Introduction to Qualitative Modelling

Wendy Olsen

wendy.olsen@manchester.ac.uk

Please cite as:
Olsen W. 2004 Introduction to Qualitative Modelling. Paper presented as part of the Focusing on the Case Workshop series, December, 2004.

In this document, I explore the use of regression modelling and other forms of modelling that allow for qualitative variables. In this introductory document I introduce the notion of a 'qualitative variable' and the three main forms of regression analysis which use them. I conclude by mentioning related non-regression techniques and how causality is addressed in each.

A qualitative variable is a categorical or ordinal measure, expressed as integers in a column, which is recorded for each of many comparable cases. The cases, in the rows, can fit into groups and may differ considerably from each other (e.g. countries of the world, along with regions of very large countries such as China and India). An indicator which is categorical will show mutually exclusive 'values' across the cases, each value having a 'Value Label' (e.g. state-socialist government) that summarise the nature of the type into which this case fits. In Ragin's terminology, a categorical variable places the cases into crisp sets. Gender and ethnicity are classic nominal variables.

Ordinal variables also have mutually exclusive categories, but their measurement metric is arbitrary so the values need not be integers; taking the mean of these numbers is still meaningless (grin). For instance, Likert scales are 'ordinal', and so are grouped income categories such as 0-3000, 3001-6000, >7000.

The standard linear regression model can have a categorical variables among the independent variables. You need to 'binarise' the categories into k-1 dummies for k categories. The regression model best suited to binary or nominal outcomes is the logistic regression model. Here we model:

Y is the logarithm of the odds of Y occurring.
The odds of Y are the ratio of cases where Y occurred to cases where Y did not occur.
The logarithm of the odds of Y creates a continuous variable for any set of cases, i.e. for sub-groups.

For a discussion of the possible advantages of the logistic model, see Olsen and Morgan (mimeo). Actually the log of 1 is 0 and the log of 0 does not exist, so for any particular case the log of the odds of Y [called the logit of Y' does not exist. However statisticians calculate the predicted log-odds of Y and this lies on a scale from 0 to infinite. We can then un-log the odds (called the anti-logarithm; most calculators give this) and work out the error for each case in odds terms. The error, as in all regressions, is the distance from the actual outcome to the predicted outcome. We square the errors and add them up in linear regression, but in logistic regression the optimal solution is found in a different way. We use the minimised likelihood function, and books by Menard and others describe this procedure (Menard, 1995). So

there is no r-squared value for logistic regression. It is considered a non-linear estimate, since the log-odds functions are curved rather than straight line slopes.

The second form of regression analysis groups variants of logistic regression into one basket. Probit analysis analyses the probability of the outcome occurring; tobit analysis analyses the probability under some restrictions (see Gourieroux, 1997, and Long, 1991). You can also extend logistic models to the multinomial outcome using polytomous logistic regression. See Agresti (1984 and 1996).

The third form of qualitative regression takes the model to a multi-level plane. you can have data on individuals, households, regions, and countries all in one model. Two main types of multi-level regression model can be used (and each can be applied both to linear regression with a continuous dependent variable, or to logistic regression with a binary dependent variable):

nested levels, e.g. individuals within households [complex surveys; this is like a 1-to-n database relation, or n-to-1 which is equivalent]

or

non-nested levels, e.g. firms and regions, such that each firm is in some regions and each region has some firms, but they are not nested and not every firm is in every region. This is like a n-to-n relation in 'database management', and you simply index each table of data to the other.

The software to do multi-level modelling with either ML-Win or HLM. Each can be found on google. I will demonstrate ML-Win in the Workshop activity.

Please note that you need extra data at the macro level of analysis to make multi-level modelling really interesting. Now the 'level' of the region or household is qualitative interesting, and you need not only to indicate it (with dummy variables as in ordinary regression) but to have variables that describe the character of each region. Geographers use this a lot now. See Jones, 1997, or other articles by Jones.

References:

Strongly recommended easy reading: Menard, S. (1995). Applied logistic regression analysis. London, Sage.

Other references:

Agresti, A. (1984). Analysis of ordinal categorical data. New York ; Chichester, John Wiley & Sons : John Wiley & Sons.

Agresti, A. (1996). An introduction to categorical data analysis. New York ; Chichester, Wiley.

Menard, S. (1995). Applied logistic regression analysis. London, Sage.

Gourieroux, C. (1991) Econometrics of Qualitative Dependent Variables, CAP, Cambridge

Jones, K. (1997). Multilevel Approaches to Modelling Contextuality: From Nuisance to Substance in the Analysis of Voting Behaviour. Places and People: Multilevel Modelling in Geographical Research. G. P. Westert and R. N. Verhoeff. Utrecht, Urban Research Centre.

Long, J. Scott (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage.
Olsen, W. K. and J. Morgan (2004). A critical epistemology of analytical statistics: addressing the sceptical realist. Mimeo, British Sociological Association, March. York. For a copy contact wendy.olsen@man.ac.uk

9/9/04

www.durham.ac.uk/case.2004