THE UNIVERSITY *of York*

CENTRE FOR HEALTH ECONOMICS

THE IDENTIFICATION, ESTIMATION AND INTERPRETATION OF 'HIGHER-LEVEL EFFECTS': THE PROBLEM OF ENDOGENEITY

Nigel Rice and Matthew Sutton

CHE Technical Paper Series 7

THE IDENTIFICATION, ESTIMATION AND INTERPRETATION OF 'HIGHER-LEVEL EFFECTS': THE PROBLEM OF ENDOGENEITY

Nigel Rice[1] *and Matthew Sutton*[1]

Centre for Health Economics,
University of York

Summary

In many applications in the study of health and health care, observations are clustered within higher-level units. Recent interest in multilevel modelling has highlighted the importance of considering these 'higher-level effects'. However, for empirical applications in health economics there are at least two outstanding problems:

i) while the conceptualisation of 'higher-level effects' is appealing in an intuitive sense, their actual interpretation is often rather opaque. Explicit or implicit assumptions must be made about the relationships between individual consumption, group consumption, and individual and group characteristics. Different implications for policy will arise depending on whether endogenous, exogenous or correlated effects are assumed; and

ii) alternative ways of estimating models with 'higher-level effects' imply different assumptions concerning their distribution. If there are relatively few observations clustered within the higher-levels, there is a particularly acute trade-off to be made between controlling for correlation between higher-level effects and individual characteristics (traditionally best treated by fixed effects models) and identifying higher-level effects by preserving degrees of freedom and exploring variances (best achieved through random effects specifications).

These problems are discussed in this paper and illustrated by analysis of 'household-effects' on individual alcohol consumption using data from the *Health Survey in England.*

Keywords:
Group Influences; Interdependent Preferences; Multilevel Models; Fixed and Random Effects Models; Alcohol Consumption.

Acknowledgements:
Data from the Health Survey in England were supplied by the ESRC Data Archive. We are grateful to Paul Contoyannis, Martin Forster, Andrew Jones, Rob Manning, Roshni Mangalore and Tony Scott for helpful comments on previous drafts.

---

[1] *Seniority of authorship is not assigned. Address for correspondence: Centre for Health Economics, University of York, York YO1 5DD. Tel: 01904 433718. Fax: 01904 433644. emails: nr5@york.ac.uk and mas6@york.ac.uk.*

1. Introduction

In studies of health, health-care and health-related behaviour there has been increased interest in 'higher-level effects' - that is, the implications, both theoretical and substantive, of correctly specifying the hierarchical structure of clustered observations (Rice and Leyland, 1996). Consideration of the influences of groups on individuals and vice versa has intuitive appeal. Many empirical studies in health economics could benefit from incorporation of these factors via approaches such as multilevel modelling (Rice and Jones, 1997), or more common procedures outlined in the panel data literature (Hsiao, 1995; Baltagi, 1995; Judge et al, 1980). It has become quite common for authors of applied health-economics papers to propose further data analysis using multi-level models.

Inclusion of such 'higher-level effects' offers many potential advantages in empirical work. The first, most prominent in panel data techniques, is to make allowance for unobservable heterogeneity. Rather than trying to identify factors which influence the potentially endogenous variable but not the response variable, as in instrumental-variable techniques, panel or hierarchical data sets offer the possibility of attributing heterogeneity to unobservable effects, which are often assumed constant across the period of the survey or higher-level group. In this way, allowance for higher-level effects may reduce confounding by controlling for unobservables.

Other statistical and substantive implications of a focus on higher-level effects are appropriate corrections to standard errors and thence tests of significance (via the covariance matrix), exploration of variations across higher-level effects (random coefficient models) and the modelling of more complex higher-level structures (cross-classification of levels).

However, implicit in much of the empirical work on higher-level modelling is that inferences may be drawn to guide policy. For example, what is the implication of the finding that a substantial amount of variation in the outcomes of surgical procedures is clustered at the surgeon or hospital level (Leyland and Boddy, in press)? Alternatively, what are the resource-allocation implications of the finding that, conditional on all observable characteristics, individuals' levels of health tend to be correlated within areas (Duncan, Jones and Moon, 1996; Langford and Bentham, 1996; Congdon, 1995)?

Manski (1993) proposes the following taxonomy for effects which may explain the common observation that individuals belonging to the same group tend to behave similarly:

Endogenous effects - the outcome for an observation is affected by the outcomes for other observations contained within the same higher-level unit. Because there is two-way causation in the model, care must be taken in estimating the equilibrium equation.

Exogenous effects - outcomes for observations in the same higher-level units are correlated because summary characteristics of the higher-level units influence the outcomes for the observations contained within them.

Correlated effects - observations within the same higher-level unit have similar individual characteristics or face the same influences. If these factors are not observed by the analyst they cannot be explicitly modelled. Therefore, there remains some unexplained variation between observations which can be partitioned into variation between higher-level units.

The need to distinguish between these potentially competing explanations is important not only for statistical efficiency but because the three hypotheses have different policy implications. If effects are endogenous, policies aimed at particular individuals will not only have a direct effect on those individuals but also 'knock-on' effects on other individuals. Manski (1993) terms this effect the '*social-multiplier.*'

However, similar consequences do not occur with exogenous or correlated effects. Exogenous effects describe relationships between higher-level covariates and outcomes. They indicate that the response variable (e.g. drinking behaviour) is influenced by contextual characteristics measured at the group-level (e.g. average educational attainment in the household). In the earlier patient-surgeon example, the outcome of an individual intervention may be affected not only by the individual patient's particular characteristics, but also by the case-mix composition of the surgeon's case-load. No social-multiplier effect is indicated.

Once endogenous and exogenous effects have been removed, the existence of any correlated effects merely indicates that there are unobservable characteristics which affect both the response variable and group membership. They seem to have little interpretative value in terms of directing policy. On the other hand, correlated effects are important for statistical inference and, since they may well contain misspecified endogenous or exogenous effects, remain important for exploratory data-analysis.

In this paper we demonstrate the problems associated with the identification and thence interpretation of higher-level effects using data on individuals' drinking behaviour grouped within households. Drinking behaviour is a good example of interdependent outcomes as there has been a great deal of attention related to peer-effects on behaviour and the relationship between population and individual drinking levels (Edwards et al, 1994; Skog, 1980). Household-effects on individual drinking have been analysed in several ways and been found to be significant. Rice et al (1997) used a multilevel modelling approach to detect correlated effects within the household. Alternatively, Sutton and Godfrey (1995) entered mean consumption of alcohol, *by others in the household*, as an independent regressor in their study of individual drinking behaviour. This is a form of externality approach, which has been used in the case of household-effects on smoking behaviour (Jones, 1995). In this context, the limitations of the approach have already been identified:

> "The externality approach is adopted here and it is assumed that other people's smoking has a direct influence on an individual's decision to quit. An appropriate...variable is OTHER-SMOKERS, a binary variable that indicates the presence of other smokers in the individual's household. This variable is not ideal as....there is a potential simultaneity problem as members of the same household should all influence each other." (Jones, 1995, p.9).

Moreover, as Manski (1993) shows, if *ad hoc* formulations of higher-level effects are adopted, at best composite parameters may be estimated, at worst the equations may be misspecified. Also, policy recommendations based on incorrectly specified higher-level effects will be misleading.

To illustrate the potential for competing causes of group-effects, we summarise economic theories of the clustering of consumption choices in social groups in section 2. For compatability with the empirical part of the paper, the hypotheses are grouped according to

Manski's (1993) taxonomy of group-effects. Their application to household-effects of individual drinking behaviour is also discussed.

In section 3 we outline the identification problem and propose a testing strategy which may permit us to discriminate statistically between the different types of 'group-effects'. In section 4, we emphasise the trade-off between efficiency and consistency in parameter estimation when higher-level groups contain few lower level units. We review the competing approaches adopted in fixed-effects models and random effects models (of which multilevel modelling is an extension), and discuss their relative merits in the example presented here. In sections 5 and 6, we implement our proposed testing strategy on drinking data from the *Health Survey for England*. The implications of the results and possible solutions to the identification problem are discussed in a final section.

2. Economic explanations for higher-level effects

Traditionally, explanations for the clustering of behaviour in social groups have been associated with sociology. This is because the framework can be interpreted as assuming that economic-optimising behaviour only arises in economically favourable conditions (Lindbladh et al, 1996). In other words, individuals behave similarly because the decision-process is determined by household-level characteristics. However, Lindbladh et al (1996) also suggest that recent relaxations of the traditional economic model of behaviour-determination may offer explanations of group-effects. Economics has historically focused on perfectly-informed individuals making choices based on knowledge of risks and preferences. Recent adaptations to this model (including bounded rationality; heterogeneous 'models of the world' or expectations about the consequences of different actions; group norms; and habits) may offer alternative explanations of the clustering of consumption choices within households.

Hirshleifer's (1995) review of existing work on Becker's (1991) theory of social influence categorises economic theories of conformity into five groups: sanctions upon deviants; payoff interactions; preference conformity; parallel reasoning; direct communication. Many of the economic theories of conformity express processes which give rise to endogenous effects.

Endogenous effects

The threat of sanctions upon those who deviate from understood norms of behaviour has been proposed to give rise to fairly rigid patterns of behaviour within certain groups (Akerlof, 1980). Within-group externalities can constrain individual consumption or deter deviations from mean consumption. Sugden (1989) emphasises that these may include individual perceptions of social sanction (*psychological* externalities) as well as regulation by actual sanctions. He suggests that the human desire for approval 'is at least as fundamental as the desire for most consumption goods...[because]...we are after all, social animals, biologically fitted to live in groups' (p.95), and that conventions evolve because of the psychological externality of being the focus of ill-will, resentment or anger of one's peers. These theories may be particularly appropriate in the case of alcohol consumption. As Elster (1989) suggests, the process upon which self-help groups like Alcoholics Anonymous are based can be described as 'mutual-sanctioning', in which individuals agree to punish each other if they deviate from agreed norms.

Alternatively, it may be social norms of distribution or fairness in the allocation of the household resources which may affect the level and distribution of alcohol consumption in the household. Alcohol purchasing and consumption decisions may be based on an implicit understanding that individuals should each receive their 'fair-share'. Individual consumption thereby becomes a *pro rata* function of household consumption.

Alternative economic theories of conformity may be particularly relevant in the case of alcohol consumption. The class of theories which Hirshleifer (1995) describes as *payoff interaction* theories are based on the notion that the decisions of one agent directly impacts on the benefits of similar decisions made by others. For many, drinking is a social activity, giving rise to interdependency in the benefits of another 'round'. It is likely that one individual's decision to consume an extra drink can be reinforced (and reinforces) a similar decision by another person. It may be hypothesised that participation in 'rounds' encourages convergence of drinking patterns, since consumption-rates may rise to the speed of the fastest drinker or converge to the mean. In instances where a significant share of drinking takes place at home, or where individuals from the same household socialise together, individual and household consumption will be interdependent.

A further class of economic theories of conformity suggests a more direct reason for convergence in group actions. Jones (1984) suggests that individuals receive utility from doing the same as others. This process is distinct from concerns about deviation from group norms, instead suggesting that interdependency of group preferences is a basic feature of an individual preference set.

The notion of bounded rationality, which highlights the costs of acquiring information on all possible consumption choices together with the associated utility consequences, is another factor underlying a number of theories which express endogenous effects. Duesenberry (1949), for example, proposed that interdependency of consumption patterns arises because individuals do not have knowledge of the full range of commodities. Instead, they only identify a sub-set of what would be superior commodities for them. This sub-set is determined by the familiarity which arises through frequent contact with commodities, for example by observing the consumption patterns of other members of the household.

The role of information is important in other theories of conformity such as *direct communication* (Hirshleifer, 1995). In this case, individuals in certain groups tend to converge on superior commodities when those who identify the optimal choice convey their findings to others in their group. The theory of *informational cascades* also relies on significant costs of information acquisition or processing (Hirshleifer, 1995). It is proposed that the consumption choices of others are interpreted as 'signals' by their peers, and may eventually replace individual calculations of costs and benefits as the criteria on which consumption decisions are made. In our case, individuals may drift into high levels of alcohol consumption, deciding to trust the judgements of others rather than attempt trade-offs of utility-gains with the risks highlighted in public drinking guidelines

There is one final aspect of drinking behaviour which may encourage convergence in drinking patterns. Availability has been found to significantly influence consumption levels (Edwards et al, 1994). Acquisition costs of alcohol have also been proposed as predicting different aspects of drinking patterns, such as the frequency and intensity of drinking bouts (Berggren and Sutton, 1996). It is likely that if heavy drinkers maintain greater supplies of alcohol at home, the reduction in acquisition costs for others in the household may induce additional consumption.

Exogenous effects

Exogenous effects exist when group-aggregated characteristics significantly influence individual behaviour above and beyond the direct effect of those characteristics at the individual-level. Unlike the case of endogenous effects, changes in the level of others' drinking has no direct effect on individual consumption decisions. As mentioned earlier, the sociological notion of constraints on economic-optimising behaviour could bring about just such an effect. An alternative mechanism through which exogenous effects may arise occurs if an individual's relative position in the group affected their decisions. In our example, this may be because the effect of educational attainment has two components; an absolute effect and an effect relative to average attainment in the individual's household.

Consideration of the effects of the time-costs of drinking on individuals' allocation of time also suggests exogenous effects. Jones and Sutton (1997) propose that drinking-styles, such as whether an individual drinks during the day as well as in the evening, will influence the amount drunk. As the consumption of social occasions may be determined by household characteristics, such as whether the household contains small children, exogenous household effects may influence individual consumption.

Correlated effects

Fundamentally, correlated effects imply a different form of association between individual and group behaviour. Correlated effects arise because individuals with common unobservable characteristics that affect drinking behaviour tend to cluster together in groups. In other words, consumption and group membership are both determined by a third factor and changes in both are *caused* by influences unknown to the analyst. For example, a heavy drinking individual may consciously or unconsciously choose a partner that shares similar preferences and characteristics, such as attitudes to unhealthy lifestyles.

Moreover, various components of decision-making which directly affect consumption patterns may cluster in households. It is likely, for example, that individuals who collect in households will not only have common underlying preferences but also common 'models of the world' (Lindbladh et al, 1996). Alternatively, we may envisage that individuals differ in their latent 'booziness' or genetic pre-disposition to alcohol and that this factor is correlated with household membership. Essentially, correlated effects may emerge because of a variety of factors which tend to cluster in households, affect drinking, and are not observed in household surveys of health-related behaviour.

A thought-experiment to test whether the hypothesised process is a form of correlated effect or an alternative type of higher-level effect is whether a given individual will consume the same amount regardless of the household in which he/she is placed. If the same amount would be consumed regardless of household membership then a correlated effect may be assumed to exist.

3. Identification and an estimation strategy

Manski (1993) analyses the problem of identifying the influence of higher-level units on individual responses. We summarise his model in this section before proposing an estimation strategy in light of the main findings.

The response variable, y, is assumed to be a function of three vectors of characteristics, (x, z, u), in which x denotes an individual's reference group, and z and u are attributes that directly affect y. The values of z are observed, but the values of the characteristics u remain unobserved to the analyst. Assuming that the unobservables can be expressed as a function of the group-identifiers, the following linear form for the mean regression of y on the observable characteristics is derived:

$y = \alpha + \beta E(y|x) + E(z|x)'\gamma + z'\eta + u$,
where $E(u|x,z) = x'\delta$ and therefore

$$E(y|x,z) = \alpha + \beta E(y|x) + E(z|x)'\gamma + x'\delta + z'\eta \quad (1)$$

In the example we consider, $E(y|x,z)$ is the expectation of individual alcohol consumption y, conditioned on household membership x and individual characteristics z. $E(y|x)$ represents an appropriate measure of household alcohol consumption, often estimated as a mean across all members of the household. $\beta \neq 0$ implies that there exists an endogenous effect, since individual consumption varies with $E(y|x)$, average consumption in the household. The direct effect of individual socio-economic characteristics z on individual drinking consumption are provided through the estimate of η. $E(z|x)$ is the mean (or some other appropriate measure) of the individual characteristics across the household. If $\gamma \neq 0$, the model expresses an exogenous effect, since individual drinking levels vary with mean household characteristics. The term $x'\delta$ denotes individual drinking consumption varying with household membership. Therefore, if $\delta \neq 0$, the model expresses correlated effects, such that household members tend to behave similarly because they have similar individual unobserved characteristics u.

To address the problem of the endogeneity of $E(y|x)$ in equation (1), Manski (1993) integrates through by z to derive a "social equilibrium" equation (integrating over households with respect to z):

$$\mathrm{E}(y|x) = \alpha + \beta \mathrm{E}(y|x) + \mathrm{E}(z|x)'\gamma + x'\delta + \mathrm{E}(z|x)'\eta \quad (2)$$

Provided that $\beta \neq 1$, inserting (2) into (1) leads to the following reduced-form model:

$$\mathrm{E}(y|x,z) = \alpha/(1-\beta) + \mathrm{E}(z|x)'[(\gamma + \beta\eta)/(1-\beta)] + x'\delta/(1-\beta) + z'\eta \quad (3)$$

which expresses the response variable as a function of mean characteristics at the group-level, group identifiers and individual characteristics. The composite parameters $\alpha/(1-\beta)$, $(\gamma+\beta\eta)/(1-\beta)$, $\delta/(1-\beta)$ and η are identified if the regressors $[1, E(z|x), x, z]$ are linearly independent in the population. Although evidence that $(\gamma+\beta\eta)/(1-\beta)$ is non-zero indicates that some social effect is present, it is not possible to distinguish between the two social effects (β and δ).

Manski (1993) proposes that pure endogenous or exogenous effects can only be estimated if it is assumed that correlated effects do not exist and that the higher-level effects are either all endogenous or all exogenous. In this case a two-stage estimation procedure can be used to produce sample inferences on (γ, η). For a pure exogenous-effects model, a first stage is to use sample data on (z, x) to estimate $E(z|x)$ non-parametrically. In the second stage, the parameters (γ, η) are estimated by the least squares fit of y to $[1, E_N(z|x), z]$ where $E_N(z|x)$ is the first-stage estimate of $E(z|x)$. The pure endogenous effects model can be estimated analogously.

In the reduced form specification, equation (3), only the composite parameters $(\gamma + \beta\eta)/(1-\beta)$ and $\delta/(1-\beta)$ on the higher-level effects are identified. The Appendix contains details of the sets of possible realisations for θ_1 and θ_2, representing the composite parameters $(\gamma + \beta\eta)/(1-\beta)$ and $\delta/(1-\beta)$ respectively, which prove to be extremely unhelpful in identifying the endogenous, exogenous and correlated effects. However, if we are prepared to make the following two assumptions:

a) some of the individual characteristics have significant effects on the response variable (i.e. $\eta \neq 0$) and
b) the magnitudes of the exogenous effects are not offset by the magnitudes of the corresponding individual characteristics, weighted by the magnitude of the endogenous effect (i.e. $\gamma \neq -\beta\eta$),

this information can be simplified to the following four possible scenarios based on whether θ_1 and/or θ_2 are non-zero:

	Endogenous effect	*Exogenous effect*	*Correlated effect*
Scenario i). $\theta_1 = 0,\ \theta_2 = 0$	No	No	No
Scenario ii). $\theta_1 \neq 0,\ \theta_2 = 0$	Endogenous and/or Exogenous effects		No
Scenario iii). $\theta_1 = 0,\ \theta_2 \neq 0$	No	No	Yes
Scenario iv). $\theta_1 \neq 0,\ \theta_2 \neq 0$	Endogenous and/or Exogenous effects		Yes

Table 1: Interpretations of the composite parameters in the reduced form

In summary: scenario i) can be taken to imply no higher-level effects; scenario ii) implies exogenous and/or endogenous effects but no correlated effects; scenario iii) suggests only correlated effects are present; and scenario iv) suggests that exogenous and/or endogenous effects exist and there are additionally correlated effects. Under scenarios i) and iii), the parameters can be taken to be structural parameters and no further estimation is required. Under scenario ii), it may be possible to estimate pure endogenous and pure exogenous effects models and test between them using a non-nested test. Under scenario iv), it is not clear how estimation can proceed beyond the reduced-form model.

4. Efficiency and consistency in parameter estimation

Empirical applications in health and health care tend to view higher-level effects as being of either intrinsic importance in themselves, or as nuisance parameters to be conditioned upon whilst estimating the impact of other covariates of interest. For example, in studies concerned with efficiency measurement, higher-level parameter estimates are fundamental to the objectives of the study (Greene, 1993). In other applications, interest may lie solely in the covariate estimates, once the effects of higher-level groupings have been 'removed'.

Econometric techniques available to estimate models containing higher-level effects have been discussed extensively in the panel data literature (Hsaio, 1986; Baltagi, 1995; Judge et al, 1980). Such models are principally concerned with estimation when the sample design consists of multiple observations most often measured on individuals. However, the principles of the techniques are applicable to other hierarchical designs such as measurements made on individuals clustered within defined groupings. The methods available to estimate such models can be conveniently categorised into fixed and random effects models. Whilst the former are essentially concerned with *point* estimation of higher-level and individual effects, the latter place much greater emphasis on the *distributional* parameters of higher-level effects.

The advent of multilevel models, exploiting further the use and applications of random effects models, may have particular additional advantages over more traditional techniques, such as extending the possible number and complexity of levels within the hierarchy of interest (Goldstein, 1995; Longford, 1993; Bryk and Raudenbush, 1992; Rice and Jones, 1996). However, the choice of model specification adopted in empirical applications warrants careful consideration and may lead to vastly different estimated effects. Criteria for choosing between different specifications include both theoretical and practical considerations and we discuss some of these below.

For the particular example presented in this paper, we have many higher-level groups represented by household membership. It may be the case that empirical specifications contain only a limited number of higher-level groups and that the choice of specification is limited to fixed effects only. However, we assume this not to be the case and that multiple higher-level groupings are present, such that consideration of fixed and random effects specification is warranted. The literature on panel data techniques places much emphasis on the relative merits of treating higher-level units as random or fixed effects (Judge et al, 1980; Hsiao, 1986). For example, if we exclude endogenous and exogenous higher-level effects for simplicity, model (1) may be rewritten in forms familiar to the panel data literature:

$$y_{ij} = \beta' z_{ij} + u_j + e_{ij} \qquad (4)$$
$$y_{ij} = \beta' \mathrm{z}_{ij} + \delta \mathrm{x}_j + \mathrm{e}_{ij} \qquad (5)$$

where; $i = 1,\ldots,T$, and $j = 1,\ldots,M$. In model (4), u_j represent the set of j higher-level effects modelled as random terms and correspond to x_j in equations (3) and (5). Typically, i would represent longitudinal measurements and j, individuals. We can extend this classification to include the analysis of any hierarchy: for our example, i represents individuals and j households. In model (4) the household specific effects are specified as random effects and in (5) they are specified as a set of j fixed effects (j-1, if an intercept term is specified).

An important consideration in choosing between fixed and random specifications is when an explanatory variable is correlated with the higher-level effects. In such circumstances, random- and fixed-effects approaches may lead to vastly different estimates (Hsiao, 1983; Hausman, 1985). The difference in the two approaches can be viewed as one of whether the conditional distribution of u_j, given z, is equal to the unconditional distribution of u_j. In the case where z is correlated with u the two approaches will yield different estimators, since $E(u_j)$ will not be constant but equal to some function of z (i.e. $E(y_{ij}|z_{ij},x_j) \neq E(y_{ij}|z_{ij})$).

The situation can be extended to multilevel models (Rice and Jones, 1996). When the random effects u_j and explanatory variables z_{ij} are correlated, and group sizes are relatively small, the iterative generalised least squares estimator (Goldstein (1986)) for the parameters β will be inconsistent as the number of higher-level groups becomes large. However, treating the effects u_j as fixed and applying a dummy variable estimator leads to consistent estimates. When group sizes are large, the two estimators are equivalent (Blundell and Windmeijer, 1996).

In the situation where an explanatory variable is correlated with the higher-level effects, and the sole concern of the analyst is the estimation of the explanatory variable parameters (and their properties), a fixed-effects specification has advantages.[2] However, in the multilevel formulation, intrinsic interest lies in the estimation of higher-level variances after conditioning on the set of explanatory variables. Higher-level effects are not viewed as nuisance parameters, but are of central importance. In situations where explanatory variables are correlated with higher-level effects, a random-effects specification is at the cost of inconsistent parameter estimation. However, it has benefits in terms of investigating higher-level effects and greater efficiency, since fewer parameters are estimated (σ^2_w for random-effects, j-1 set of dummies for fixed-effects). This latter consideration is particularly important in situations such as the example used here, where there are many higher-level units, with few level 1 observations in each. In such cases, higher-level effects may be poorly estimated and this is, in part, the motivation behind using a random specification with shrunken higher-level estimates (Rice and Jones, 1996).

5. Analysis

Data from the *Health Survey for England* (HSE; White et al, 1993) are used to implement the proposed estimation strategy. Data collection was performed throughout 1993 and on into early 1994, and consisted of approximately 17,000 interviews with adults (aged 16 or over) living in 9,700 households in England. The sample was distributed relatively evenly across the 14 English Regional Health Authorities, and was obtained by sampling households from two to three electoral wards of residence in each Authorities' area. The survey is relatively unusual in seeking responses from all adults in each household[3], providing an opportunity to explore effects within households. The 1993 survey focused on cardiovascular disease and associated risk factors, including alcohol consumption, as well as general health and various long-standing illnesses.

---

[2] At the time of writing, one of the authors is working on an estimation procedure to produce consistent estimates in the presence of correlated predictors and random effects (Rice, Jones and Goldstein, 1997).

[3] Whilst the General Household Survey used by Sutton and Godfrey (1995) contains self-reported information on all individuals within a household, health-specific interview surveys, such as the *Health and Lifestyle Survey*, rarely do.

Empirical distributions of alcohol consumption tend to be highly skewed with long tails towards high consumption levels. This relationship has been found to hold even in very homogeneous populations where one may expect more symmetrical distributions (Edwards et al, 1994). The HSE data also display strong skewness and transformation to a more symmetrical distribution was sought using a logarithmic transformation. Due to some individuals reporting alcohol consumption between zero and unit, we added a constant value of unity to each observation before taking the logarithmic value.

Excluding data on households from whom incomplete alcohol consumption data were obtained, and on individuals where missing entries were recorded on the covariates required for the analysis, results in data for 15,429 individuals clustered within 8,737 households. Table 2 displays summary statistics for the variables used. As specified in equations (1) and (3), for each individual level variable the corresponding household-level mean was also calculated.

6. Results

The estimation results for the reduced-form equation (3) are shown in Table 3. A semi-log formulation is adopted over a double-log specification on the basis of a superior RESET test. The equation has been estimated using OLS and Fixed- and Random-Effects Models in LIMDEP (Greene, 1991) and a Multilevel Model using Mln (Rasbash et al, 1995).[4]

RESET test-statistics are provided for each of the models. There is little evidence of misspecification in any of the models, although the Fixed-Effects estimator performs least well (RESET t-statistic: $0.01 < P < 0.05$). Further, plots of standardised residuals against Normal scores showed no serious deviations from the assumption of Normality in the multilevel model. A Hausman test (Hausman, 1978) of Fixed versus Random-Effects specification indicates that the Random-Effects Model is consistent ($\chi^2(36) = 0.0001$; $p > 1.0$). The Random-Effects Model and the Multilevel Model estimation procedures are more efficient than the Fixed-Effects estimates, and this can be seen by comparing the associated parameter standard errors with those of the fixed effects model.

At least one of the household-effects variables is significant in each of the different model specifications. A simultaneous test of significance of all household level variables in the multilevel model constructed by defining a (18×37) contrast matrix C, and puting $f = C\beta$ (β being a matrix of estimated coefficients) and testing $H_0 : f = k$: $k = \{0\}$ resulted in $\chi^2_{18} = 121.9$, $p < 0.001$. This suggests that exogenous and/or endogenous effects are present.

However, as the random-effects parameters in the Multilevel Model indicate, there remains substantial correlations between individuals grouped in the same household. The LM test-statistic is highly significant in the Random-Effects Model ($\chi^2(1) = 1396$; $p < 0.001$). As a result, there is no justification for moving onto either a pure endogenous or pure exogenous model (which requires an assumption of no correlated effects), and we are left with estimation of the composite parameters from the reduced-form equation. It is not possible to identify

[4] We note in passing, as a reflection of the computing costs of the different models, that the LIMDEP estimations had to be undertaken using a mainframe version of LIMDEP with an expanded maximum number of groups. The PC version of LIMDEP permits only 2,000 groups, limiting empirical work on hierarchical data-sets quite severely.

whether group-effects are produced by either an exogenous or endogenous process or, indeed, a combination of the two.

7. Discussion

In this paper we have emphasised the difficulties in both the estimation and interpretation of higher-level effects. If individuals influence and are influenced by group behaviour, the disentanglement of the separate effects that group-specific characteristics have on individual outcomes is bedevilled by problems of identification. If the intention of empirical analyses of individual behaviour and health outcomes is, either explicitly or implicitly, to inform evaluation of policy instruments, then consideration of the issues surrounding parameter identification is warranted. In many empirical analyses, only composite parameters are identified and cannot be interpreted as indicating endogenous, exogenous or correlated effects.

We have demonstrated the problems of identification using the example of household-effects on individual alcohol consumption. Although, higher-level effects are observed to be statistically significant, we do not know whether the parameter estimates represent endogenous or exogenous effects, since we are unable to avoid correlated effects caused by clustering of individuals within households on unobservable characteristics. Unless the effects of unobservables can be removed, non-nested tests of pure endogenous models against pure exogenous models cannot be undertaken. In policy terms, without this investigation, it is not known whether a '*household multiplier*' effect of interventions on individual drinking can be expected.

An alternative solution to the identification problem would be to instrument for the endogenous higher-level effect. This may present a convenient solution, but given that the effect of concern in our application is at the household level, it is unlikely that a suitable set of instruments (which satisfied the requirement that they are correlated with the independent variable but not the dependent variable) could be obtained. Further, the use of instrumental variables may be very inefficient.

A number of papers suggest moving away from a simultaneous specification of interdependent preferences (Pollak, 1976; Alessie and Kapteyn, 1991). In these models, endogenous taste changes are introduced through lagged consumption in the population or peer-group. Pollak (1976) argues that using lagged consumption ensures a smooth adjustment process and does not assume instantaneous adjustment. Many of the information-based theories would suggest time-lags in the process by which decisions converge. However, specific time-periods are not proposed and, given that drinking surveys ask respondents to recall consumption over periods up to one year (Sindelar, 1993), simultaneous specifications may remain necessary.

A final potential solution to the investigation of higher-level effects is to consider the group-formation process. With additional information on how households form, disband and re-form, for example, it may be possible to distinguish between exogenous, endogenous and correlated effects. However, this is unlikely to be available in cross-sectional household surveys.

The potential problems of misspecifying and misinterpreting models with group-related variables are large. For example, inclusion of household income in studies of individual behaviour is common. If the estimated equation is interpreted as a reduced-form equation for a model which may contain endogenous effects, then the interpretation of the parameter estimate on household income is unclear. It may be that, rather than representing the effect of group income on behaviour, a significant parameter may simply indicate interdependency of behaviour within households. In the absence of additional data or analysis of group-formation, the problems of identification and interpretation are persistent, and caution must be taken in the interpretation of coefficients on group characteristics and identifiers.

References

Akerlof, G. A theory of social custom, of which unemployment may be one consequence. *Quarterly Journal of Economics* 1980, **90**: 599-617.

Alessie, R. Kapteyn, A. Habit formation, interdependent preferences and demographic effects in the almost ideal demand system. *Economic Journal* 1991; **101**: 404-19.

Baltagi, BH. *Econometric Analysis of Panel Data.* New York: John Wiley & Sons, 1995.

Becker, G. A note on restaurant pricing and other examples of social influences on price. *Journal of Political Economy* 1981, **99**: 1109-16.

Berggren, F. Sutton, M. *Are Frequency and Intensity of Participation Utility-Bearing Aspects of Consumption? An Analysis of Drinking Behaviour.* Centre for Health Economics Technical Paper 1. York: Centre for Health Economics, 1996.

Blundell, R. Windmeijer, F. Correlated cluster effects and simultaneity in multilevel models. *Health Economics* 1997; **6(4)**: in press.

Bryk, AS. and Raudenbush, SW. *Hierarchical Linear Models.* Newbury Park: Sage, 1992.

Congdon, P. The impact of area context on long term illness and premature mortality: an illustration of multilevel analysis. *Regional Studies* 1995; **29**: 327-44.

Duesenberry, JS. *Income, Saving, and the Theory of Consumer Behaviour*, Cambridge, Mass. 1949.

Duncan, C. Jones, K. Moon, G. Health-related behaviour in context: A multilevel modelling approach. *Social Science and Medicine* 1996; **42**: 817-30.

Edwards, G. et al. *Alcohol Policy and the Public Good.* Oxford: Oxford University Press, 1994.

Elster, J. Social norms and economic theory. *Journal of Economic Perspectives* 1989; **3**: 99-117.

Goldstein, H. Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika* 1986; **73**: 43-56.

Goldstein, H. *Multilevel Statistical Models.* London: Edward Arnold, 1995.

Greene, WH. *LIMDEP Version 6.* Econometric Software Ltd, 1991.

Greene, WH. The econometric approach to efficiency analysis. In: Fried, HO., Lovell, CAK., and Schmidt, SS. (eds.) *The Measurement of Productive Efficiency.* New York: Oxford University Press, 1993.

Hausman, JA. Specification tests in econometrics. *Econometrica* 1978; **46**: 1251-71.

Hausman, JA. The econometrics of nonlinear budget sets. *Econometrica* 1985; **53**: 1255-82.

Hirshleifer, D. The blind leading the blind. Social influence, fads, and informational cascades. In: Tomassi, M. and Ierulli, K. (eds.) *The New Economics of Human Behaviour.* Cambridge:Cambridge University Press, 1995.

Hsiao, C. *Analysis of Panel Data.* Cambridge: Cambridge University Press, 1986.

Jones, AM. *A Microeconometric Analysis of Smoking in the UK Health and Lifestyle Survey.* Centre for Health Economics Discussion Paper 139. York: Centre for Health Economics, 1995.

Jones, AM. Sutton, M. *The Influence of Drinking Styles on Alcohol Consumption Levels.* Mimeo: University of York, 1997.

Jones, SRG. *The Economics of Conformism.* Oxford: Basil Blackwell, 1984.

Judge, GG. Griffiths, WE. Hill, RC. Lee, TC. *The Theory and Practice of Econometrics.* New York: Wiley, 1980.

Langford, I.H. Bentham, G. Regional variations in mortality rates in England and Wales: An analysis using multilevel modelling. *Social Science and Medicine* 1996; **42**: 897-908.

Leyland, AH and Boddy, FA. Measuring performance in hospital care: the example of length of stay in gynaecology. European Journal of Public Health, in press.

Lindbladh, E. Lyttkens, CH. Hanson, BS. Ostergren, PO. Isacsson, SO. Lindgren, B. An economic and sociological interpretation of social differences in health related behaviour. An encounter as a guide to social epidemiology. *Social Science and Medicine,* 1996; **43**: 1817-27.

Longford, NT. *Random Coefficient Models.* Oxford: Clarendon Press, 1993

Manski, CF. Identification of endogenous social effects: The reflection problem. *Review of Economic Studies* 1993; **60**: 531-42.

Pollak, RA. Interdependent preferences. *American Economic Review* 1976; **66**: 309-20.

Rasbash, J. Yang, M. Woodhouse, G. Goldstein, H. *Mln: command reference guide*. 1995 London: Institute of Education.

Rice, N. Carr-Hill, CA. Dixon, P., Sutton, M. The influence of households on drinking behaviour: A multilevel analysis. *Social Science and Medicine - In press.*

Rice, N. Jones, AM. Multilevel models and health economics. *Health Economics* 1997; **6**: - *In press.*

Rice, N. Jones, AM. Goldstein, H. Multilevel models where the random effects are correlated with the fixed predictors: a conditioned iterative generalised least squares estimator (CIGLS). *mimeo: Centre for Health Economics, University of York.*

Rice, N. Leyland, AH. Multilevel Models: applications to health data. *Journal of Health Services Research and Policy* 1996; **1**: 154-64.

Sindelar, JL. Measurement issues in alcohol survey data, in: Hilton, ME. Bloss, G. *Economics and Prevention of Alcohol-Related Problems.* US Department of Health and Human Services, 1993: 201-228.

Skog, O-J. Social interaction and the distribution of alcohol consumption. *Journal of Drug Issues* 1980, **10**: 71-92.

Sugden, R. Spontaneous Order. *Journal of Economic Perspectives* 1989; **3**: 85-97.

Sutton, M. Godfrey, C. A grouped data regression approach to modelling the economic and social determinants of individual drinking behaviour. *Health Economics* 1995, **4**: 237-47.

White, A. Nicholas, G. Foster, K. Browne, F. and Carey, S. *Health Survey in England 1991*: A survey carried out by the Social Survey Division of the OPCS on behalf of the Department of Health, London: HMSO 1993.

Appendix I

Note: $\theta_1 = (\gamma + \beta\eta)/(1-\beta)$, **and** $\theta_2 = \delta/(1-\beta)$

Scenario i). $\theta_1 = 0,\ \theta_2 = 0$:

$\theta_1 = 0$ implies either:

a. $\gamma = -1 \times \beta\eta$ and $\gamma, \beta, \eta \neq 0$
b. $\gamma = \beta\eta = 0 \Rightarrow \gamma = \beta = 0$, or $\gamma = \eta = 0$, or $\gamma = \beta = \eta = 0$

Information about η can be obtained through the associated individual level estimates and hence the hypothesis that $\eta = 0$ can be tested directly:

if $\eta = 0$, then: either, $\beta = 0$ and $\gamma = 0$, $\Rightarrow$ no endogenous or exogenous effects
or, $\beta \neq 0$, and $\gamma = 0$ $\Rightarrow$ endogenous but no exogenous effects

if $\eta \neq 0$, then: either, $\gamma = -1 \times \beta\eta$, $\Rightarrow$ both exogenous and endogenous effects
or, $\gamma = \beta = 0$, $\Rightarrow$ neither exogenous nor endogenous effects

$\theta_2 = 0$ implies:

a. $\delta = 0$: no correlated effects are present.

Scenario ii). $\theta_1 \neq 0,\ \theta_2 = 0$:

$\theta_1 \neq 0$ implies either:

a. $\gamma \neq -1 \times \beta\eta$ and $\gamma, \beta, \eta \neq 0$
b. $\gamma, \beta \neq 0, \eta = 0$
c. $\gamma, \eta \neq 0, \beta = 0$
d. $\beta, \eta \neq 0, \gamma = 0$
e. $\gamma \neq 0, \beta = \eta = 0$

Information about η can be obtained through the associated individual level estimates and hence the hypothesis that $\eta = 0$ can be tested directly:

if $\eta = 0$, then: either $\beta \neq 0$, and $\gamma \neq 0$ $\Rightarrow$ endogenous and exogenous effects
or, $\beta = 0$, and $\gamma \neq 0$ $\Rightarrow$ exogenous effect only

if $\eta \neq 0$, then either $\beta \neq 0$, and $\gamma = 0$ $\Rightarrow$ endogenous effect only
or, $\gamma \neq 0$, and $\beta = 0$ $\Rightarrow$ exogenous effect only
or, $\beta \neq 0$, and $\gamma \neq 0$ $\Rightarrow$ endogenous and exogenous effects

$\theta_2 = 0$ implies:

a. $\delta = 0$: no correlated effects are present.

Scenario iii). $\theta_1 = 0,\ \theta_2 \neq 0$:

$\theta_1 = 0$ implies either:

a. $\gamma = -1 \times \beta\eta$ and $\gamma, \beta, \eta \neq 0$
b. $\gamma = \beta\eta = 0 \Rightarrow \gamma = \beta = 0$, or $\gamma = \eta = 0$, or $\gamma = \beta = \eta = 0$

Information about η can be obtained through the associated individual level estimates and hence the hypothesis that $\eta = 0$ can be tested directly:

if $\eta = 0$, then: either, $\beta = 0$ and $\gamma = 0$, $\Rightarrow$ no endogenous or exogenous effects
or, $\beta \neq 0$, and $\gamma = 0$ $\Rightarrow$ endogenous but no exogenous effects

if $\eta \neq 0$, then: either, $\gamma = -1 \times \beta\eta$, $\Rightarrow$ both exogenous and endogenous effects
or, $\gamma = \beta = 0$, $\Rightarrow$ neither exogenous nor endogenous effects

$\theta_2 \neq 0$ implies:

a. $\delta \neq 0$: correlated effects are present.

Scenario iv). $\theta_1 \neq 0,\ \theta_2 \neq 0$:

$\theta_1 \neq 0$ implies either:

a. $\gamma \neq -1 \times \beta\eta$ and $\gamma, \beta, \eta \neq 0$
b. $\gamma, \beta \neq 0, \eta = 0$
c. $\gamma, \eta \neq 0, \beta = 0$
d. $\beta, \eta \neq 0, \gamma = 0$
e. $\gamma \neq 0, \beta = \eta = 0$

Information about η can be obtained through the associated individual level estimates and hence the hypothesis that $\eta = 0$ can be tested directly:

if $\eta = 0$, then: either $\beta \neq 0$, and $\gamma \neq 0 \Rightarrow$ endogenous and exogenous effects
or, $\beta = 0$, and $\gamma \neq 0 \Rightarrow$ exogenous effect only

if $\eta \neq 0$, then either $\beta \neq 0$, and $\gamma = 0$ $\Rightarrow$ endogenous effect only
or, $\gamma \neq 0$, and $\beta = 0 \Rightarrow$ exogenous effect only
or, $\beta \neq 0$, and $\gamma \neq 0 \Rightarrow$ endogenous and exogenous effects

$\theta_2 \neq 0$ implies:

a. $\delta \neq 0$: correlated effects are present

Table 2: Descriptive Statistics

Variable	Mean	SD
Log Alcohol Consumption	1.76	1.278
Age	0.457	0.183
Age Squared	0.242	0.181
Male	0.470	0.499
Single	0.343	0.475
Economic Status:		
Work - Part Time	0.141	0.348
Unemployed	0.057	0.232
Inactive	0.369	0.483
Social Class:		
SC - I	0.049	0.216
SC - IIIN	0.237	0.425
SC - IIIM	0.199	0.399
SC - IV	0.154	0.361
SC - V	0.063	0.242
SC - Armed Forces	0.004	0.065
SC - Student	0.046	0.210
Qualifications:		
Degree	0.098	0.298
A Level	0.196	0.397
GCSE	0.330	0.470
Foreign	0.013	0.115

Table 3:Reduced Form Results

n = 15429	OLS		Fixed Effects Model		Random Effects Model		Multilevel Model	
Variables	Parameter Estimate	Standard Error	Parameter Estimate	Standard Error	Parameter Estimate	Standard Error	Parameter Estimate	Standard Error
<u>*Fixed Part:*</u>								
Constant	1.469•	0.089	-	-	1.424•	0.117	1.446	0.119
Age	1.267•	0.330	1.398•	0.601	1.397•	0.601	1.398•	0.583
Age Squared	-2.127•	0.331	-2.755•	0.635	-2.754•	0.635	-2.755•	0.615
Male	0.761•	0.022	0.777•	0.023	0.777•	0.023	0.777•	0.022
Single	0.060•	0.022	0.057	0.060	0.057	0.060	0.057	0.058
Economic Status[a]*:*								
Work - Part Time	-0.138•	0.032	-0.099•	0.036	-0.099•	0.036	-0.099•	0.035
Unemployed	-0.181•	0.042	-0.021	0.053	-0.021	0.053	-0.021	0.051
Inactive	-0.310•	0.029	-0.106•	0.037	-0.106•	0.038	-0.106•	0.037
Social Class[b]*:*								
SC - I	0.027	0.049	0.050	0.058	0.050	0.058	0.050	0.057
SC - IIIN	-0.069•	0.029	-0.071•	0.036	-0.071•	0.036	-0.071•	0.035
SC - IIIM	-0.078•	0.031	0.034	0.039	0.034	0.039	0.034	0.038
SC - IV	-0.181•	0.034	-0.022	0.043	-0.022	0.043	-0.022	0.042
SC - V	-0.223•	0.045	-0.059	0.058	-0.059	0.058	-0.059	0.056
SC - Armed Forces	0.375•	0.144	0.007	0.169	0.007	0.169	0.007	0.164
SC - Student	-0.222•	0.058	-0.334•	0.071	-0.334•	0.071	-0.334•	0.069
Qualifications[c]*:*								
Degree	0.121•	0.042	-0.059	0.059	-0.059	0.059	-0.059	0.057
A-Level	0.155•	0.031	-0.001	0.040	-0.001	0.040	-0.001	0.039
GCSE	0.097•	0.025	0.004	0.032	0.004	0.032	0.004	0.031
Foreign	-0.228•	0.083	-0.131	0.106	-0.131	0.106	-0.132	0.103
Household Effects:								
Average Age			-1.510	1.431	-0.085	0.738	-0.096	0.729
Average Age Squared			2.141	1.290	0.729	0.759	0.730	0.749
Proportion Male			0.119	0.090	0.064	0.045	0.055	0.046
Proportion Single			0.080	0.326	0.022	0.065	0.023	0.063
Proportion Part - Time			-0.102	0.198	-0.095	0.062	-0.097	0.062
Proportion Unemployed			-0.568•	0.177	-0.271•	0.079	-0.256•	0.079
Proportion Inactive			-0.420•	0.145	-0.302•	0.055	-0.298•	0.054
Proportion SC - I			0.383	0.200	0.002	0.092	-0.024	0.092
Proportion SC - IIIN			0.052	0.124	0.002	0.055	0.001	0.055
Proportion SC - IIIM			0.243	0.130	-0.119•	0.058	-0.143•	0.058
Proportion SC - IV			0.083	0.143	-0.209•	0.064	-0.224•	0.064
Proportion SC - V			-0.192	0.179	-0.222•	0.085	-0.221•	0.085
Proportion A.Forces			0.185	1.128	0.754•	0.276	-0.758•	0.275
Proportion Student			0.370	0.524	0.095	0.117	0.088	0.116
Proportion Degree			0.477•	0.192	0.212•	0.082	0.200•	0.082
Proportion A-Level			0.180	0.149	0.170•	0.059	0.164•	0.059
Proportion GCSE			0.298•	0.112	0.118•	0.048	0.110•	0.048
Proportion Foreign			0.328	0.344	-0.144	0.156	-0.167	0.155
<u>*Random Effects:*</u>								
Level 1:								
Individual Variance	1.028		0.807		0.811		0.759•	0.013
Level 2:								
Household Variance					0.527		0.573•	0.017
Intra-household correlation					0.394		0.430	
Ramsey RESET: t-value	-1.355		2.000		0.300		0.597	
R-squared (adjusted)	0.179		0.506		0.184		-	
Log-Likelihood	-24149.6		-14412.7		-		-23347.6	

Note: a. Dummy variables contrasted against baseline of Full Time Employment
b. Dummy variables contrasted against baseline of SC - II
c. Dummy variables contrasted against baseline of No Formal Qualifications
• Indicates significance at $P< 0.05$