LOGIT Modelling

Jurgen A. Doornik, Nuffield College, Oxford.
Steve Moyle, Oxford University Computing Laboratory.

The main feature of the CoIL challenge data is that the observed response is discrete: we only observe whether or not the customers buy caravan insurance. Models for such data were developed in the 1940s and 1950s in the Bio-statistics literature (see references in Cox and Snell, 1989), in particular LOGIT and PROBIT. In the 1970s such techniques became popular in Economics, where a utility maximizing interpretation was added. By now these techniques are part of the standard statistical toolbox.

LOGIT models assume that there is an underlying unobserved continuous variable which determines the response. This then, is modelled as a linear function of observed characteristics and a random term. Together they determine the predicted probabilities of a particular response ("buy" or "not buy"). Maximum likelihood is used to estimate the coefficients on the observed characteristics in the linear function.

Our objective was to make a submission with as little effort as possible, which is why LOGIT Modelling was adopted. The main problem we faced were the large number of poorly documented attributes (to our benefit one of the authors is a native Dutch speaker). The other problem, which is common to marketing data, is the low incidence of positive outcomes.

The basic procedure was standard LOGIT analysis as implemented in version 10 beta of PcGive (Hendry and Doornik, 1999). Because of the large number of attributes in the data set, this was complemented with the PcGets automatic search procedure. Roughly, the following steps were taken:

1) LOGIT modelling of CARAVAN using the whole data set to get a feel for the data. Splitting attributes according to their value gave potential improvements for: PPERSAUTO, APERSAUTO, PBRAND, PPLEZIER.
2) It was observed that the zip-code attributes (attribute 1-43) did not have much explanatory power by regressing the product ownership attributes (44-85) on all zip-code attributes.
3) PcGets (Hendry & Krolzig, 1999) was used to reduce the model. Strictly speaking, using regression methods when the dependent variable is discrete is incorrect, but it was found to be very helpful.
4) Starting from the most general LOGIT model, the model was simplified, partially guided by the PcGets results. With hindsight we found that the model was over-simplified for forecasting purposes, and this might be an avenue for future research.

Note that the first two steps are based on standard statistical procedures, while the remaining two steps involve machine learning methods (for example, general-to-specific model search as implemented in PcGets).

The model (see figure 1) was used to select the 800 records from the evaluation data based on ranking customer records by predicted probability of purchasing caravan insurance.

Compute z as: z = -4.1 – 0.13* MOPLLAAG - 0.3* MINK123M + 0.17* MINKGEM + 0.54(ALEVEN + AFIETS + ABYSTAND)-(PBRAND ==2 or 6) + 0.8(PBRAND ==3 or 4) + 2.4(PPLEZIER >0) + 1.5(PPERSAUT ==6) +0.28(PWAOREG - PLEVEN);
Then the probability of purchasing CARAVAN insurance is: P(CARAVAN) = exp(z)/(1 + exp(z));

Figure 1: Simplified LOGIT Model for caravan insurance.

Conclusions

With relatively little effort, it was possible to generate a reasonable predictive model. Most scope for improving the model is likely to be in the data pre-processing and transformation steps. We found PcGets very helpful in formulating a final simplified model, however, for forecasting purposes it may be that the default termination criteria are too strict.

References

Cox, D.R. and Snell, E.J. (1989), *Analysis of Binary Data* (2nd edition), London: Chapman & Hall.

Cramer, J.S. (1991), *The LOGIT model: an introduction for economists*. London: Edward Arnold.

Hendry, D.F. and Krolzig, H.-M. (1999), `Improving on 'Data mining reconsidered' by K.D. Hoover and S.J. Perez' *The Econometrics Journal*, **2**, pp. 202-219.

Hendry, D.F. and Doornik, J.A. (1999). *Empirical Econometric Modelling Using* PcGive (2nd edition), London: Timberlake Consultants Press.

Acknowledgements

The first author gratefully acknowledges support from UK Economic and Social Research Council (grant R000237500). The second author is supported by European Union project IST-1999-11495:SolEuNet.