The effect of population size of paternal groups and herds on optimal estimation of the heritability index for gestation length in cattle

Jolanta Różańska-Zawieja[1#], Alicja Szabelska-Bereşewicz[2],
Zbigniew Sobek[1], Anna Nienartowicz-Zdrojewska[1],
Joanna Zyprych-Walczak[2], Idzi Siatkowski[2]

[1]Poznan University of Life Sciences, Department of Genetics and Animal Breeding,
ul. Wołyńska 33, 60-637 Poznań, Poland; [#]e-mail: jolek@up.poznan.pl
[2]Poznan University of Life Sciences, Department of Mathematical and Statistical Methods,
ul. Wojska Polskiego 28, 60-637 Poznań, Poland

Estimation of genetic parameters is a crucial element in the process of population improvement. In the case of farm animals this process is based on a sample which is a subset of the whole population. For this reason it is important to know the effect of the size of paternal groups and of the herd on the accuracy of estimations of these parameters, particularly heritability. The aim of this study was to show the effect of the population size of paternal groups and herds on the accuracy of estimation of the heritability index (h^2). The material for the analyses comprised data of Holstein-Friesian cattle born in 2005-2010 and subject to use value assessment in Poland. The trait analysed was gestation length. Calculations using a linear mixed model were performed using the R 3.1.3 platform. The analyses showed that calculations concerning daughters in paternal groups are more accurate when the daughters are in a smaller number of herds, but of greater size. An increase in the size of paternal groups at the expense of their number does not have such a negative effect on the accuracy of the estimate as in the case of a reduced number of small herds. Limiting the estimate to only the largest herds reduces its accuracy.

KEY WORDS: heritability / linear model / dairy cattle / gestation length

The process of population improvement requires knowledge of the genetic parameters of traits subject to improvement, particularly their heritability coefficients (h^2). While assessing h^2 it is crucial to fit the model to the population by including significant effects influencing the values of these traits [11]. Values of genetic parameters are determined based on a sample which is a subset of the whole population. For this reason it is important to know the effect of the population size of paternal groups and herds on the accuracy of estimation of h^2 values.

Data collection, particularly in the case of large farm animals, requires a long period of time (sometimes several years). For this reason it is useful to know at least approximate quantitative requirements for the population, to ensure the estimates of the parameter will meet a predictable level of accuracy [3, 9].

The aim of this study was to show the effect of population size (number of animals) in paternal groups and herds on the accuracy of estimated h^2.

Materials and Methods

The dataset contained information on cows that gave birth during the years 2005-2010. The trait analysed was gestation length. The data concerned Polish Black-and-White Holstein-Friesian (PHF) cows. The initial dataset comprised records on 491,336 cows. In the first stage, i.e. analysis of the effect of the size of herds used for the calculation, we used data pertaining to cows which came from sires with more than 1,000 calves. The resulting data set contained herds ranging in size from at least 400 to at least 1,700 cows. For the most efficient herd size (at least 400 cows), an analysis was carried out to identify the optimal number of paternal groups, using data in the range of at least 400 to at least 1,700 calves from each sire. As a final step we analysed the size of herds and paternal groups at the same time. This resulted in datasets where the size of herds varied from at least 400 to at least 1,100 cows in the herd and the number of calves from each sire varied from at least 400 to at least 1,500.

For these datasets we used the following mixed model:

$$Y_{ijklmno} = s_i + L_j + G_k + M_l + P_m + o_n + B_o + e_{ijklmno}$$

where:

$Y_{ijklmno}$ – is the gestation length

s_i – is the random effect of the *i*-th bull

L_j – is the fixed effect of the *j*-th lactation number ($j = 1, …, 11$)

G_k – is the fixed effect of the sex of the calf ($k = 1, 2$)

M_l – is the fixed effect of the body mass of the *l*-th calf

P_m – is the fixed effect of the *m*-th year ($m = 2005, …, 2010$)

o_n – is the random effect of the *n*-th herd

B_o – is the fixed effect of the *o*-th season of the calf's birth ($o = 1, 2$) (summer, winter)

$e_{ijklmno}$ – is the random error

The following phenotype variance (σ_γ^2) parameter was derived from the estimated variance components:

$$\sigma_\gamma^2 = \sigma_s^2 + \sigma_o^2 + \sigma_e^2$$

where σ_s^2 is the variance of the effect of the bull, σ_o^2 is the variance of the effect of the herd and σ_e^2 is a residual variance. Using these parameters the heritability coefficient (h_S^2) was calculated as follows:

$$h_S^2 = \frac{\sigma_s^2}{\sigma_\gamma^2}$$

We used the variance explained (R^2) as a relevant summarizing statistic of mixed-effects models. The use of R^2 in this type of model is rare, but it is often used in linear models (LM) and generalized linear models (GLM). R^2 has the very useful property of providing an absolute value for the goodness-of-fit of a model, which cannot be given by information criteria [8]. For mixed-effects models R^2 may be categorized into two types. Marginal R^2 represents the variance explained by fixed factors and the conditional R_c^2, which is interpreted as variance explained by both fixed and random factors (i.e. the entire model), and is calculated according to the following equation:

$$R_c^2 = \frac{\sigma_f^2 + \sum_{l=1}^{u} (\sigma_l^2)}{\sigma_f^2 + \sum_{l=1}^{u} (\sigma_l^2) + \sigma_e^2 + \sigma_d^2}$$

where σ_f^2 is the variance of fixed effect components, σ_l^2 is the variance of the lth random effect, σ_e^2 is the variance due to additive dispersion, and σ_d^2 is the distribution-specific variance (details Nakagawa and Schielzeth [8], p. 137).

The estimates were to verify what sizes of random effects would be appropriate to obtain the best model for estimating the heritability coefficient. The choice of the model was made based on the conditional R_c^2.

All calculations were performed using the R 3.1.3 [10], lme4 [2], data.table [4], MuMIn [1] and dplyr [13] packages.

Results and Discussion

First, we focused on the impact of herd size on estimation of the coefficients R_c^2 and h_S^2. For a fixed number of repetitions of the sires (of at least 1,000 calves from one sire) we checked how the values of these coefficients changed for different herd sizes. A summary of the characteristics of the datasets considered is presented in Table 1.

Table 1

Summary of datasets obtained for the fixed number of repetitions for each bull in the dataset (>1,000) and each number of repetitions for the herd in the dataset (range 400 – 1,700)

Size of the herd	Number of cows	Number of sires	Number of herds
>1700	3584	69	2
>1600	3584	69	2
>1500	3584	69	2
>1400	5064	79	3
>1300	7823	85	5
>1200	7823	85	5
>1100	12 378	97	9
>1000	13 431	102	10
>900	16 287	128	13
>800	20 421	143	18
>700	26 364	161	26
>600	31 709	170	34
>500	40 660	188	50
>400	54 255	208	81

For such datasets we performed analyses to calculate R_c^2 values as well as the heritability coefficient h². The results obtained are shown in Table 2.

Table 2

R_c^2 values, variances of random effects and heritability coefficient for the number of repetitions for each bull in the dataset (>1,000) and each number of repetitions for the herd in the dataset (range 400 – 1,700)

Size of the herd	R_c^2	σ_s^2	σ_o^2	σ_e^2	h_S^2
>1700	0.198	0.555	5.362	27.442	0.017
>1600	0.198	0.555	5.362	27,442	0.017
>1500	0.198	0.555	5.362	27.442	0.017
>1400	0.200	0.662	5.015	26.515	0.021
>1300	0.215	0.738	5.199	26.075	0.023
>1200	0.215	0.738	5.199	26.075	0.023
>1100	0.205	0.622	4.597	25.140	0.020
>1000	0.202	0.629	4.398	25.083	0.021
>900	0.199	0.661	4.241	25.463	0.022
>800	0.201	0.989	3.891	25.087	0.033
>700	0.214	1.054	4.295	25.124	0.035
>600	0.211	1.126	4.078	25.120	0.037
>500	0.227	1.180	4.768	25.261	0.038
>400	0.239	1.350	5.161	25.452	0.042

The values of R_c^2 in Table 2 show that the lower the restrictions on the number of repetitions in the herd, the better model is, and that as R_c^2 increases that of h_S^2 increases as well.

Next, we focused on the impact of the number of calves from each bull on estimation of the coefficients R_c^2 and h_S^2. For a fixed herd size (at least 400 cows) we checked how the values of these coefficients change for different numbers of calves from each sire. A summary of the characteristics of the datasets considered is presented in Table 3.

Table 3

Summary of datasets obtained for the fixed herd size (at least 400) and each number of calves from each sire in the dataset (range 400 – 1,700)

Number of calves	Number of cows	Number of sires	Number of herds
>1700	13659	6	294
>1600	18549	9	318
>1500	21681	11	321
>1400	24576	13	324
>1300	27264	15	328
>1200	34729	21	343
>1100	43776	29	346
>1000	51045	36	346
>900	59480	45	347
>800	70591	58	347
>700	77992	68	347
>600	31709	170	34
>500	40660	188	50
>400	54255	208	81

For such datasets we performed analyses to calculate R_c^2 values and the heritability coefficient. The results are shown in Table 4.

The values for the heritability coefficient increase as the number of calves from each sire decreases. However, the heritability coefficient values are higher than the data in Table 2, falling within the range of 0.055 and 0.114.

As the next step we calculated the h_S^2 and R_c^2 values for each combination of the number of repetitions for each herd (from >400 to >1,100 cows) as well as each bull (from >400 to >1,100 calves) in the dataset. The results are presented in Figure 1 and Figure 2, respectively.

The optimal solution would be to maximize the h_S^2 coefficient and R_c^2 values at the same time. Figure 1 shows that for a number of repetitions for each herd ranging from

Table 4

R_c^2 values, variances of random effects and heritability coefficient for the fixed number of repetitions for each herd in the dataset (>400) and each number of repetitions for the bull in the dataset (in range 400 – 1,700)

| Number of calves	R_c^2	σ_s^2	σ_o^2	σ_e^2	h_S^2
>1700	0.140	1.596	1.409	25.783	0.055
>1600	0.142	1.754	1.377	26.112	0.060
>1500	0.154	2.229	1.351	26.068	0.075
>1400	0.149	1.925	1.436	26.180	0.065
>1300	0.178	3.001	1.439	26.185	0.098
>1200	0.171	2.609	1.448	25.813	0.087
>1100	0.160	2.460	1.261	25.879	0.083
>1000	0.163	2.447	1.260	25.659	0.083
>900	0.171	2.604	1.299	25.617	0.088
>800	0.164	2.390	1.244	25.645	0.082
>700	0.166	2.449	1.250	25.757	0.083
>600	0.173	2.649	1.277	25.582	0.090
>500	0.177	2.764	1.309	25.605	0.093
>400	0.196	3.479	1.338	25.675	0.114
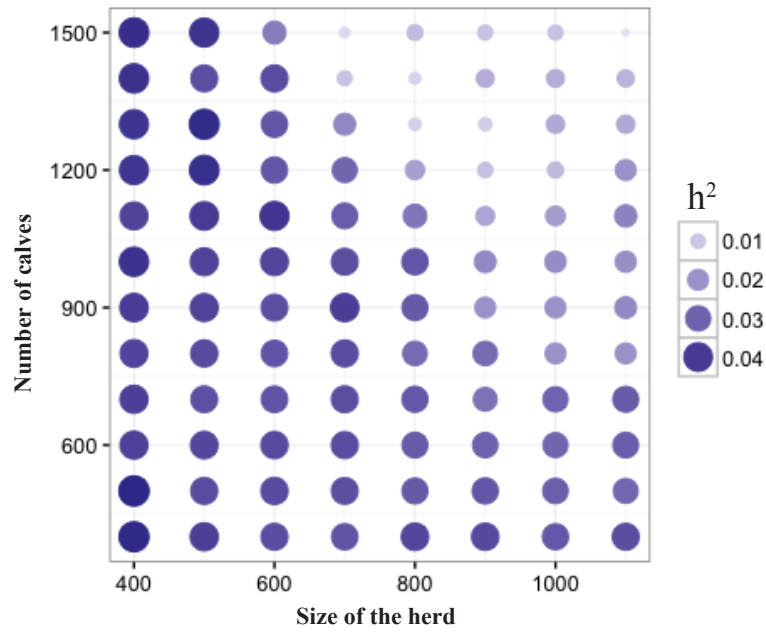
Fig. 1. Heritability coefficient for the number of repetitions for each herd in the dataset (range 400 – 1,100) and each number of repetitions for the bull in the dataset (range 400 – 1,500)
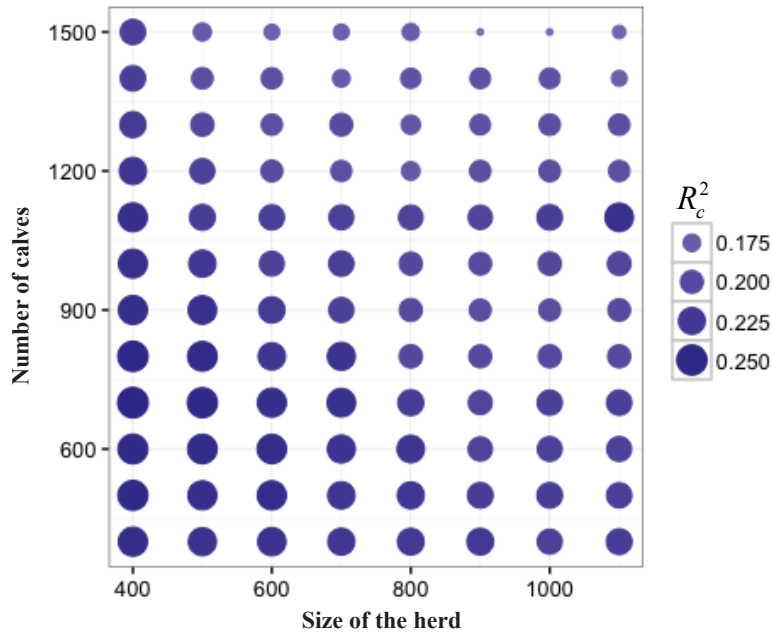
14

Fig. 2. R_c^2 values for the number of repetitions for each herd in the dataset (range $400 - 1{,}100$) and each number of repetitions for the bull in the dataset (in range $400 - 1{,}500$)

>400 to >600 the h_S^2 coefficient remains at a similar level, while for a smaller number of repetitions for the herd and calves the h_S^2 coefficient decreases.

In contrast, Figure 2 shows that the highest values of R_c^2 are obtained for the smallest number of repetitions of herd and bull together.

The study investigated the effect of population size of paternal groups and herds on the value of the heritability coefficient h_S^2 for gestation length in cattle. For this purpose a linear mixed model was used, which the authors found to be most accurate in the preliminary investigations based on the conditional R_c^2. The highest value of h_S^2 was obtained for the variant of over 400 calves from a sire in a herd ($h_S^2 = 0.114$). An increase in the herd size to over 1,700 calves from the sire caused a considerable reduction in the number of paternal groups, which caused a decrease in the h_S^2 value to 0.055.

Khan and Khan [7] presented results of heritability coefficients in native cattle. These values were compared to numerous studies conducted worldwide. The results of these estimations varied in terms of both values and accuracy. According to Khan and Khan [7], these differences resulted both from the diversity of the world cattle population and the varied sample population sizes. Thus in our study we attempted to show the extent of the effect of the number of objects and their size in the population on the accuracy of the estimation.

15

Rzewuska and Strabel [12] showed that for milk yield traits the effect of lactation number on the value of h² is so great that it fully justifies the introduction of the effect of lactation number to the model.

Depending on the specific character of a given population the model may be fitted using additional effects, as in the case of a study by Ismael et al. [5], where the effect of geographical location was included. This extension of the model was justified when the authors had 811,285 records from primiparous cows from 7,458 herds from 3,432 sires.

Jenko et al. [6] also investigated the effect of the number of observations in herd-year classes on the values of estimated h². Analyses were conducted on a population of over 42,000 animals. The conclusions drawn by the authors indicated low susceptibility to diversification of estimations depending on the adopted criterion of population size in the herd-year class. The authors showed that an increase in the population size in herd-year classes caused a minimal increase in the heritability coefficient estimations.

In the present study, the coefficient of determination (R_c^2) was used to show that if calculations concern paternal groups of sires whose daughters are found in a smaller number of herds, but of greater size, the estimations obtained are more accurate and h² values increase slightly.

For the investigated trait it was shown that an increase in the population size of paternal groups at the expense of their number does not decrease the estimation accuracy as much as restriction of the number of small herds. Limiting the estimation to only the most numerous herds reduces the accuracy of estimation.

REFERENCES

1. BARTOŃ K., 2016 – MuMIn: Multi-Model Inference. R package version 1.15.6. https://CRAN.R-project.org/package=MuMIn.
2. BATES D., MAECHLER M., BOLKER B., WALKER S., 2014 – _lme4: Linear mixed-effects models using Eigen and S4_. R package version 1.1-7. URL: http://CRAN.R-project.org/package=lme4.
3. BAUSELL R.B., LI Y.-F., 2002 – Power Analysis for Experimental Research. A Practical Guide for the Biological, Medical and Social Science, Cambridge University Press.
4. DOWLE M., SHORT T., LIANOGLOU S., SRINIVASAN A. with contributions from SAPORTA R. and ANTONYAN E., 2014 – data.table: Extension of data.frame. R package version 1.9.4. URL: http://CRAN.R-project.org/package=data.table.
5. ISMAEL A., STRANDBERG E., BERGLUND B., KARGO M., FOGH A., LØVENDAHL P., 2016 – Genotype by environment interaction for the interval from calving to first insemination with regard to calving month and geographic location in Holstein cows in Denmark and Sweden. *Journal of Dairy Science* 99, 5498-5507.
6. JENKO J., GORJANC G., KOVAČ M., DUCROCQ V., 2013 – Comparison between sire-maternal grandsire and animal models for genetic evaluation of longevity in a dairy cattle population with small herds. *Journal of Dairy Science* 96, 8002-8013.

7. KHAN M.A., KHAN M.S., 2016 – The heritability estimates of linear type traits in sahiwal cows. *Journal of Animal and Plant Sciences* 26 (1), 25-33.
8. NAKAGAWA S., SCHIELZETH H., 2013 – A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4 (2), 133-142.
9. NORMAN H.D., WRIGHT M.T., KUHN S.M., HUBBARD J.B., COLE J.B., VANRADEN P.M., 2009 – Genetic and environmental factors that affect gestation length in dairy cattle. *Journal of Dairy Science* 92, 2259-2269.
10. R Core Team, 2015 – R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.
11. RÖNNEGÅRD L., LEE Y., 2013 – Exploring the potential of hierarchical generalized linear models in animal breeding and genetics. *Journal of Animal Breeding and Genetics* 130 (6), 415-416.
12. RZEWUSKA K., STRABEL T., 2013 – Genetic parameters for milk urea concentration and milk traits in Polish Holstein-Friesian cows. *Journal of Applied Genetics* 54 (4), 473-482.
13. WICKHAM H., FRANCOIS R., 2015 – dplyr: A Grammar of Data Manipulation. R pack-age version 0.4.1. URL: http://CRAN.R-project.org/package=dplyr.