Uncovering Multivariate Structure Between Milk Production Variables and Udder Health Variables Using Canonical Correlations.

[a]Ersbøll, B.K. & [b]Bruun, J. [a]Informatics and Mathematical Modelling, Building 321, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark. [b]Department of Animal Science and Animal Health, Division of Epidemiology, The Royal Veterinary and Agricultural University, Grønnegårdsvej 8, DK-1870 Frederiksberg C, Denmark

Summary

Present surveys and studies often include measuring large numbers of variables. Commonly the statistical analyses are performed multiple times for different combinations of variables with the inherent risk of inflating the type I error-rate. This study investigates canonical correlation analysis on a dataset consisting of milk production related variables on one hand and udder health variables on the other. Canonical correlation analysis is applied to the variable groups “production related variables” and “udder health variables”. The canonical correlation analysis results in conclusions similar to those from the combination of factor analysis and regression analysis.

Introduction

Present surveys and studies usually include measuring large numbers of variables. Common ways of analysis include measures of association such as: odds ratios and pair-wise correlations, modelling some sort of structure such as: (multivariable) linear or logistic regression analysis. These analyses are usually performed multiple times for different combinations of variables with the inherent risk of inflating the type I error-rate, thereby getting false significances. Although not so common, principle components analysis and factor analysis have also been used to uncover structure. One advantage being that the smaller number of components/factor-scores selected can be used as e.g. regression variables, thereby reducing the chance of false significances considerably.

In this study we will pursue this line of analysis even further by investigating the impact of so-called canonical correlation analysis on a dataset consisting of milk production related variables on one hand and udder health variables on the other. The objective of canonical correlation analysis is to construct new variable(s) as a linear combination of e.g. the milk production related variables and other new variable(s) as a linear combination of the udder health variables in such a way that they pair-wise have maximal correlation between them.

Materials and methods

Data

The data was collected in 89 Danish dairy herds on a random sample of around 20 cows in each herd, in total 1,774 cows were included. The herds were selected as a convenience sample in a region defined by four postal codes situated south of the stream Kongeå in the south of Jutland.

For the present analysis the same variables as considered in a factor analysis by Bruun[1] was used. Production related variables (parity, ln(somatic cell count), energy corrected milk yield and days in lactation) were extracted from the Danish Cattle Database and clinical registrations on teat- and udder health variables were performed.

The clinical registrations considered were: teat skin roughness, udder-asymmetry between hind quarters, udder-asymmetry between fore quarters, udder tissue texture, udder-asymmetry between fore and hind quarters, presence of udder lumps, deep vs. normal udder-shape, small vs. normal udder-shape, short vs. normal teat shape, presence of manure on teats, presence of manure on udder, presence of manure on thigh, and udder-shape between hind legs.

Due to missing information on at least one variable, 306 cows had to be excluded from the analyses.

Canonical Correlation Analysis

Canonical correlation analysis was developed by Hotelling[2,3]. A good recent overview of the application and theory is given in Johnson and Wichern[4].

Given a set of variables, a linear combination of these means first multiplying each variable by an individual weight and then adding them together to form a new variable. Given two sets of variables, a canonical correlation analysis is designed to find a linear combination of the first set of variables and another linear combination of the second set of variables such that these two linear combinations have maximal correlation. This correlation is called the first canonical correlation. The two linear combinations are the first pair of canonical variables. It is now possible to find a second pair of canonical variables, uncorrelated with the first, which have (second-) highest correlation. This can be repeated as many times as there are variables in the smaller set. Canonical correlation analysis has resemblances with both multivariable regression and principal components. It is interesting to note, that the first canonical correlation is at least as high as the multiple correlation between any one variable in one set and all the variables in the other. In this study the procedure cancorr which runs under the SAS system, version 8.2 was used.

Results

Since there are four production related variables and thirteen clinical registration variables the maximal number of canonical correlations is four. The canonical correlations are: 0.52, 0.23, 0.18, and 0.07 respectively. A formal likelihood ratio test shows that the first three of these are highly significant (p-value<0.0001). The first pair of canonical variables shows that parity (weight +0.95) correlates with deep vs. normal udder-shape (weight +0.55) together with small vs. normal udder-shape (weight –0.59). The second pair of canonical variables shows that milk yield (weight +0.60) together with days in lactation (weight –0.57) correlates with udder tissue texture (weight +0.68). The third pair shows that somatic cell count (weight +0.51) together with milk yield (weight –0.83), and days in lactation (weight –0.94) correlates with presence of manure on teats (weight +0.51) together with precence of manure on thigh (weight +0.53).

Discussion

Canonical correlation analysis is applied to the variable groups “production related variables” and “udder health variables”. The canonical correlation analysis results in conclusions similar to those from the combination of factor analysis and regression analysis.

The first pair of canonical variables were comparable to the results of a combination of factor- and regression-analysis, where parity was significant in a regression of production related variables against a factor interpreted as shape of the udder[1]. Similarly, the third pair corresponds with the regression of a factor interpreted as hygiene, where the somatic cell count and days in lactation were both significant.

In addition the canonical correlation analysis as the second pair of canonical variates correlated a generalized milk-yield variable with udder tissue texture.

An important conclusion is that canonical correlation analysis allows for a sensible analysis of sets of correlated variables. This is not guaranteed, neither for a factor analysis on one set followed by a regression on single variables from the other set, nor for regressions on factors from separate factor analyses of each set.

References

1. Bruun, J.: 2003, Risk factors for different health measures in Danish dairy cows – with metritis, antibacterial drug use and udder health as examples. PhD thesis, Epidemiolgy, Department of Animal Science and Animal Health, Frederiksberg, Denmark.
2. Hotelling, H.: 1935, The Most Predictable Criterion. Journal of Educational Psychology, 26, 139-142.
3. Hotelling, H.: 1936, Relations Between Two Sets of Variables. Biometrical, 28, 321-377.
4. Johnson, R.A. and Wichern, D.W.: 2002, Applied Multivariate Statistical Analysis, 5th edition, Prentice Hall