Estimating the Between Country Genetic Correlation using Four Different Methods

Sijne van der Beek
NRS, P.O. Box 454, NL-6800 AL Arnhem, The Netherlands.
Email: svdb@nrs.nl

Abstract

The genetic correlation between two countries was estimated with four procedures. In the first procedure raw data were analysed. In the second procedure, data were first precorrected and then analysed. In the third procedure, the INTERBULL procedure, deregressed proofs were analysed. The fourth procedure is a simple modification of the INTERBULL procedure. Data were simulated and analysed using the four procedures. Apart from sire-daughter relationships, all simulated animals were unrelated. Alternatives with weak and with strong connections between countries, and with various number of fixed effects were studied. The results for the four methods were very similar. Since the potential bias of the INTERBULL method was not found here, also reasons for this bias could not be inferred from the present study.

Introduction

Ideally, the genetic correlation between countries is estimated using raw data. This is, however, not feasible due to computational limitations but also because such an analysis requires the person analysing the data to have detailed knowledge on the background of the raw data, which is not likely to be the case. Alternatively to analysing raw data, data can be corrected within country for fixed effects and/or random effects after which the corrected data can be analysed to obtain genetic correlations. Corrected data can be further summarized by taking daughter yield deviations or their equivalent: deregressed proofs.

INTERBULL uses deregressed proofs. Genetic correlations are estimated using an EM-algorithm. Since only deregressed proofs are available, residuals can not be computed and therefore in the INTERBULL EM-algorithm an update for the residual variance is obtained from the update of the genetic variance and the heritability as supplied by the participating country.

Sigurdsson et al. (1995) showed that this algorithm can lead to underestimates when data from two countries is weakly connected. Koerhuis (1996) analysed raw data from two countries with weak connections and found a genetic correlation of 0.95. It thus seems that underestimation only occurs for a combination of weak connections with either precorrected data or an algorithm in which assumptions have to be made.

The aim of this study is to investigate potential reasons for the underestimation of genetic correlations with the INTERBULL EM-algorithm. Data were simulated and analysed using four procedures. In the first procedure raw data was analysed, in the second procedure raw data was first corrected for fixed effects and then analysed, in the third procedure deregressed proofs were analysed using the INTERBULL EM-algorithm, and in the fourth procedure also deregressed procedure were analysed but using a modified EM-algorithm.

Method

Simulation of data

A two-country case was simulated. Sires were generated with either having daughters in both countries or having daughters in only one of the two

countries. For simplicity all sires and dams were unrelated to each other. Each dam had one offspring. One generation of sires, dams and daughters was simulated.

Model of analysis

Data were analysed using the model:

$\mathbf{y} = \mathbf{Xb} + \mathbf{Zs} + \mathbf{e}$

where

y is the vector of observations
b is the vector with fixed effects
s is the vector with random sire effects and
e is the residual

Sire effects were predicted using this model. These sire effects were deregressed using the equation

$$\mathbf{d} = \hat{\mathbf{s}} + \mathbf{G}\mathbf{R}^{-1}\hat{\mathbf{s}}$$

where

d is the vector with deregressed proofs
$\hat{\mathbf{s}}$ is the vector of predicted sire effects
G is the genetic covariance matrix and
R the residual covariance matrix

In fact, for the simple design studied here, deregressed proofs are equal to the average of daughter yields corrected for fixed effects:

$$d_i = \frac{1}{n_{di}} \sum_{j=1}^{n_{di}} (y_{ij} - x_{ij}' \hat{b})$$

where

d_i is the deregressed proof of sire i
n_{di} is the number of daughters per sire, and
$(y_{ij} - x_{ij}' \hat{b})$ is the precorrected observation for daughter ij

Estimation of correlations

Four procedures were used to estimate genetic correlations.

In procedure RAW, raw data are analysed using a multitrait sire model that includes the same fixed effects as were simulated. An EM-algorithm is used to obtain estimates for the sire (co) variances and the residual variances.

In procedure PRE, data are first analysed within a country using a general linear model with the true values for the sire variance and residual variance. Then precorrected data $(y_{ij} - x_{ij}'b, \hat{})$ are analysed using a multi trait sire model that includes only country means as fixed effects.

In procedure ITB, first within country sire effects are predicted and subsequently deregressed. Deregressed proofs are then analysed in a multi trait sire model using the EM-algorithm of Sigurdsson et al. which is also used by INTERBULL. In this algorithm, first sire (co) variances are updated. Then, the updated sire variances and a priori specified heritabilities are used to update the residual variances.

In procedure MOD, which is a modification of the ITB procedure, also deregressed proofs are analysed. The procedure differs from procedure ITB in the way residual variances are updated. In the EM-algorithm, first mixed model equations are solved. Solutions for sire effects are used to update the sire variances. In the procedures using raw or precorrected data, residual variances are updated using the residuals that can be computed from the observations and the current solutions for fixed and sire effects. When only deregressed proofs are known, residuals cannot be computed are therefore procedure ITB multiplies the sire variances by an priori specified ratio of residual and sire variance. Procedure MOD uses another approach. Instead of using prior knowledge on the heritability, prior knowledge on the residual variance is used. Before starting the EM-algorithm, the prior for the residual variance is used to simulate residuals.

These residuals are then added to the deregressed proofs to generate ›phantom‹ observations. For each daughter of a sire a ›phantom‹ observation is generated. The phantom observations are corrected so that the daughter average of the phantom observations is equal to the deregressed proof. Now, the EM-algorithm uses the phantom observations. The mixed model equations and equations to update the sire (co) variances are exactly equal to the algorithm of Sigurdsson et al. However, given the phantom observations and solutions from the mixed model, residuals can be computed that are used to update the residual variance. In fact, the only difference between procedure MOD and procedure PRE is that phantom observations are used instead of precorrected observations.

Alternative situations studied

The alternatives simulated varied in connectedness of the data, and number of fixed effect classes for the one fixed effect simulated per country. For all alternatives, 50 sires were simulated that could have daughters in two countries. For the well-connected data set, all sires had 10 daughters in both countries.

For the weak connected data set, 10 sires had 10 daughters in both countries, 10 sires had 10 daughters in country 1 and 1 daughter in country 2, 10 sires had 1 daughter in country 1 and 10 in country 2, 10 sires has 10 daughters in country 1 and 0 daughters in country 2 and 10 sires had 0 daughters in country 1 and 10 in country 2. The number of fixed effect classes was either 2 or 20. The sire variance was 0.225 and the residual variance 0.775. The genetic correlation was simulated at 0.90.

Results and discussion

In Table 1 estimates of the genetic correlation are given for the four procedures studied. In general, estimated correlations are a bit lower than the simulated correlation. Four alternative 4 the average estimate was substantially lower for all methods. This was largely due to one replicate for which all methods gave an estimated correlation of below 0.20. The most important observation is that all procedures give similar results. This was unexpected: for the alternatives with weak connections. Apparently, the conditions simulated here are uncomparable with a real international data set. Possibly, the data of the countries were still too much connected. Alternatively, maybe underestimation only occurs when much of the connections between countries are due to relationships between sires, where in this study connections were only due to a sire having daughters in two countries.

Table 1. Estimated genetic correlation for four procedures and difference from estimate using raw data (average from 20 replicates are given). True genetic correlation is 0.90.

Altern.	# fixed effect classes	Connectedness	RAW	PRE	ITB	MOD
1	2	STRONG	0.865	0.865	0.848	0.871
2	20	STRONG	0.900	0.881	0.869	0.890
3	2	WEAK	0.880	0.880	0.884	0.871
4	20	WEAK	0.820	0.797	0.792	0.807

Conclusion

For the data studied here, no significant differences between the four methods were found.

References

Sigurdsson, A., Banos, G. and Philipsson, J. 1995. Estimation of international (co)variance components. Chapter V in: A Sigurdsson, Multiple trait genetic evaluation of dairy cattle within and across country, dissertation.