Session 15 – Abstract 4833

Detection and Correction of Outliers for Fatty Acids Contents Measured by Mid-Infrared Spectrometry Using Random Regression Test-Day Models

H. Soyeurt [1], P. Dardenne [2], and N. Gengler [1,3,*]

[1] Gembloux Agricultural University, Animal Science Unit, Passage des Déportés 2, 5030 Gembloux, Belgium
soyeurt.h@fsagx.ac.be
[2] Walloon Agricultural Research Center, Quality Department, Chaussée de Namur 24, 5030 Gembloux, Belgium
[4] F.N.R.S., Rue d'Egmont 5, 1000 Brussels, Belgium
** Presenter*

Gembloux Agricultural University

1. Introduction and Objective

The use of **fixed limitations based on** the study of **extreme values** for a specific traits (in this case, **the saturated fatty acids (SAT)**) **is not flexible enough.** Some cows can produced naturally low or high contents of SAT. In the same way, some cows can produce milk with an acceptable content of SAT and be sick. The proposed method is to **detect outliers based on residual limitation.**

2. Materials and Methods

Animal Population

A total of 58,443 test-day SAT spectra recorded from 16,470 first parity Luxembourg Holstein cows in 699 herds. Records were collected from October 2007 to January 2009. Spectra were obtained by a FOSS MilkoScan FT6000. SAT content (g/dl of milk) was estimated by mid-infrared spectrometry using the Belgian MIR calibration equation.

Methodology

Residuals for SAT were estimated by solving this random regression test-day BLUP model :

Fixed effects : herd*date of test; stage of lactation; and age.

Random effects : herd*calving year, animal additive, permanent environment regressed using second order Legendre Polynomials. Residual variance was assumed to be constant through the lactation.

A pre-filter was applied to detect abnormal values based on the study of extreme SAT values. Records showing SAT content superior to 4.61 g/dl of milk and inferior to 1.67 g/dl of milk were deleted. The data set contained 57,574 test-day SAT records. Variance components were estimated by REML. The residual standard deviation (se) was equal to 0.29 g/dl of milk.

Different thresholds of residual limitation were studied on un-filtered data: 0.2 se, 0.4 se, 0.6 se, 0.8 se, 1.0 se, 1.2 se, 1.4 se, 1.6 se, 1.8 se, 2.0 se, 4.0 se, 6.0 se, 7.0 se, and 8.0 se.

Optimum residual limitation should be determined by considering 2 criteria: the correlation calculated between observed and predicted values for SAT content; and the amount of remaining data. The outliers were replaced by the predicted values.

3. Results and Discussion

Table 1. Descriptive statistics of dataset before and after the deletion.

	N	Mean	SD	Min	Max	Skewness	Kurtosis
Total	58,443	2.92	0.6	0.33	9.96	0.73	2.81
Outlier detection [1]							
Fixed limitation	57,574	2.91	0.54	1.68	4.61	0.42	-0.08
0.2*se	13,775	2.88	0.48	0.74	8.59	0.63	2.83
0.4*se	25,401	2.88	0.48	0.74	8.59	0.54	1.73
0.6*se	34,838	2.89	0.49	0.74	8.59	0.49	1.24
0.8*se	42,053	2.89	0.5	0.74	8.59	0.49	1.02
s_e	47,158	2.89	0.51	0.74	8.59	0.49	0.91
1.2*se	50,673	2.89	0.52	0.74	8.59	0.48	0.81
1.4*se	53,004	2.9	0.53	0.74	8.59	0.47	0.79
1.6*se	54,525	2.9	0.54	0.74	8.59	0.47	0.74
1.8*se	55,634	2.9	0.55	0.74	8.59	0.47	0.72
2*s_e	56,334	2.9	0.56	0.74	8.59	0.47	0.74
3*s_e	57,827	2.91	0.58	0.59	8.59	0.49	0.89
4*s_e	58,169	2.91	0.59	0.44	9.26	0.53	1.25
5*s_e	58,332	2.92	0.59	0.41	9.26	0.56	1.48
6*s_e	58,334	2.92	0.6	0.33	9.26	0.57	1.57
7*s_e	58,415	2.92	0.6	0.33	9.26	0.61	1.82
8*s_e	58,426	2.92	0.6	0.33	9.26	0.63	1.88

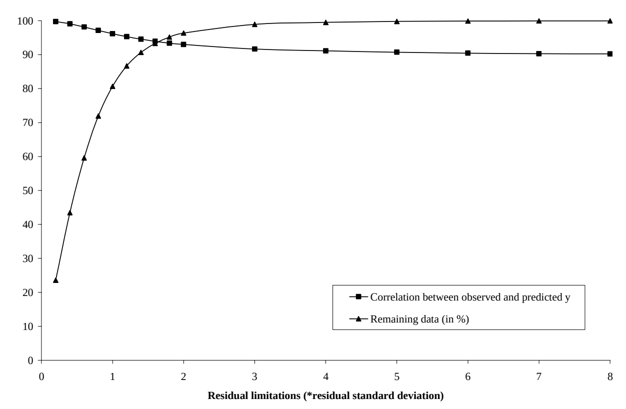

Figure 1. Evolution of the correlation between observed and predicted values and the percentage of remaining data after deletion based on several residual limitation.

Larger number of outlier at the beginning of lactation.

The estimation of the sum of squared residuals and the sum of absolute residuals were inferior using the corrected data than the initial data set suggesting a better model performance.

The ranking of animals stayed unchanged (Spearman correlation superior to 0.999).

The results obtained in this study showed also the evolution of estimated breeding values across lactation.

4. Conclusion

The proposed method is **efficient to detect and correct abnormal values** and to estimate **robust breeding values**.

The authors acknowledge also the support of the European project entitled RobustMilk (Grant agreement n°211708). This project is financially supported by the European Commission under the Seventh Research Framework Programme. The contents of this paper are the sole responsibility of the publishers, and they do not necessarily represent the views of the Commission or its service.