi EJ, Everett RW, Searle SR (1984) Phenotypic and genetic components of milk and two measures of somatic cell concentration. J Dairy Sci 67:2028–2033

Mondardes HG, Hayes JF (1985) Genetic and phenotypic relationships between lactation cell counts and milk yield and composition of Holstein cows. J Dairy Sci 68:1250–1256

Rothschild MF, Henderson CR, Quaas RL (1979) Effects of selection on variances and covariances of simulated first and second lactations. J Dairy Sci 62:996–1002

Schaeffer LR, Wilton JW, Thompson R (1978) Simultaneous estimation of variance and covariance components from multitrait mixed model equations. Biometrics 34:199–208

Searle SR, Rounsaville TR (1974) A note on estimating covariance components. Am Stat 28:67–68

Thompson R (1973) The estimation of variance and covariance components with an application when records are subject to culling. Biometrics 29:527–550