Multivariate outlier detection of dairy herd testing data

Ho Chang Choi[1] H. P. Edwards[2] C. Hassell Sweatman[3]
V. Obolonkin[4]

(Received 27 January 2016; revised 16 May 2016)

Abstract

This paper describes the challenge presented by the Livestock Improvement Corporation regarding the need to detect multivariate outliers in very large datasets of dairy herd milk testing data. Various approaches and techniques were applied to a subset of one dataset in order to establish the potential of both manual and automatic detection of outliers in large datasets using multivariate statistical techniques.
Keywords: statistics, multivariate statistics, outlier detection

Contents

http://journal.austms.org.au/ojs/index.php/ANZIAMJ/article/view/10512 gives this article, © Austral. Mathematical Soc. 2016. Published May 31, 2016, as part of the Proceedings of the 2015 Mathematics and Statistics in Industry NZ Study Group. ISSN 1445-8810. (Print two pages per sheet of paper.)

1 Introduction

Livestock Improvement Corporation (LIC) provides herd testing services for 72% of all New Zealand dairy farmers. Herd testing is the process by which dairy farmers estimate the productive output of individual cows in their herds. LIC staff visit each farm two–four times a year to collect milk samples and deliver them to LIC laboratories. Every year LIC analyses over 20 million individual milk samples collected from approximately 3.5 million cows. The subsequent data set includes the main production trait measurements like milk yield, fat, protein and lactose content, as well as somatic cell count and some corresponding metadata.

The primary purpose of herd testing is to provide dairy farmers with information they require to make culling and replacement stock selection decisions. Herd testing data also contributes to the 'industry good' National Animal Evaluation system for dairy sire evaluation, developed and operated by LIC on behalf of New Zealand dairy farmers. LIC uses the data for the analysis of the dairy cow population in order to assist farmers with on farm selection decisions, as well as the evaluation of LIC's own sires in its sire proving scheme. LIC also uses the herd testing data in wide range of research and development projects in the fields of genomics, farm management and farm automation.

LIC's challenge was to develop methodologies for multivariate outlier detection of the herd testing data, and in particular methodologies which would enable accurate distinction between "erroneous" outliers (i.e., errors in measurement and recording) and "genetic" outliers (i.e., cows with good or possibly bad genetic traits for milk production). They also wanted to be able to detect and classify outliers in real time.

2 Datasets

Although LIC had many different datasets (and large datasets), a relatively simple dataset of 886,000 records was provided for further study. This dataset contained several key response variables of interest such as milk volume, fat percent and protein together with covariates such as time and date, location (using map coordinates), and herd size. From this dataset a subset of 30,000 cases was randomly generated for evaluation and testing. To reduce the complexity of the problem, attention was further restricted to those records taken from herds over one milking season (2012–13) and where each cow had two milk samples were taken (morning and afternoon). This resulted in a working dataset of 6760 cases, which was further reduced to 6743 cases after cases with zero measurements were removed. Thus there were four response variables of interest: am milk volume, pm milk volume, fat percent and protein percent. This enabled the challenge group of mathematical scientists from wide backgrounds to work collectively on commercially sensitive data and produce results which would be indicative of what LIC might reasonably expect if the same methods were applied to other full datasets.

3 Robust Multivariate Outlier Detection

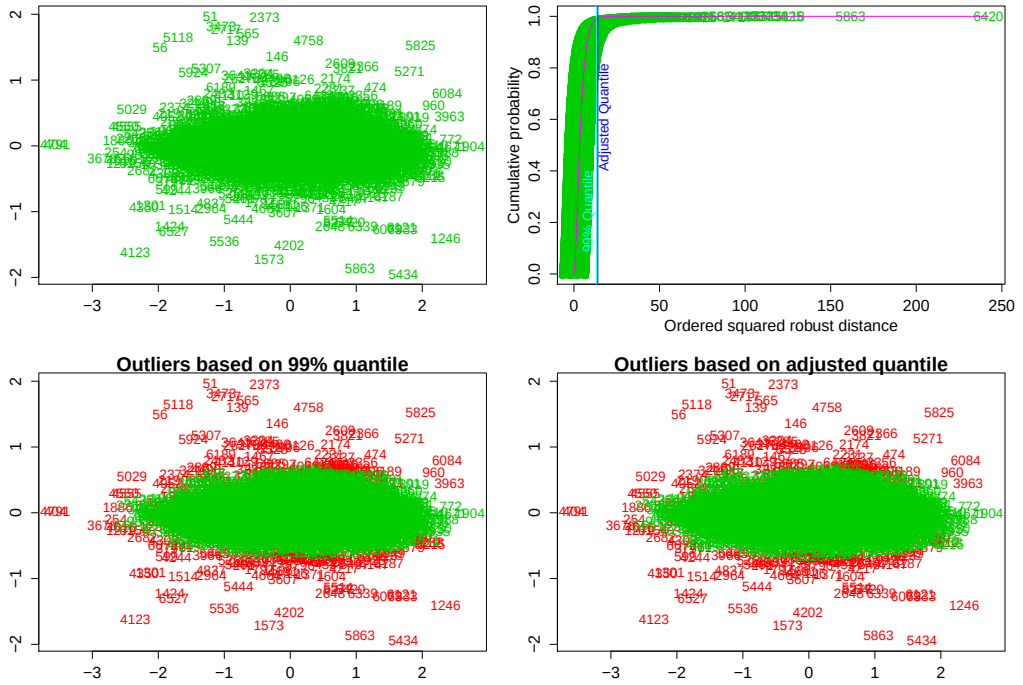
The R package `mvoutlier` was used to test the feasibility of outlier detection on the working dataset. The package `mvoutlier` implements robust mul-

tivariate outlier detection procedures described by Filzmoser, Garrett and Reimann [1] and Filzmoser, Maronna and Werner [2] using a high exclusion criterion. These methods are robust in the sense that robust estimates of location and scatter based on the minimum covariance estimator are used to determine Mahalanobis distance measures, as the classical Mahalanobis measure is too highly sensitive to outliers [4]. However they still assume that the underlying data set follows a multivariate normal distribution. For this reason, the data were transformed using square root transformations on the milk volume measurements and log transformations on the fat and protein percent measurements. The R package MVN was used to test for multivariate and univariate normality and the results indicated that (in comparison to the untransformed data) the bulk of the transformed data values were normally distributed apart from the extreme values. This is what might be expected in a dataset that is suspected of containing outlying values.

Firstly, the adaptive outlier detection method of Filzmoser et al. [1] (labelled AQ below) was applied and the resulting adjusted quantile plot and related scatterplots are shown in Figure 1. This plot shows that the adjusted and unadjusted quantiles are almost identical which means that that the two outlier detection methods produce almost the same set of 252 outliers. Figure 2 shows the outliers in a univariate scatterplot matrix. This plot produces separate univariate scatterplots of each component (on the same standardised scale for comparison) using colours and symbols to represent direction and degrees of outlierness. Specifically, the colours use a heat map type of scale to indicate Euclidean distances from the coordinate-wise minimum (blue closest, red furthest away) while the plotting symbol indicates the size of the robust Mahalanobis distance measure (cross means big, circle means little).

The two computationally fast algorithms PCOut and Sign described by Filzmoser et al. [2] were also applied to the transformed data. Both of these methods are computationally efficient for higher dimensional data and provide a useful point of comparison with the adjusted quantile methodology. In addition, PCOut provides a means of distinguishing between location outliers (those which come from a distribution with a different location) and scatter

Figure 1: Adjusted quantile plot of the transformed data.



outliers (those which come from a distribution with a different scatter matrix). Results of the PCOut and Sign procedures are shown in Figure 3 and Figure 4 respectively. PCOut identified 1071 outliers (a surprisingly large value given that this is out of 6743 values in total), including 560 values identified as location outliers and 309 values identified as scatter outliers. Sign identified 272 outliers. Table 1 shows how many values are identified by at least two of these procedures.

The very high number of outliers produced by PCOut is probably because the data had not been deseasonalised. Milk production in New Zealand peaks in late spring and early summer and drops in winter so measurements

Figure 2: Univariate scatterplots of the transformed data.

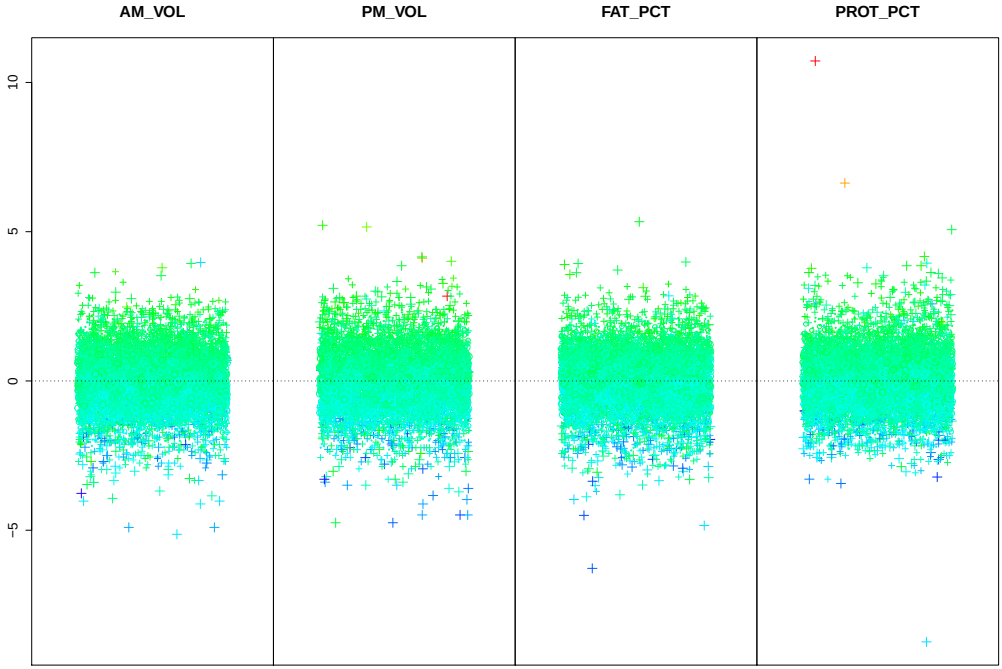
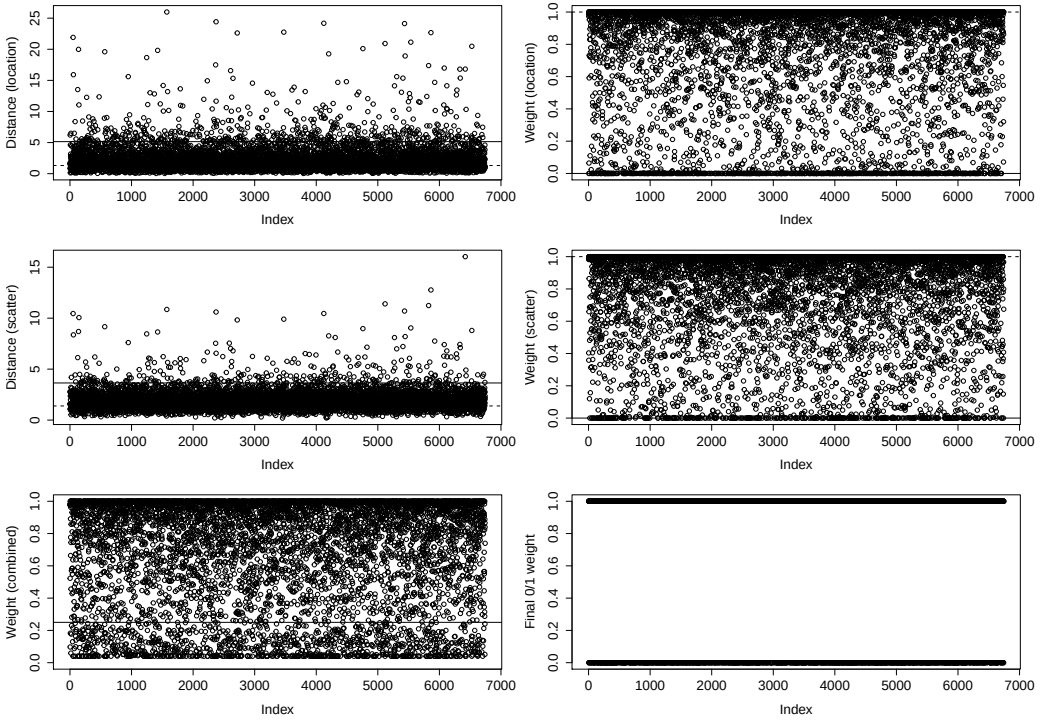


Table 1: Number of outliers detected by each procedure, or by a combination of two procedures.

Procedure	AQ	PCOut	Sign
AQ	252	252	128
PCOut		1071	231
Sign			272

Figure 3: PCOut procedure results. The upper left and middle left plots show location outliers and scatter outliers respectively above respective horizontal lines.

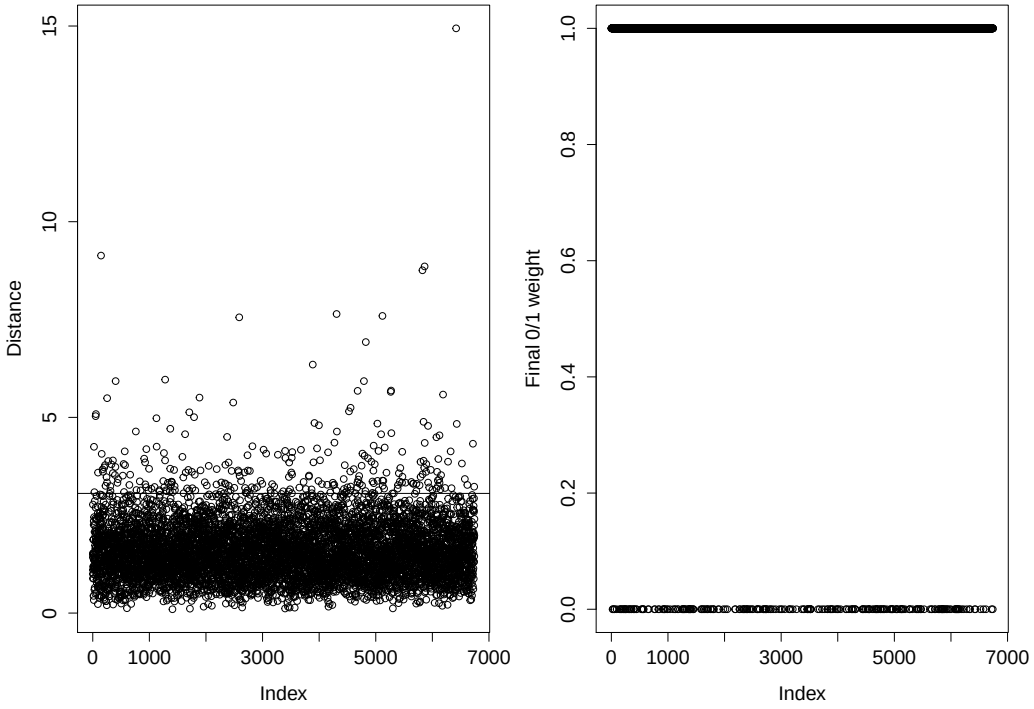


taken at either of these times are more likely to be detected as outliers. In addition PCOut produces two weights for each data point, indicated the relative likelihood of it being a

1. location outlier, and a
2. scatter outlier.

Data points are then classified as outliers of one of these types if the corresponding weight exceeds a threshold value, or if the combined weights exceed a third threshold, or possibly both (which is why the total number of outliers

Figure 4: Sign procedure results. The left plot shows outliers above the horizontal line.



detected exceeds 869). As the object of the analysis was primarily to indicate directions of future analysis, default settings were used for these thresholds regardless of whether they were appropriate for LIC's situation or not.

All 128 values identified as outliers by the Sign procedure were also identified as outliers by the PCOut procedure. Thus as an initial step a total of 128 values in the test data set could be investigated for possible reasons why they appear to be outliers.

Finally, methods based on the minimum covariance determinant distance measure are computationally demanding in higher dimensions, so that any

Table 2: Repeated outlier cow data.

ANML_ KEY	DATE	AM_ VOL	PM_ VOL	FAT_ PCT	FAT_ MASS	PROT_ PCT	PROT_ MASS
23861721	120730	12.0	7.1	4.75	45.3625	4.65	44.4075
23861721	121211	1.7	1.4	12.79	19.8245	4.61	7.1455
23862760	120821	8.0	2.0	2.71	13.5500	3.67	18.3500
23862760	121016	3.9	8.4	3.29	20.2335	3.89	23.9235
26853386	121204	19.6	16.2	3.08	55.1320	3.45	61.7550
26853386	130115	18.2	13.7	3.23	51.5185	3.32	52.9540
22323844	130107	17.3	8.5	3.04	39.2160	3.34	43.0860
22323844	130304	14.7	7.1	3.18	34.6620	3.33	36.2970

further work based on some of LIC's higher-dimensional data sets should incorporate the principal components decomposition methods described by Filzmoser et al. [2].

4 Repeated Outliers

In order to distinguish between erroneous outliers and genetic outliers we considered cows which appeared at least twice per milking season as multivariate outliers. This was possible since LIC staff visit each farm two–four times per year. Cows of interest for breeding or culling might be expected to appear as outliers every time they are measured. This strategy aimed to eliminate unremarkable cows which appear just once per milking season as an outlier due to an erroneous measurement entry. Four cows appeared exactly twice as outliers. No cows appeared more than twice. The data for these repeated outliers are given in Table 2.

The fat mass and protein mass were not used in the outlier analysis but calculated later.

This strategy achieved the aim of eliminating outliers due to obvious technical errors. It delivered a small number of cows of biological interest for further study.

- The first cow (23861721) yielded consistently high fat and protein percentages. The afternoon fat percentage is high due to low milk volume, and corresponds to a low fat mass.
- The second cow (23862760) yielded a low fat percent, and low fat and protein masses.
- The third cow (26853386) yielded high milk volumes, resulting in high fat and protein masses, although the fat and protein percentages were not extreme.
- The fourth cow (22323844) yielded consistently low fat and protein percentages, although milk volume yields were not extreme.

Cows may qualify as outliers for different reasons at different times of year. Our strategy aimed to detect cows which perform well or badly all season. The effectiveness of this strategy is limited by the timing and the number of farm visits per season.

High fat (or protein) percent does not correspond to high fat (or protein) mass if milk volume is low. High fat (or protein) percent is desirable when milk is to be dried. In other circumstances, high fat (or protein) mass may be more desirable.

The dataset contained herd map references (not shown). These could be used to locate the cows of interest and could also be used to investigate possible relationships between biological outliers and geographical location.

One could use multiple methods of identifying cows classed as repeated outliers. For example, unions of the repeated outlier sets would yield larger

sets of cows of interest whereas intersections of the repeated outlier sets would yield fewer cows of interest.

5 Extreme Value Analysis

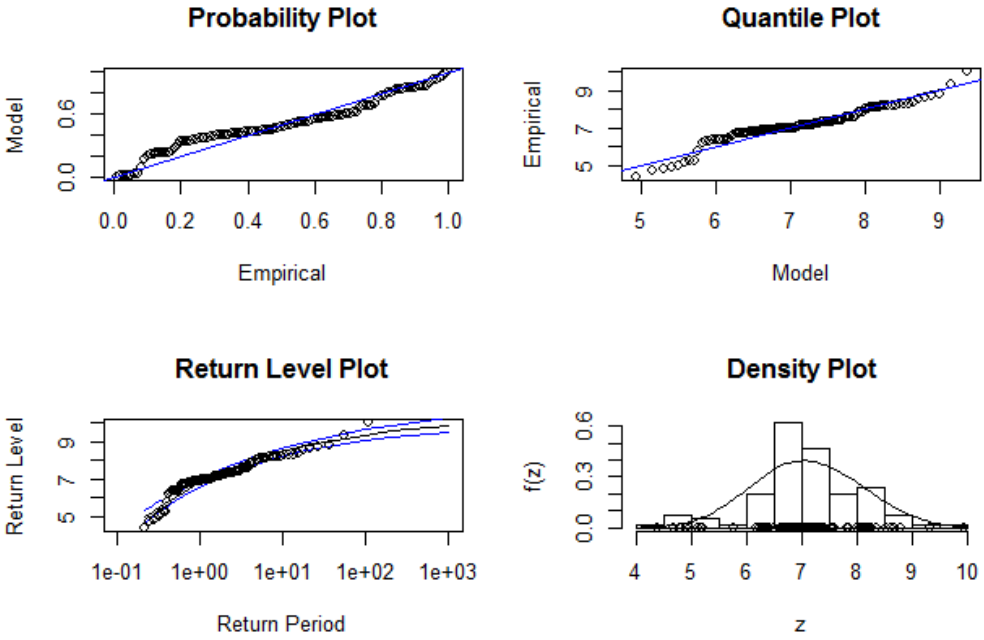
Extreme value analysis was considered as a further approach to successful identification of genetic outliers. In this approach the data is used to fit an extreme value distribution to the dataset after extreme outliers (in this case as identified by the `mvoutlier` package using the 99% criterion) are removed and the resulting fit then examined. Furthermore, a theoretical upper bound can be calculated from the fitted distribution which is then used to indicate what levels of extremes might be expected for a genetic outlier. The extreme value R packages `evd` and `ismev` were used to produce the results in this section. Gilleland, Ribatet and Stephenson [3] discussed extreme value analysis packages in R.

Firstly, we consider extremes with high fat percentage. The estimated mean, standard deviation and the shape parameter of the fitted extreme value distribution are 6.77%, 0.96%, and −0.26, respectively. We now look at the diagnostic plots to do basic analysis, and these are shown in Figure 5. The probability plot suggests that the model is a good fit in the extremes. The quantile level plot suggests that the high extreme values are well represented by the model. Moreover, most of the extremes are well within the confidence interval in the return level plot (shown by the lower and upper curves) which further shows that the model is a good fit. Both the density plot and shape parameter suggest that the extremes follow a negative Weibull distribution. This distribution has a theoretical upper bound

$$\hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} = 6.77 + \frac{0.96}{0.26} = 10.46\%.$$

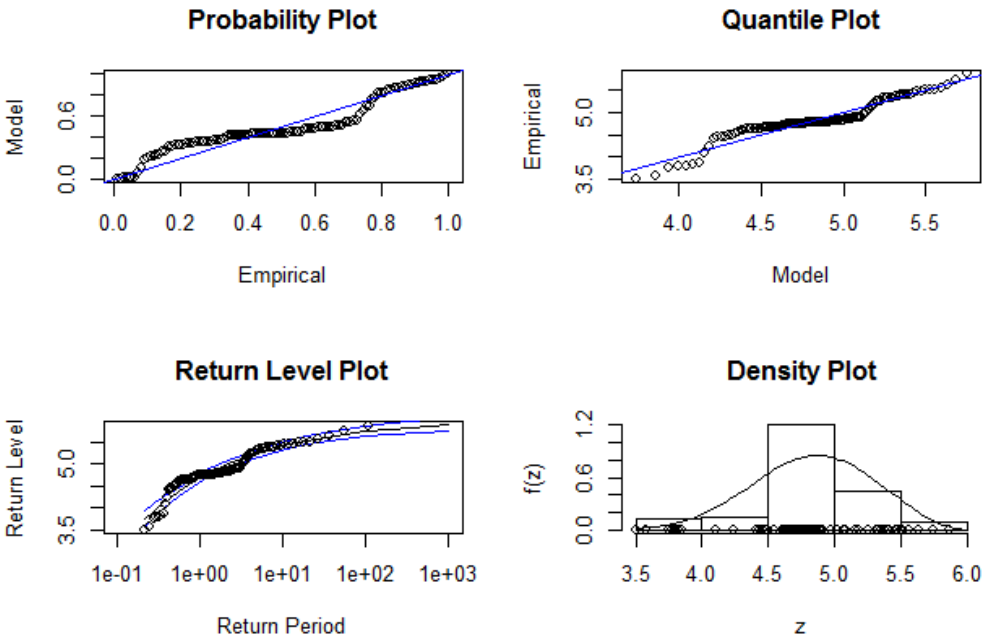
Therefore it is possible that New Zealand cows were able to produce milk with more than 10% fat in 2012–13.

Figure 5: Fat extreme value distribution diagnostic plots



Secondly, consider extremes with high protein percentage. The estimated mean, standard deviation and the shape parameter of the fitted extreme value distribution are 4.69%, 0.46%, and −0.35, respectively. The corresponding diagnostic plots are shown in Figure 6 which show similar results to fat. The probability plot suggests that the model is a good fit in extremes. The quantile level plot suggests that the high extreme values are well represented by the model. Most of the extremes are well within the confidence interval bounds in the return level plot which further shows that the model is a good fit. Both the density plot and shape parameter suggest that the extremes follow negative Weibull distribution. This distribution has a theoretical upper

Figure 6: Protein extreme value distribution diagnostic plots



bound

$$\hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} = 4.69 + \frac{0.46}{0.35} = 6.00\%.$$

This suggests that it is highly unlikely that New Zealand cows were able to produce milk with more than 6% protein in 2012–13.

There were several limitations to this analysis.

1. We only looked at sample data in 2012–13 which may not be representative of other dairy seasons in New Zealand.
2. We did not account for the ages of the cows (milk yield generally increases with age).

3. The geographic information referred to above was not used in the extreme value analysis.
4. We may have removed possible "extremes" during the process of removing the outliers using the `mvoutlier` package.

The analysis could be improved by

1. using the whole data set;
2. removing all systematic and measurement errors;
3. examining correlations between extremes with high fat and protein percentages;
4. considering any patterns in genetic covariates or geographic location if sufficient numbers of cows are found which generally produce more than average.

6 Discussion

The results presented in the above analyses (which are based on a small subset of just one of LIC's datasets) suggest that there is potential for applying many different statistical techniques in order to identify outliers. Clearly there are many other approaches that are also worthy of consideration (for example mixture modelling), and the dataset which was used did not contain any genetic covariates which could be examined or modelled. Nevertheless, the results obtained in a short space of time indicate that statistical modelling is likely to provide valuable insights into LIC's mission to identify the important genetic components of highly producing dairy cow herds in New Zealand.

Acknowledgements We are grateful to Livestock Improvement Corporation for bringing this problem to MINZ-2015. We also acknowledge and thank

the other team members who worked on this problem: Gordon Hiscott, Murray Jorgensen, Xu Dong Liu, Ana Marsanasco, Karen McCulloch, Devendra Oak, and Heather Ricketts.

References

[1] P. Filzmoser, R. G. Garrett and C. Reimann. Multivariate outlier detection in exploratory geochemistry. *Computers and Geosciences*, 31:579–587, 2005. M41

[2] P. Filzmoser, R. Maronna and M. Werner. Outlier identification in high dimensions. *Computational Statistics and Data Analysis*, 52:1694–1711, 2008. M41, M46

[3] E. Gilleland, M. Ribatet and A. G. Stephenson. A software review for extreme value analysis. *Extremes*, 16:103–119, 2013. M48

[4] P Rousseeuw. Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, W. Wertz, editor, *Mathematical Statistics and Applications Volume B*, pages 283–297, Budapest, 1985. Akademiai Kiado. M41

Author addresses

1. **Ho Chang Choi**, Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand.
mailto:hochang.choi@stats.govt.nz

2. **H. P. Edwards**, Institute of Natural and Mathematical Sciences, Massey University at Albany, Private Bag 102904, North Shore Mail Centre, New Zealand.

mailto:h.edwards@massey.ac.nz

3. **C. Hassell Sweatman**, Auckland University of Technology, Private Bag 92006, Auckland 1142, New Zealand.
   mailto:csweatma@aut.ac.nz

4. **V. Obolonkin**, Livestock Improvement Corporation, Private Bag 3016, Hamilton 3240, New Zealand.
   mailto:vobolonkin@lic.co.nz