Integrated Detection and Correction of Outliers in a Random Regression Test-Day Model

***P. Mayeres* [1]*, A. Gillon*[1] *and N. Gengler* [1,2]**

[1]*Animal Science Unit, Gembloux Agricultural University, B-5030 Gembloux, Belgium*
[2]*National Fund for Scientific Research, B-1000 Brussels, Belgium*

Abstract

Given the large amount of data collected in the field and used in breeding values evaluations systems, it is clear that the presence of errors is nearly inevitable. Quantitative tools have to be defined for detecting outliers. A method to detect and correct that has the particularity to be integrated into a random regression test-day model is presented: if the residuals is greater than *x* residual standard deviation, data are considered as outliers. This study showed that a five residual standard deviation limitation seems to be the best compromise between the correction of abnormal values and the introduction of bias into the evaluation. Relative frequency of outliers indicated that more corrections occurred at peak yield and for high levels of production. This was mainly due to the heterogeneity of variance, which is not yet taken into account. Therefore heterogeneous variance adjustment should avoid this problem. The presented system is implemented in the genetic evaluation for production in the Walloon Region of Belgium.

Introduction

"The data entering a country Genetic Evaluation System (GES) should have high quality, irrespective how "quality" is defined" [INTERBULL, 2001].

All statistical analysis relies on data: on their accuracy depends its validity. In many cases, all of the data can be individually checked and then some statistical tools exist to underline particular deviation (Cook distance, leverage…). These statistics are estimable with mixed models [Christensen *et al.*, 1992], but their cost-effective computation on large data set is impossible and therefore responsible for their poor utilization in the field of dairy cattle breeding value (BV) estimation models. Moreover, such tools are only developed for advising, and the final choice of elimination stays essentially at the appreciation of the modeler.

In the Walloon Region of Belgium, 12.545.989 tests day records of 718.709 lactating cows are analyzed to compute the breeding values of 939.995 animals (August 2003), which is, in regard of other populations a rather small data set. Still, it is obviously impossible to check all data individually. Some adapted quantitative methods for data quality assessment have to be defined, with care of not doing inadvertent selection of data or introduction of bias.

Pre-evaluation filters are common uses to eliminate the extreme erroneous data, corresponding to biological inconsistencies. In that way, recording guidelines of the International Committee for Animal Recording [2002], gives an acceptable range of the daily test values for the main dairy cattle breed.

In a recent paper, Wiggans *et al*. [2003] showed a procedure for the detection and pre-adjustment of abnormal test-day (TD) yields implemented for August 2002 USDA-DHIA genetic evaluation for yield traits. In opposition to simple pre-evaluation filters, they compare the observed yield with the predicted one in order to identify 2% of abnormal data. For that, they established an acceptance interval around the prediction, this prediction being calculated as preceding TD yield plus preceding test interval multiplied by daily yield change.

Random regression test-day model (RRTDM) directly allow predictions.

Therefore the aim of this paper is to present how outliers can be detected and corrected directly inside a genetic evaluation system based on a random regression test-day model (RRTDM).

Materials and Methods

Data

Data was provided by the Walloon Breeding Association (AWE) which manages performance recording data in the Walloon Region of Belgium. Data edition was done to keep the TD records during the first three lactation occurring between 5 and 365 DIM and to exclude unlikely ages for a given lactation or gestation lengths. Additionally, TD yields were limited to 0 to 85 kg for milk, 1.5 to 9.0 % for fat and 1.0 to 7.0 % for protein.

In August 2003, a total of 12.545.989 TD records from 718.709 lactating cows (essentially Holstein, but also dairy and dual-purpose breeds) were finally used for BV estimation.

Model

The aim of this paper being not to discuss about modeling, the description of the model will be brief. Further details will be found in the paper of Auvray and Gengler [2002]. The multilactation, multitrait RRTDM used can be written as:

$$\mathbf{y_c = Xb+Q(Wh+Za+Zp)+e}$$

where $\mathbf{y_c}$ is a vector of precorrected milk, fat and protein test day records, **b** is a vector of fixed effects (herd*test date, stage of lactation, stage of lactation*age at calving*season of calving, gestation stage), **h** is a vector of herd*period of calving environmental random regression coefficients, **a** is a vector of additive genetic random regression coefficients, **p** is a vector of permanent environmental random regression coefficient, **e** is a vector of random residuals, **X**, **W** and **Z** are incidence matrices, **Q** is the covariate matrix for the second order Legendre polynomials. The precorrection applied on the data to account for environmental effects of age inside lactation*stage*breed classes.

Residuals are supposed to be homogeneous inside lactation and inside trait, following a normal distribution of mean 0 and standard deviation σ_e.

Integrated detection and correction

The Preconditioned Conjugate Gradient (**PCG**) algorithm is used for the estimation of the production breeding values in the Walloon region [Lidauer *et al.*, 1999]. At every round of iteration residuals are calculated and used for the integrated detection and correction of outliers: at each round, residuals are estimated and checked individually.

A data is considered as deviating if its residual is greater than *x* residual standard deviations (σ_e). In this case, the residual is blocked to the σ_e limit. The σ_e used were 1.83, 2.06 and 2.25 kg for milk, 9.62, 11.61 and 12.88 g for fat and 6.51, 7.27 and 7.88 g for protein, for first, second and third lactation respectively. Six models were so computed: without residual limitation and with restriction of 3, 4, 5, 6 and 7 σ_e.

Results and Discussion

Pre-evaluation filters

The number of data excluded during the edition process for presenting an abnormal level is presented in Table 1.

Table 1. Number of milk, fat and protein test day values excluded.

Trait	Lower limit		Upper limit	
	Value	N <	Value	N >
Milk	0	0	85	315
Fat %	1.5	2972	9.0	1056
Protein %	1.0	1097	7.0	1976

The limitation on fat and protein percent follow the recommendation of ICAR [2002]. For milk however, the lower and upper limits are not 3 and 99.9 kg as suggested but 0 and 85 kg. The r eason for excluding only impossible lower yields (< 0 kg) was the presence of very low producing animals in older data. The upper limit was found by individual examination of extreme high producing animals. Both limits might be open to review in the future. As shown in Table 1 7416 values were so put to missing values, what represents only 0.02% of the initial records.

Integrated detection and correction

Contrary to the pre-evaluation filters, outliers were not simply eliminated: these were restricted to the residual limit. The idea behind this is that outliers might have different origins and the safest handling is to reduce them towards their expected value.

Table 2 compare the number of expected and observed outliers, given the theoretical residual distribution.

Table 2. Expected and observed outliers for the different models of residual limitation.

Residual limit	**Out of interval values**	
	Expected	Observed
± 3 σ_e	101.475	132.716
± 4 σ_e	2.381	38.541
± 5 σ_e	22	14.177
± 6 σ_e	0	6.691
± 7 σ_e	0	3.883

The number of deviant TD yields increased logically with decreasing residual limit. Approximately 32.000 test day values did not follow the theoretical distribution, which represent only 0.08 % of the whole data set. As explained earlier their elimination might introduce bias, it was decided to use a residual limit. The change in residual value will obviously affect especially the BV estimation of concerned animals. BVs (mean of first three lactations) of all animals were computed and compared for the 6 models (Table 3).

Table 3. Absolute difference between BVs estimated for milk, fat and protein (kg) by the model without integrated detection and those evaluated by the models with a 3, 4, 5, 6 and 7 σ_e limitation.

Residual limit	**Trait**	**BV absolute difference for animals with at least one control deviating of**							
		± 4 σ_e		± 5 σ_e		± 6 σ_e		± 7 σ_e	
		Mean	*STD*	*Mean*	*STD*	*Mean*	*STD*	*Mean*	*STD*
± 7 σ_e	Milk	-	-	-	-	-	-	72.57	130.8
	Fat	-	-	-	-	-	-	2.98	5.36
	Protein	-	-	-	-	-	-	2.32	4.14
± 6 σ_e	Milk	-	-	-	-	4.69	8.00	72.60	130.8
	Fat	-	-	-	-	0.22	0.41	2.99	5.36
	Protein	-	-	-	-	0.15	0.25	2.32	4.14
± 5 σ_e	Milk	-	-	4.93	6.98	14.38	14.66	86.72	141.5
	Fat	-	-	0.22	0.31	0.66	0.70	3.63	5.85
	Protein	-	-	0.15	0.21	0.450	0.45	2.78	4.49
± 4 σ_e	Milk	3.13	5.08	10.37	11.18	20.05	19.72	95.28	149.5
	Fat	0.15	0.23	0.48	0.51	0.96	0.95	4.07	6.20
	Protein	0.10	0.16	0.33	0.34	0.63	0.65	3.06	4.75
± 3 σ_e	Milk	13.11	13.07	23.89	22.97	33.47	31.70	112.3	162.7
	Fat	0.56	0.58	1.05	1.02	1.54	1.50	4.86	6.79
	Protein	0.42	0.40	0.75	0.71	1.06	1.00	3.61	5.17

Animals with TD yields deviating more than 7 σ_e shows the biggest differences, a logical result. When the model residual limit was reduced, variations in BVs became larger for a given class of animals as the correction increased.

Given the results a limit of 5σ_e seems to be a good compromise between the correction of abnormal test day values and the introduction of bias. This allows to limit the major outliers, which were responsible of large erroneous BV estimation. The choice of 4 σ_e will reduce 24.464 extra data, but these did not introduce great changes in BVs.

The principle of the detection and correction of outliers used is based on the assumption of goodness of fit of the model. If these is not correct, predictions will be erroneous, and some data will incorrectly be defined as outliers as well as some outliers will not be detected. Particularly the homogeneity of the variance is suspicious: with no correction for variance heterogeneity in the model, we can expect some problems for outliers detection (e.g. animals with extreme high or low production level, at peak yield…).

Theoretically, the distribution of outliers will be independent of the level of production or of the stage of lactation. This is studied in Figures 1 and 2.

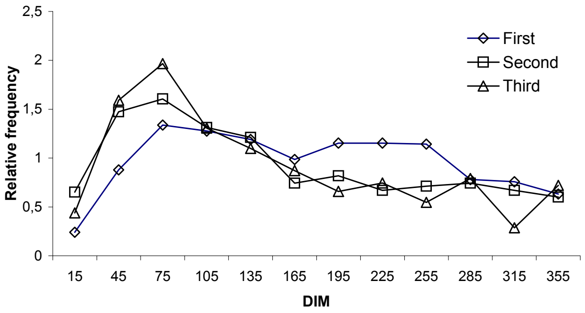

Figure 1. Evolution of the relative frequence (ν) of outliers (=$\nu_{outliers}$ / $\nu_{population}$) across DIM for milk in the first three lactations.

The number of deviant TD yields increase at peak yield, it maybe due to a biggest variance at this stage of lactation. It is interesting to observe that the frequency of outliers did not grow at early and late stages of lactation. This shows that models has a sufficient fit for these data. However this can be due to a problem in the detection of outliers for the lower productions of the beginning and the end of the lactation.

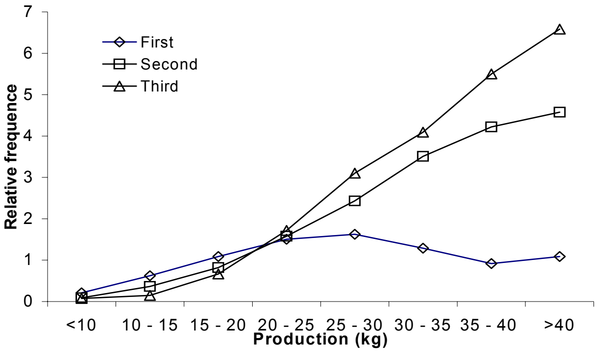

Figure 2. Evolution of the relative frequence (ν) of outliers (=$\nu_{outliers}$ / $\nu_{population}$) by level of production for milk in the first three lactations.

The findings shown in Figure 2 seem to confirm this hypothesis. Very few corrections are made for low production, contrary to high levels. This could be mainly due to the higher variance for high production animals. With the 5 σ_e limitation, an animal having a mean production of 15 kg in first lactation has less chance to be out of the 15 ± 5*1.83 interval than another with the 35 ± 5*1.83 interval.

However, it is important to note that the number of outliers for high production animals remains low, many appearing for animals with intermediary production, the most represented in the Walloon population (Figure 3).

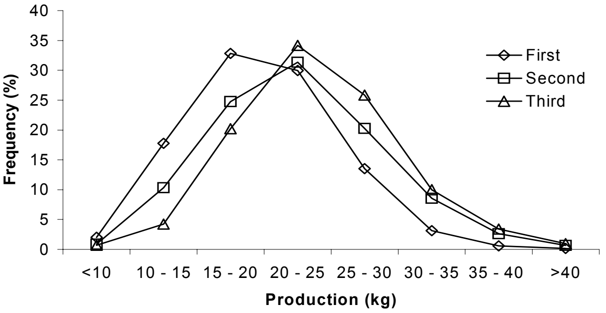

Figure 3. Evolution of the frequence (ν) of outliers by level of production for milk in the first three lactations.

Conclusions

The procedure of integrated detection of outliers in the Walloon Region BV estimation was presented. Until now a limitation of 6 σ_e was used, but at the look of these results, a 5 σ_e restriction seems more convenient.

Some problems, mainly due to the heterogeneity of variance, were outlined. Considerable work is actually in progress to take into account this issue in our evaluation model. Integrating both issues could lead to a more flexible solution than using a strict interval as the one proposed by ICAR [2002] for injured animals (going from 60% to 150%).

Acknowledgments

The authors gratefully acknowledged the support of the Walloon Breeding Association (**AWE**) and the Walloon Regional Ministry of Agriculture (Projects : RW1009, D31-1039).

Nicolas Gengler, who is Chercheur Qualifié of the National Fund for Scientific Research (Brussels, Belgium), acknowledges support through Grant 2.4507.02 F (2) of the National Fund for Scientific Research.

References

Auvray, B. & Gengler, N. 2002. Feasibility of a Walloon test-day model and study of its potential as tool for selection and management. *Interbull Bulletin 29,* 123-127.

Christensen, R., Pearson, L.M. & Johnson, W. 1992. Case-deletion diagnostics for mixed models. *Technometrics 34,* 38-45.

ICAR. 2002. International agreement of recording practices. International committee for animal recording (ICAR), Rome, Italy. Online. Available: http://www.icar.org. Accessed Aug. 10, 2003.

INTERBULL. 2001. Interbull guidelines for national and international genetic evaluation systems in diary cattle with focus on production traits. *INTERBULL,* Uppsala, Sweden.

Lidauer, M., Strandén, I., Mäntysaari, E.A., Pösö, J. & Kettunen, A. 1999. Solving large test-day models by iteration on data and preconditioned conjugate gradient. *J.Dairy Sci. 82,* 2788-2796.

Wiggans, G.R., VanRaden, P.M. & Philpot, J.C. 2003. Technical Note: Detection and adjustement of abnormal test-day yields. *J.Dairy Sci. 86,* 2721-2724.