

In G.J. Rowlands, M.N. Kyule and B.D. Perry (eds) Special Issue: Proceedings of the 7th International Symposium on Veterinary Epidemiology and Economics, Nairobi, 15th-19th August, 1994. *The Kenya Veterinarian* 18(2), p 165-170, 1994

COMPARISON OF FOUR MULTIVARIATE TECHNIQUES FOR CAUSAL ANALYSIS OF EPIDEMIOLOGICAL FIELD STUDIES

Pfeiffer, D.U. and Morris, R.S.^a

Multivariate analysis techniques have become standard epidemiological tools in exploring causal relationships, but in relatively few cases have different analytical methods been compared on the same dataset. Four multivariate analysis techniques which can be used for the analysis of a single dependent and multiple independent variables were used on identical data - stepwise regression, path analysis using standard regression, path analysis using structural equation modelling and classification tree analysis. The comparison is based on data collected during a case-control study of tuberculosis breakdowns in cattle herds in New Zealand (Pfeiffer 1994).

MATERIALS AND METHODS

The case-control study data was collected between December 1988 and May 1990 by interview using a questionnaire with 134 items from 95 case herds, 95 random and 95 control herds matched on type of enterprise. The questionnaire referred to farm-specific data, general information on the interviewee, general stock information, stock management information and tuberculosis data. Only the data from the case herds and random controls was used as the basis of the comparison of the multivariate analysis techniques. The general approach to the analysis of the data was based on a univariate and a multi-variate analysis. The initial univariate analysis step consisted of a screening process which was aimed at identifying variables which were statistically significantly associated with case-control status of a herd. A significance level of $p < 0.15$ was used in this analysis step to ensure that potentially important variables were included during the following steps of the analysis. Logistic regression was used for this analysis. The multivariate analysis was conducted using four different methods: Stepwise logistic regression, path analysis using standard regression procedures, path analysis using structural equation modelling and classification tree analysis. Stepwise logistic regression selects a set of variables purely on the basis of statistical significance. Path analysis combines the biological understanding of the researcher with the power of statistical analysis. A causal web including direct as well as indirect effects can be represented by such a model. The first step in the analysis is the development of a null hypothesis path model depicting the hypothesized causal relationships between variables. The hypothesized paths represented in this model are then tested using statistical analysis techniques resulting in a final path model. Two different approaches for path analysis were evaluated. A path model using standard regression procedures (including ordinary least squares and logistic regression) as used in most published studies of this kind in veterinary epidemiology has to be recursive and the analytical technique does not provide an overall assessment of model fit. Relationships between variables are shown in the structure of the model. If logistic regression was used, the regression coefficients can be interpreted as odds ratios. Interaction terms can be included in the model and it is possible to quantify direct, indirect and interaction effects. As a newer alternative, path analysis using structural equation modelling was also applied to the same data. The path model is analyzed as a single model and therefore provides an estimate of overall model fit. The technique allows quantification of direct, indirect and total effects. The model can be non-recursive and it is

^a Department of Veterinary Clinical Sciences, Massey University, Palmerston North, New Zealand

possible to include latent variables. The effects in the model have to be linear and additive. Classification tree analysis uses a technique called recursive partitioning to develop a binary classification tree as a hierarchical-type representation of the data space. Statistical analyses were conducted using the software packages PC-SAS version 6.04 (SAS Institute, Cary, North Carolina, U.S.A.), LISREL version 7.16 (Scientific Software, Mooresville, Indiana, U.S.A.) and CART version 1.1 (California Statistical Software, Lafayette, California, U.S.A.).

RESULTS

Stepwise logistic regression

The final regression model included a set of 6 important factors and two interaction terms. TB breakdown herds had the following characteristics when compared with the control herds: They had a lower proportion of young stock in their herd, they bought stock from a larger number of different herds, the management was using more labour units, they were better informed about the government tuberculosis control scheme, they were closer to the next endemic area, their managers were less likely to work part-time on the farm. The interaction terms implied that herds which had more labour units and were better informed about the control scheme were at reduced risk of breakdown. If they had more labour units and had a part-time manager, they were more likely to break down with tuberculosis infection. This analysis was based on 177 observations. The final model was 70.7% sensitive and 62.2% specific.

Path analysis using ordinary and maximum likelihood regression

The final path model based on this technique included 6 variables with direct and nine variables with indirect effects. The direct effects on TB breakdown status were identical to the effects identified using stepwise logistic regression. The indirect effects indicated that herds which purchase more cattle and those which have more beef cattle are more likely to buy from more different herds. Farms with a larger pasture area employed more permanent labour units and also had more total labour units. Herds with more cattle livestock units had more total labour units employed. Herds with more total cattle livestock units had a smaller proportion of young stock in the total herd. Knowledge about the control scheme was better if a manager did know more about the mechanisms of disease spread and other employment at the same time. The distance to the next herd with a TB breakdown was not important. The number of observations used for the individual regressions varied between 174 and 182.

Path analysis using structural equation modelling

The final path model in this case included 21 variables, of which 12 had a direct effect and 9 had indirect effects on the outcome variable. Most of the factors emphasized the importance of herd management and herd characteristics. Herds with larger beef components were more likely to be case herds. Herds with a larger beef component, bought more cattle, which they sourced from more different herds. If cattle were bought from more different herds, it was likely that the farm used a management system where animals are kept to a significant extent on “run offs”. If cattle were mainly grazed “off farm”, they were more likely to break down with tuberculosis infection. Behaviour patterns of the person managing the herd did have an impact on the risk of breakdown. Managers who preferred the interaction with cattle to working with machinery, who considered themselves less sociable but more persevering were more likely to have TB reactors in their herd. Better knowledge about the epidemiology of tuberculosis and the disease control scheme was found in managers in charge of a herd with

TB reactors. The closer the next endemic area the more likely it was that the herd had reactors. Farms with reactors were using more labour units. If the cattle had access to bush, they were more likely to come from a herd with TB reactors. The squared multiple correlation for this model was 0.642. A total of 182 observations was used for the analysis. The q-plot of the normalized residuals indicates that there are some large positive and negative residuals.

Classification tree analysis

The final classification tree was based on nine variables. Eighty-four percent of 38 herds which had more than 865 livestock units and a proportion of less than 28% young stock, were case herds. Eighty-five percent of 21 farmers who had less than 866 LSU beef cattle, who did not believe in the effectiveness of the control scheme and who bought cattle, had TB reactor cattle. Another branch of the classification tree characterized 78 herds of which 24% were case herds. These herds had less than 866 LSU beef cattle, did think that the current government disease control scheme was effective, had less than 1830 LSU dairy cattle and a farm size of less than 378 hectare. The final classification tree was able to correctly classify 47% of herds as case herds. The probability of correctly classifying control herds was 0.67. the sample size was 182 observations.

DISCUSSION

The objective of this analysis was to improve the understanding of the factors involved in the causal web related to tuberculosis breakdowns in cattle herds. The three techniques involving stepwise regression approach and both path analyses allowed similar insights into the system. All three methodologies identified the increased risk for herds (especially beef herds) which rely on stock purchases to keep up their herd numbers. Farms which had reactors were also more likely to be larger operations with larger herd size, pasture area and more employees. But the path analyses provide a more complete representation of the underlying causal structure. The path analysis using standard regression techniques mainly added indirect effects to the stepwise regression model. The structural equation model provided the most complete view of the system under study. It included grazing cattle mainly "off farm" as a major risk factor. This factor was commonly seen as representing an additional risk, as these cattle could be grazed in areas where they are exposed to tuberculosis infection in wildlife populations. The structural equation model also included a cluster of variables describing farmer personality factors which did appear to have an effect on case-control status of herds. The regression model and the path models included the distance to the next endemic area as a risk factor. Classification tree analysis does not provide information on the structure of a causal web. The classification tree which was developed on the basis of this analysis can be used to classify herds into risk groups within the framework of an expert system. The tree mainly included variables related to the type of enterprise as reflected in the number of beef and dairy cattle and the proportion of young stock in the herd. The technique cannot estimate indirect effects and in this example provided very limited insights into the underlying biological system. In terms of meeting the objective of this analysis the multivariate techniques most adequate would be both path analyses. The structural equation model provides the most comprehensive view of the causal web under study, but is also more difficult to implement and cannot provide quantitative estimates of risk.

REFERENCES

- Pfeiffer, D.U. 1994. The role of a wildlife reservoir in the epidemiology of bovine tuberculosis. Unpublished Ph.D. Thesis, Massey University, Palmerston North, New Zealand, 496pp.