# Modelling A Hierarchical Event: An example of young people and academic qualifications using BHPS data

## *BHPS-2003 Conference,* July, Essex University.

Vernon Gayle
*Department of Applied Social Science, University of Stirling*
*Stirling, FK9 4LA*
*Scotland,*
*vernon.gayle@stirling.ac.uk*
Tel:01786467708
Fax:01786467689

This paper introduces an application of the continuation ration model for longitudinal data. Standard log-linear models do not take ordinality into account, thereby potentially they may disregard useful information. The paper employs a simple social science example to demonstrate that a) the continuation ratio model is clearly beneficial when the outcome of interest has a substantively important hierarchy because this is taken into account in the model b) the continuation ratio model is appropriate when results will not be invariant to reversing the order of the response categories.
The model and associated notation are outlined. Data from a cohort of young people who were 16 in BHPS Wave A is used. A model is estimated for data from Waves A, C and E.
Results are reported for the estimated effects of wave and parental qualifications.
The paper indicates a number of benefits that are associated with this methodology, for example the ability incorporate both time-constant and time-varying explanatory variables, to undertake analysis of state dependence and to increase flexibility in representing residual heterogeneity.

*DRAFT*

# Introduction

A large amount of data analysed within sociological studies consists of categorical outcome variables that can plausibly be considered as having a substantively interesting order (for example levels of attainment of educational qualifications). Standard log-linear models do not take ordinality into account, thereby potentially they may disregard useful information. The continuation ratio model is particularly suitable for data with ordinal, or 'hierarchical', outcomes (Fienberg and Mason 1979; McCullagh and Nelder 1989). Berridge (1992) provides a macro for fitting the continuation ratio model in GLIM and Gayle (1996) has demonstrated the fitting of this model to tabular data within sociological research. The work of Berridge and dos Santos (1996a; 1996b) has developed the continuation ratio model for the analysis of longitudinal data by extending it to the analysis of recurrent events.

The continuation ratio model should be chosen when the outcome variable can be conceptualised as being hierarchical. If the substantive understanding of the structure of the outcome variable is such that results should not be invariant to reversing the order of the response categories then the continuation ratio model is suitable. Furthermore, when the response variable includes a natural baseline category to which all the other categories may be referred the continuation ratio model is appropriate.

The proportional odds (or cumulative logit) model is sometimes considered as a suitable approach for ordinal outcomes. The motivation for the proportional odds model is provided by an appeal to the existence of an underlying continuous and perhaps unobservable random variable. The continuation ratio model is more appropriate than the proportional odds model in cases where the categories of the response variable really are discrete, for example in situations where the response variable is a series of ordered categories and movement from one to another denotes a shift or change from one state to another. In such cases the response categories cannot be thought of as coarse groupings of some finer scale.[1]

[1] See Ian Plewis' comments in McCullagh (1980) discussion.

# Modelling Ordinal Recurrent Events

SABRE is a software package for the statistical analysis of binary recurrent events (**S**tatistical **A**nalysis of **B**inary **R**ecurrent **E**vents). It was developed at Lancaster University, U.K., and is a 'GLIM-like' software package (see Barry, Francis, Davies, and Stott 1998). The continuation ratio model for recurrent events is now incorporated into SABRE Software. In this section I present the model and introduce some of the associated mathematical notation.

Consider a hierarchical (or ordinal) outcome that comprises four categories. In a longitudinal set of survey data, for example, we would commonly have a sequence of ordinal outcomes for each respondent. It is, conceptually at least, straightforward to consider modelling these ordinal recurrent events as a series of conditionally independent binary outcomes via binary logistic regression.

Assuming that an individual $i$ , with vector of explanatory variables $\underline{x}_{it}$ is in category $j(i, t)$ for the $t$-th event. The probability of this individual being in category $j$, given the outcome is in category $j$ or higher is given by Equation 1.

**Equation 1.** The Continuation Ratio Model

$$\frac{\exp(\theta_j + \underline{\beta}'\underline{x}_{it})}{1+\exp(\theta_j + \underline{\beta}'\underline{x}_{it})}; \; j = 1,...,3$$

In Equation 1 the parameter $\theta_j$ is the intercept (cut point) specific to the $j$-th partition of the original outcome, $j$ = 1,...,3, and $\underline{\beta}'$ is the vector of regression coefficients associated with the vector of explanatory variables. $\underline{\beta}'$ is not subscripted by $j$ as we are assuming that the effects of the explanatory variables are common to all three partitions. This assumption can be tested in the modelling process (see Table 2 below).

If the outcome corresponding to the $j$-th partition of the $t$-th event for the $i$-th individual is denoted by $y_{ijt}$ , it equals 1 if $j$ equals $j(i, t)$, it equals 0 for values of $j$ less than $j(i, t)$, and is undefined for values of $j$ greater than $j(i, t)$ if $j(i, t)$ =1,...,3, and equals 0 for all values of $j$ if $j(i, t)$ = 4.

To clarify this Figure 2 reports a sequence of ordinal outcomes for the $i$-th individual.

**Figure 2.** A Sequence of Ordinal Outcomes

1 2 2 3 4

In Figure 3 this sequence is expressed in terms of $y_{ijt}$'s.

**Figure 3.** Ordinal Sequence of Outcomes Expressed as $y_{ijt}$

| Event Number *t* | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Ordinal Outcome | 1 | 2 | 2 | 3 | 4 |
| $y_{i1t}$ | 1 | 0 | 0 | 0 | 0 |
| $y_{i2t}$ | | 1 | 1 | 0 | 0 |
| $y_{i3t}$ | | | | 1 | 0 |

*Note that $y_{i4t}$ is not used in the subsequent equations and is therefore omitted from this table

The likelihood of the *i*-th individual resulting in category *j*(*i*, *t*) on the *t*-th event is expressed in Equation 2.

**Equation 2.** Likelihood of an Individual Resulting in a Category at a Given Event

$$L_{ij(i,\,t)t}(\underline{\theta}, \underline{\beta}; \underline{x}_{it}) = \begin{cases} \prod_{j=1}^{\mathrm{j(i,\,t)}} \dfrac{[\exp(\theta_j + \underline{\beta}'\underline{x}_{it})]^{y_{ijt}}}{1+\exp(\theta_j + \underline{\beta}'\underline{x}_{it})} & , \quad \mathrm{j(i,\,t)} = 1,2,3. \\ \prod_{j=1}^{3} \dfrac{1}{1+\exp(\theta_j + \underline{\beta}'\underline{x}_{it})} & , \quad \mathrm{j(i,\,t)} = 4 \end{cases}$$

In order to handle residual heterogeneity a case specific error, represented by the parameter ω and the variable ε, can be incorporated into the logistic framework (see Equation 3).

**Equation 3.** Case-Specific Random Error

$$L_{ij(i,\,t)t}(\underline{\theta}, \underline{\beta}, \omega; \underline{x}_{it}, \varepsilon_i) = \begin{cases} \prod_{j=1}^{\mathrm{j(i,\,t)}} \dfrac{[\exp(\theta_j + \underline{\beta}'\underline{x}_{it} + \omega\varepsilon_i)]^{y_{ijt}}}{1+\exp(\theta_j + \underline{\beta}'\underline{x}_{it} + \omega\varepsilon_i)} & , \quad \mathrm{j(i,\,t)} = 1,2,3. \\ \prod_{j=1}^{3} \dfrac{1}{1+\exp(\theta_j + \underline{\beta}'\underline{x}_{it} + \omega\varepsilon_i)} & , \quad \mathrm{j(i,\,t)} = 4 \end{cases}$$

In order to eliminate this case specific error, the full likelihood may be integrated over the variable ε (see Equation 4).

**Equation 4.** Full Integrated Likelihood

$$Li(\underline{\theta}, \underline{\beta}, \omega; \underline{x}_i) = \int \left[ \prod_{t=1}^{Ti} Lij(i,t)t(\underline{\theta}, \underline{\beta}, \omega; \underline{x}_{it}, \varepsilon) \right] \mathrm{f}(\varepsilon) d\varepsilon$$

In Equation 4 f($\varepsilon$) is the probability density function (mixing distribution) of the error term and $T_i$ is the length of the sequence, which can vary from one respondent to another. If we assume $\omega\varepsilon \sim N(0,\omega^2)$, then the above likelihood integral may be evaluated numerically.

## Data

The data employed in this example is a panel of young people born in 1975 who were aged 16, and therefore part of the adult survey, in Wave A. The outcome variable was derived from wQFACHI the highest academic qualification. In order to keep the example simple I have restricted the analysis to only three waves of data (BHPS Waves A, C and E) although it is possible to track the majority of these young people through to Wave J. It would also be possible to add other cohorts of 16 year olds as they enter the BHPS adult survey. This would allow us to explore cohort, and possibly period, effects as well as ageing effects.

**Figure 4. wQFACHI**

| |
|---|
| Higher Degree |
| 1st Degree |
| HND, HNC, Teaching |
| A' Level |
| O'Level |
| C.S.E. |
| None of these |

**Figure 5.** Highest Academic Qualification by Wave

| | WAVE A | WAVE C | WAVE E |
|---|---|---|---|
| No qualifications | 41 | 6 | 4 |
| C.S.E. | 31 | 14 | 13 |
| O'Level | 69 | 48 | 25 |
| A' Level | 1 | 45 | 66 |
| n | 142 | 113 | 108 |

Figure 5. indicates that these young people become better academically qualified over three waves (a five year period between age 16 and 21).

The derived outcome variable has the following categories, no qualifications, C.S.E. level qualifications, O'level qualifications and A'level qualifications. In this analysis 'no qualifications' represents a starting point or base category. Movement from one category to another only takes place in one direction. For example a young person may have no qualifications in Wave A, but have O'levels in Wave C and the A'levels in Wave E. Conversely however, a young person will not have A'levels at Wave A, O'levels at Wave C and then no qualifications at Wave E! This means that results will not be invariant to reversing the order of the response categories. Therefore the

continuation ratio model is appropriate for this analysis. Furthermore, the categories of the response variable represent discrete states.

**Figure 6.** The Hierarchical Outcome

| | |
|---|---|
| Cut point 1 | No Qualifications V Some Qualifications |
| Cut point 2 | C.S.E. V O'Levels & A'Levels |
| Cut point 3 | O'Levels V A'Levels |

I have argued that if we consider a hierarchical (or ordinal) outcome that comprises four categories, in a longitudinal set of survey data, we would commonly have a sequence of ordinal outcomes for each respondent. It is, conceptually at least, straightforward to consider modelling these ordinal recurrent events as a series of conditionally independent binary outcomes via binary logistic regression.

In essence the continuation ratio model is undertaking a combined analysis.

The outcome has three cut points.

1. No qualifications versus some qualifications.

2. Given that the young person has some qualifications, C.S.E. level qualifications versus O'level and A'levels.

3. Given that the young person has at least C.S.E. level qualifications, O'levels versus A'levels.

**Figure 7.** Explanatory Variables

| |
|---|
| WAVE<br>BHPS Wave A<br>BHPS Wave C<br>BHPS Wave E<br><br>GENDER<br>Males<br>Females<br><br>PARENTAL QUALIFICATIONS<br>Either parent has a formal academic qualification (Wave A)<br>Neither parent has a formal academic qualification (Wave A)<br><br>PARENTAL SOCIAL CLASS<br>Highest Social Class of either parent (Registrar General) I to V (Wave A)<br><br>HOUSING TENURE<br>Parents own family home (Wave A)<br>Parents do not own family home (Wave A) |

In this example I model the effects of a restricted set of explanatory variables. I have chosen gender, two parental variables and housing tenure. Obviously, in a full analysis more explanatory variables would have been tested.

## Results

**Figure 8.** Continuation Ratio Models

| **Variables** | **Deviance** |
|---|---|
| Cut points | 917 |
| Cut points + Wave | 735 |
| Cut points + Wave + Gender | 734 |
| Cut points + Wave + Parental Qualifications | 723 |
| Cut points + Wave + Parental Social Class | 721 |
| Cut points + Wave + Housing Tenure | 721 |

Gender, parental social class and housing tenure were not significant.

**Figure 9.** Model of Best Fit

| Parameter | Estimate | Standard Error |
|---|---|---|
| | | |
| Cut Point 1 | -2.1694 | 0.33877 |
| Cut Point 2 | -0.60578 | 0.29721 |
| Cut Point 3 | 2.9912 | 0.39864 |
| | | |
| Wave A | - | - |
| Wave C | -2.8210 | 0.34060 |
| Wave E | -3.9465 | 0.41237 |
| | | |
| Either parent has a formal academic qualification | - | - |
| Neither parent has a formal academic qualification | 1.3537 | 0.40281 |
| | | |
| Scale | 1.6535 | 0.24670 |

The odds of a young person having no qualifications rather than some qualifications in Wave A are .11 [i.e. ln (–2.1694)]. Given that they have some qualifications, the odds of a young person having C.S.E. level qualifications rather than O'levels or A'levels is .54 [i.e. ln (–0.60578)]. Given that they have qualifications that are at least of C.S.E. level, the odds of a young person having O'levels rather than A'levels is ∞ [i.e. ln (2.9912)] .

The odds of a young person having no qualifications rather than some qualifications in Wave C are .006 [i.e. ln (–2.194 – 2.8210)]. Given that they have some qualifications, the odds of a young person having C.S.E. level qualifications rather than O'levels or A'levels is .03 [i.e. ln (–0.60578–2.8210)]. Given that they have qualifications that are at least of C.S.E. level, the odds of a young person having O'levels rather than A'levels is 1.18 [i.e. ln (2.9912–2.8210)].

The odds of a young person having no qualifications rather than some qualifications in Wave E are .002 [i.e. ln (–2.194 –3.9465)]. Given that they have some qualifications, the odds of a young person having C.S.E. level qualifications rather than O'levels or A'levels is .01 [i.e. ln (–0.60578–3.9465)]. Given that they have qualifications that are at least of C.S.E. level, the odds of a young person having O'levels rather than A'levels is .38 [i.e. ln (2.9912–3.9465)].

As we would expect neither parent having a formal academic qualification has a negative effect on a young person gaining qualifications. For example the odds of a young person having no qualifications rather than some qualifications in Wave A are .11 [i.e. ln (–2.1694)] if either of their parents have a qualification. However the odds of a young person having no qualifications rather than some

qualifications in Wave A increase to .44 [i.e. ln (–2.1694 + 1.3537)] if their parents do not have a qualification.

The results from the model of best fit also indicate that there is a significant amount of residual heterogeneity (scale = 1.6535).

## Conclusion

Standard log-linear models do not take ordinality into account, thereby potentially they may disregard useful information. In this paper I have demonstrated, through a simple example, that the continuation ratio model is clearly beneficial when the outcome of interest has a substantively important hierarchy because this is taken into account in the modelling. I have also demonstrated that the continuation ratio model is appropriate when results will not be invariant to reversing the order of the response categories.

The ability to fit the model using an existing software package is a great benefit especially to researchers whose interests are substantive rather than methodological. Due to the construction of the SABRE environment the continuation ratio model for recurrent events fits into the general framework of GLIM analysis. Gilchrist (1985) asserts that not only does GLIM bring a wealth of interesting theoretical problems together but it also encourages an ease of data analysis sadly lacking from traditional statistics. Operating under an established framework is not only of benefit to researchers but also to those trying to evaluate research results.

There are a number of other benefits to this approach. In the simplified analysis above the explanatory variables were time-constant. SABRE has the ability to fit both time-constant and time varying explanatory variables. It would also be possible to undertake more advanced analysis that explored the effects of past behaviour on current behaviour (state dependence). In some social science analyses investigating the existence or the effects of cumulative inertia, for example, might be appropriate.

A further benefit of a SABRE is that when considering data on recurrent events there will be individuals for whom there will be zero (or very low) probabilities of change in outcome from one event to the next. These individuals are termed as ‘stayers’. An awareness of the issue of ‘stayers’ is important for technical reasons. A limitation of a parametric modelling approach is that the tail behaviour of the normal distribution is inconsistent with ‘stayers’ and they will tend to be

underestimated (see Spilerman 1972). Recurrent events may be analysed using other software but SABRE is specifically designed to handle stayers and this feature increases SABRE's flexibility in representing residual heterogeneity (Barry, Francis, Davies, and Stott 1998).

**References**


Barry, J., Francis, B., Davies, R.B. and Stott, D. (1998) *SABRE Software for the Analysis of Binary Recurrent Events – A users guide* (Lancaster University: Centre for Applied Statistics).

Berridge, D. (1992) Fitting the continuation ratio model in GLIM4. *Springer-Verlag Lecture Notes in Statistics,* **78,** 1-7.

Berridge, D. and dos Santos, D. (1996a) Fitting A Random Effects Model To Ordinal Recurrent Events Using Existing Software. *Journal of Statistical Computing Simulation,* **55**, 73-86.

Berridge, D. and dos Santos, D. (1996b) Modelling Ordinal Recurrent Events. *Survey and Statistical Computing,* 233-240.

Fienberg, S. and Mason, W. (1979) The Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological Methodology,* 1-67.

Gayle, V. **(**1996) Modelling Tabular Data with an Ordered Outcome. *Sociological Research Online,* 1,
<http://www.socresonline.org.uk/socresonline/1/3/4.html>

Gilchrist, R. (1985) Introduction: GLIM and Generalized Linear Models. *Springer Verlag Lecture Notes in Statistics, **32**,* 1-5.

McCullagh, P. (1980) 'Regression Models for Ordinal Data' (with discussion), *Journal of the Royal Statistical Society B,* **42**, 109-142.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models* (London: Chapman Hall).

Spilerman, S. (1972) Extensions of the Mover-Stayer Model. *American Journal of Sociology,* **78**, 599-626.