Exploring the influence of others: Modelling social connections in social surveys

Dave Griffiths (University of Stirling)

Paul S. Lambert (University of Stirling)

Vernon Gayle (University of Edinburgh)

11th June 2015

Please do not cite or reference without prior permission of the authors (contact david.griffiths@stir.ac.uk).

This paper is funded by the 'Is Britain Pulling Apart?' project, led by Prof. Paul S. Lambert, as part of the ESRC Secondary Data Analysis Initiative (SDAI) Phase 1.

Abstract

The influence of others on our behaviour and outcomes is a central tenant of sociological thought and, therefore, a sociological approach should consider the potential for respondents in large-scale surveys containing data on multiple household members. Analysts typically either ignore such structure amongst respondents, or construct household-level units in multilevel models. We examine other clustering units within the household and suggest that using an 'Alter Explanatory Variable' (AEV) is a potential method of capturing the effect of social contacts whilst producing easily interpretable results.

Introduction

Many of the landmark studies in British sociology concerned the role of personal networks, such as Bott's (1957) *Family and Social Network* and Young and Wilmot's (1957) *Family and Kinship in East London*. These studies left little doubt that the behaviour of others with whom social contact is held, both within and outside the household, strongly shaped the experience of individuals. In the intervening period, a rich body of high quality nationally representative social survey datasets have been developed and made available for research purposes. However, the extent to which researchers exploiting survey datasets have incorporated appropriate information about other people in analyses of individual level processes has been surprisingly partial, and conventional social survey analysis is sometimes criticised for having an individualist focus upon respondents, to the neglect of their relationships and social contacts (cf. Gilbert 2008). In some instances, however, this oversight involves the lack of use of relevant data, rather than its unavailability. In this paper, we discuss a range of methods by which it is possible to incorporate more information on the 'influence of others' within the household in statistical analyses, including some examples that can lead to important changes of interpretation and new insights about interdependencies in social processes.

The ‘influence of others’ clearly involves both people who may share the same household as a respondent, and other social contacts who do not. Many social surveys have asked respondents to describe certain characteristics of their closest friends (e.g. Halsey et al. 1978; Blackburn et al. 1980; University of Essex 2010, 2013), or deliberately collected information of the social ties between respondents within specific populations studied (e.g. West and Sweeting 1995; Harris et al. 2009; Christakis and Fowler 2009). Such strategies are underpinned by the assumption that individuals’ behaviours and outcomes are often influence by the people around them, for instance through cultural consumption patterns being influenced by partner’s interests (Upright 2004) or through the contagion of one person’s bad mood flowing through a household (Christakis and Fowler 2009). Our focus below is on a body of data which frequently collects data from socially connected individuals, but frequently overlooks this design in subsequent analysis, namely data from other survey members within the respondent’s household. In this paper we explore the implications of ignoring this element of household survey design and propose methods for making stronger use of such data.

Household connections in social surveys

Large-scale social surveys typically take one of two strategies of recruiting participants from randomly selected households. Some surveys, such as the British Social Attitudes Survey (Park et al. 2004), randomly select one individual per household to take part, sometimes asking basic socio-demographic and/or socio-economic questions about other household members. Other surveys, such as national censuses, Understanding Society (UKHLS) and the Labour Force Survey in the UK, gather full detailed information from all household members. In both cases it is plausible to construct additional variables that give information about some or all of the (other) individuals connected to. This strategy has been employed in various circumstances, but is perhaps most familiar in socio-economic studies in the form of ‘household income’ measures (e.g. Jenkins 2011), and alternatively when researchers have constructed measures of social class based upon household-level profiles (e.g. Davies and Elias 2010; Marshall et al. 1995) and/or by distinguishing between the class influences of individuals and their spouses or other alters (e.g. Wright 1997, c10). However, such measures of household context tend not to be applied systematically across application areas. It is possible that, because many previous studies have proceeded without using such indicators (and because the construction of such measures is potentially burdensome), most social science analysts are not in the habit of routinely constructing and testing for the influence of additional explanatory variables based upon the characteristics of other household sharers.

The principle of homophily, that people tend to share characteristics with those that they interact with, is widely accepted across the social sciences (McPherson et al. 2001; Brashears 2008; Mackinnon et al. 2011). One component of homophily, within-household homophily, is equally widely recognised; two people living within the same household are relatively more likely to share a similar educational level, ethnicity, political outlook, and long-term socio-economic and health prospects, for example (e.g. McPherson et al. 2001). A number of social mechanisms may lie behind such homophily, including processes of attraction or adaptation (Brynin et al. 2008; Levitt and Leonard 2013), for instance couples attending the same cultural events may have similar interests which brought them together (attraction) or attend events together they have differing interest in (assimilation) (Upright 2004). Alternatively, mechanisms of co-dependence could occur, for instance with the time spent on housework chores depending on cohabitee's attitudes towards housework and cleanliness.

The influence of others can be observed in Figure 1, which shows correlates between members of the same couple or household across a range of measures. The inter-cluster correlation can be interpreted as the average level of correlation between the response values for any two individuals within the same cluster. Figure 1 demonstrates that high correlations are often found, even after controlling for basic socio-demographic effects. Whilst in some instances correlations might be expected (for instance, couples attending the cinema together, eating the same meals or feeling financially secure), similarity in terms of running, volunteering and smoking suggests that our behaviours might be influenced by those around us, whether through exposure to the harms/benefits of such actions or through access to undertaking such activities.

FIGURE 1 ABOUT HERE

Whilst we can assume that all people are influenced by their social connections, surveys collecting information on one individual per household are unable to capture such effects. However, household surveys which interview all members can generally only explore within-household ties and produce a statistical problem, namely that the 'clustering' of individuals within households generates non-independence of cases which ought not to be ignored from analysis[1]. Whilst this

[1] Many statistical models, such as regression models, include the assumption that cases are independent, which would be violated, leading to erroneous results, if multiple individual records came from the same household (unless some additional parameter for the relationship has been specified to account for that) (e.g. Berry 1993). However, even arithmetic statistical results are at risk of bias when analysis is applied to multiple cases from the same household. For instance, we can see that the characteristics of larger household units clearly have a greater influence on the descriptive results, so that chance sampling variations in selecting households play out unevenly for larger and smaller households.

point is reasonably well known amongst methodologists, it is not always satisfactorily addressed in research applications. Indeed, within this journal from 2009 to 2014 there were six articles which contained detailed data on multiple household members, none of which mentioned how they controlled for such structure[2]. Most secondary social surveys provide recommended sampling weights and related guidance which is designed to address the statistical shortcomings from ignoring household clustering alongside other aspects of the survey design (e.g. Davern et al. 2009; Rafferty 2009; Lynn and Kaminska 2010; Cleveland et al. 2011), and in one social science discipline, econometrics, it is common practice to apply 'robust standard errors' to recognise clustering of multiple records within the household (e.g. UCLA Statistical Consulting Group 2014). However, many solutions equate, to all intents and purposes, to down-weighting the influence of multiple respondents from the same household (or selecting only one person per household). This is often described as treating the household clustering as a 'nuisance', re-adjusting the data so that it can be treated as if it were from a sample of independent cases.

Hierarchical, multilevel models are often used to recognise the role of household structure. Literature in political science highlights that the household provides a valuable level for understanding voting behaviour, for instance showing how family relations can produce shared values and outlooks (Zuckerman and Kotler-Berkowitz, 1998) that can be captured by using tools such as random effects models (e.g. Johnston et al. 2005). Steele et al. (2013) utilise statistical modelling tools that demonstrate the important influence of shared household environment on choices of housing tenure and location, and Chandola et al. (2005) demonstrate how random effects could be used to improve analysis of health related outcomes by treating the household as a clustering factor in analysis , concluding that '*contextual effects on health at the household level may need to be assessed before recommending policies on improving health through focusing on larger units of aggregation, like neighbourhoods, wards or districts*' (Chandola et al. 2005: 174). A characteristic of these studies is an effort to use a statistical modelling framework that makes the influence of other individuals within the household more central to the analysis. However, the 'random effects' approaches considered hitherto have tended to involve a fairly narrow range of permutations, and have focussed conceptually upon controlling for statistical 'similarity' of responses within households (whereas, as previously noted, in some scenarios there may also be a need to consider devices that capture responsive or dependence relationships, that might even

[2] See Supplementary Appendix for discussion of these papers. We accept that controls for household structure might have been tried and not reported in many reports due to the minor effect observed. This note is not a criticism of those papers, but rather an example of how often these processes are unreported or overlooked.

result in dissimilarity rather than similarity within the household). Similarly, the use of random effects models at household level have, typically, focused on data from a single wave of data, perhaps due to the added complexity in the changing composition of households that individuals inhabit and the difficulty in capturing household units which shift between waves. Accordingly, our argument is that much survey research can and should be enhanced by applying a wider variety of statistical modelling approaches that are designed to recognise the influences of other respondents from the same household.

In this paper, we explore two potential methods for capturing the influence of others within the household from social surveys. Firstly, we explore other intra-household structures to determine appropriate clustering units for multilevel models. Secondly, we adopt an 'alter explanatory variable' to distinguish effects 'between' and 'within' households.

Optimising configuration of within-household alters

There can be considerable variation in operational definitions of the household between nations, and in some instances between different surveys within the same country (e.g. Hoffmeyer-Zlotnik and Warner 2014; Casimir and Tobi 2011)[3]. In the UK, most social surveys define households in line with the Office for National Statistics' Census definition, which is "*a group of people (not necessarily related) living at the same address who share cooking facilities and share a living room or sitting room or dining area*" (ONS 2009:4; ONS 2012). This definition highlights physical aspects of the home, but it means that a household could comprise a number of different combinations of individuals - for instance, a couple living alone; two parents and their children; four unrelated adults, three of whom are renting bedrooms from the fourth who is the home owner. However, the social mechanisms related to the 'influence of others' might sometimes more closely reflect emotional and kinship ties rather than physical proximity. For instance, some households might usefully be thought of as comprising smaller subgroups of closely connected individuals, sometimes described as 'concealed households' (ONS 2014), and formal criteria for defining component 'families' within households have sometimes been specified (e.g. Haskey 2010).

[3] Moreover, Hoffmeyer-Zlotnik and Warner (2014) highlight that the allocation of individuals to household units according to an agreed definition is itself sometimes handled inconsistently by survey respondents, data collection agencies, and research analysts (e.g. Gerber et al. 1996; Hoffmeyer-Zlotnik and Warner 2008).

Studies incorporating social relations within large-scale social survey analysis have typically operationalised the household as a unit (Zuckerman and Kotler-Berkowitz 1998; Johnston et al. 2005; Steele et al. 2013; Chandola et al. 2005). Table 1 lists six different configurations of social connections that can be specified according to information that is available in most household surveys. Each configuration represents different sorts of relationship with 'others' that could be responsible for shaping individual behaviours and outcomes, with some overlap with concepts of intra-household relationships that have been discussed in previous studies (e.g. Haskey 2010; Hoffmeyer-Zlotnik and Warner 2014). The Table uses data from Wave 2 of the UK Household Longitudinal Study (University of Essex 2013). For instance, it shows that the 54,597 different individuals interviewed in the UKHLS came from 30,476 different households, and from 38,726 different 'couples'[4].

TABLE 1 ABOUT HERE

Four of the configurations listed in Table 4 are quite well known. The 'person' level regards all individuals as being independent, ignoring household sharers. The 'couple' level combines people with their 'partners', if they have any. To be clear, those without a cohabiting partner are located in a 'couple' group comprising solely of themselves; for instance, two married adults with one child in the household would be defined as one 'couple' group comprising the two parents, and another 'couple' group containing solely the child. The 'economic family' configuration defines groups whose financial situation is intertwined, such as couples and any financially dependent children (this configuration, which is also often known as the 'consumer unit' and 'benefit unit', is sometimes used to define measures of 'family income', that are not equivalent to measures of 'household income' – cf. Jenkins 2011: 80). Fourthly, the 'household' level is typically defined as all individuals living within a specified physical environment, irrespective of any family or emotional relationships. The two other groupings are rarely studied to our knowledge. The 'inner family' is a slightly more inclusive concept than the 'economic family', and incorporates caring relationships. For instance, it includes children who are financially dependent, and anybody who is cared for by other group members, but

[4] Note that these figures refer to the population of people who gave a full interview to the UKHLS. In this table, if one person gave a full interview but all of their household sharers declined the interview or were unavailable, then they are treated as a single person household in terms of the records that they contribute to the UKHLS.

it excludes co-resident children who have their own partners or are financially independent (the financial independence of children is itself a phenomena that may be related to family and parenting strategies – e.g. Lareau, 2003). Lastly, the 'wider family' contains any individuals linked by connections, however tenuous, that involve blood, marriage, guardianship or care. They could include, for instance, three generational families, aunts, cousins and so forth. Whilst these six configurations capture a wide range of household structures, they are not necessarily exhaustive. For instance, the definitions treat step-family and adopted family ties as equivalent to other family ties, but other permutations could potentially be defined recognising complex 'blended' family formations. In addition, some individuals may be substantially influenced by family members who do not live in the same household and therefore configuration; however since existing social surveys rarely have data on extra-household family relationships, such possibilities are not incorporated in our review below.

The Intra-cluster correlation (ICC) figures in Table 1 indicate that non-negligible patterns of variation in responses may be associated with the respective configuration patterns. These statistics capture the 'variance component' associated with the unit. In the two-level case, for instance, it suggests that 63% of variance in the 'smoking' outcome can be linked to patterns of differences between economic families rather than to individual variation within economic families (another interpretation of equal validity is that the average correlation in smoking behaviour between different cases within the same economic family is 0.63). In addition, the '6-level' model is an ambitious application of a random effects model which seeks to identify patterns of variance at all six of the hierarchical configurations. In practice, a model with such a high number of levels is difficult to identify and is rarely used in applied statistics. The results shown are from a model with imperfect convergence, but they are nevertheless suggestive that certain proportions of the variations in the outcome can be linked to distinctive levels net of each other: for example, 24% can be linked the household, but a further 10% can be linked to the wider family distinctively from the household; net of these, only small proportions can be linked distinctively to the couple and economic family units (6% and 2% repsectively), and no pattern at all (0% of the variance) can be distinctively attributed to the inner family, net of the influence operating through the other configuration units.

One plausible hypothesis is that different configurations are optimal when studying different social processes. In fact, Figure 1 above also suggested this phenomena, because it showed that the difference between the ICC for a 'couple' and a 'household' configuration was itself different for different outcome measures. Table 25 therefore summarises ICC values and model fit patterns for an indicative range of outcome measures using random effects for different configurations of clusters. The results of two-level random effects models (columns 2-4) suggest that across all the outcomes, every one of the configurations provides improved statistical fit compared to the model where it was ignored. For one outcome, the 'best' statistical performance (in terms of highest explained variance) is with the rarely used 'inner family' configuration, although for all other outcomes, it is either the conventional 'household' or 'couple' configurations that capture the most variance. The table also suggests different patterns for different outcomes; for some outcomes, the patterns linked to the configuration are particularly strong, as seen in the high values of the ICC for the measures of financial anxiety, voting and cinema attendance, as well as moderately high values for several other measures related to lifestyle and preferences. There are, perhaps, substantive rationales for these differences- i.e., subjective wellbeing (GHQ) is most strongly clustered amongst the emotional ties of the inner family, whereas financial anxiety is clustered within couples.

TABLE 2 ABOUT HERE

In addition, since both the 'household' and the 'couple' units are especially commonly exploited, it is interesting to ask whether 3- or 4-level random effects models can usefully identify any additional influence of other configurations over and above these key configurations. The second panel of Table 5 tests this, reporting whether or not the given configuration was associated with a significant variance component over-and-above a distinctive variance component for the couple or household configuration (i.e. in a three level model) or both (i.e. in a four level model). In most but not all cases, there is evidence that the extra configuration offers improved model fit over and above a control for one other structure (i.e. the household or the couple). However, there is only one permutation where a configuration matters distinctively over and above the separate effects of both household and couple configuration (the 'economic family' configuration for the outcome measure of financial anxiety). This pattern is potentially insightful, revealing that there may be distinctive empirical patterns captured by the 'economic family' that are not fully incorporated in patterns linked to the couple or family. In the other cases, however, it was not possible to distinguish such separable patterns.

The important point to bear in mind is that most studies that use individual level data from the surveys with a household sampling design such as the UKHLS work with the sample of (in this case 54,597) individual records, without adjustment or acknowledgement that one or more other records in the dataset may come from another person within the same configuration. The specification and brief evaluation of different household-based configurations above suggests many opportunities.

Modelling information about the household

Thus far, our discussion has focussed on random effect multilevel models. It follows that other forms of model adaptation, such as fixed effects specifications, are equally likely to lead to differences in results compared to models which ignore household structure. The 'fixed effects' model allows us to focus on patterns that influence differences in the outcome within the higher level unit (i.e., household), providing a 'within effects' influence, i.e. a statement about how differences in the X variables within a cluster tend to be related to differences in the Y variable (Allison 2009). The difference between 'between effects' and 'within effects' in statistical models is a complex one, about which solutions are not entirely agreed upon (e.g. Allison 2009), but there are various devices available that may allow us to calculate separate 'between effects' and 'within effects' parameters (or indeed to better interpret intermediate parameters that may be influenced by both processes). Moreover, attention to household clustering patterns in data is probably of the most substantive importance if the patterns of influence upon the outcome are indeed expected to work in different ways when conceptualised as differences 'between households', or differences amongst people 'within households'. For instance, an example of a 'between effect' would be that households with individuals with higher average paid working hours tend to have higher average levels of subjective well-being; a related but substantively conflicting 'within effect' could be the finding that, when comparing different people from the same household, those who work longer hours tend on average to have lower subjective well-being than their cohabitees who work shorter hours. It is possible that both patterns could co-exist, and in such situations, the parameters from a model that ignores clustering might be substantively misleading because they will reflect some unspecified combination of 'between' and 'within' effects.

Social surveys often contain information relating to the household, such as total household income, access to motor vehicles and occupancy rating, which are, by their very nature, shared by all household members. However, variables which are structurally shared by all cluster members offer no deviation and therefore cannot be modelled using 'fixed effects'. Similarly, as 'fixed effects' models analysed the differences within higher level units, they cannot analyse clusters comprising a single case. In terms of household analyses, therefore, all respondents living alone (or with people not including in the survey) would be discarded from the data. Similarly, analysis at the couple level would ignore anyone without a partner.

Group means are sometimes produced, summarising data on all household members, including the respondent. An alternative approach would be to create an 'alter explanatory variable' (AEV) which produces the group mean of an individual's cohabitees. Thus, it is not a measure of the household that individual's reside in, but the characteristics of the residents they cohabite with. The average can be used both for continuous measures (e.g. the average well-being score for all household sharers) and also for categorical outcomes that are represented by dummy variables (e.g. the proportion of alter's within the household who smoke). These AEV's exclude the individual and therefore often differ for people within the same household; for instance, if only one member of a three-person household smokes, they will be shown as living exclusively with non-smokers whereas the other members see 50% of their co-residents smoke. Thus, the AEV captures the composition of social ties rather than the composition of the unit analysed. An appealing feature of this approach is the complexity associated with the fixed effects model can be incorporated within models which can be easily interpreted, and constructed, by people lacking sophisticated knowledge of multilevel analysis. Similarly, the AEV measure enables within-household effects to be observed in multiple wave data as the shape of their clusters change, which cannot readily be undertaken using a random effects framework.

In addition, models with explanatory variables about the household can potentially be combined with random effects multilevel models. This generates what is often described as a 'combination' model or 'hybrid' model (e.g. Allison 2009), because it recognises the household structure both in terms of adjustments to the 'error' terms, and in terms of additions to the explanatory variables.

However, generating an AEV raises difficulties in dealing with missing data where individuals have no alters within their household. In a 'complete cases' approach to analysis, this would mean that many cases would be excluded from analysis because they have missing data on the relevant alters' score

– for instance, individuals living alone would be dropped from the analysis. We suggest using mean imputation to circumvent these issues, centring the mean as zero and imputing that score for those without any connections in that configuration, either through having no alters or their alters having no valid responses (Kreft and De Leeuw 1995). This means that the AEV effects should be interpreted as the difference from the mean, where observable.

The AEV as an explanatory variable

Cheng et al. (2007) analysed eating patterns, from 1975 to 2000, using time diaries. Their research focused on the amount of time individuals spent eating and drinking per day, concluding that over the years people reduced the time they spent enjoying their meals. Using the 2000-01 National Time Use Survey (IPSOS-RSL 2003), a survey with a household cross-sectional design, Cheng et al.'s analysis did not feature controls for the clustering of cases into household. Potential household level measures were also not considered at length, although the presence of pre-school and school-age children within the household was explored as an explanatory factor in the analysis.

We have replicated this study to explore how constructing an AEV measure alters the findings. This study was chosen for two reasons. Firstly, it presents an outcome variable which might be clustered within households; families who eat together might feasibly spend similar times on their meals. Secondly, we wish to explore the methodological consequences of adopting an AEV measure, rather than challenging existing empirical work without fully exploring additional explanatory factors. As Cheng et al. conducted a comparative study, using 1975 data without a household framework, it was both impossible for their research to adopt clustering techniques at both time-points and also our findings should not impinge on their findings.

Table 3 replicates the Cheng et al. (2009: 46, Table II, column 1) analysis, including several treatments to control for household effects. Column 1 fits a model without any household controls, using the variables included in the initial study. The subsequent columns show, respectively, an AEV explanatory factor, a random effects model, a hybrid model (random effects including the AEV measure) and a fixed effects model, all at the household level[5]. Each model shows a marked improvement in model fit when household clustering is treated. Including the AEV measure showed

[5] Analysis of each within-household composition suggested the strongest model fit was associated with the entire household.

a decrease in BIC, a measure of parsimony, indicating the model is substantively improved by the inclusion of the additional variable. Whilst both the random effects and hybrid models show an improved model fit over not controlling for household structure, the improvement in the models compared to sophistication of the model was not as marked as when using the AEV measure alone. The fixed effects model drops the 2,007 cases living with no other valid respondents as it analyses differences within units, also omitting the presence of children in household from the analysis as there can be no within-unit variance. Thus, the smaller sample size prevents easy comparison of BIC, although the R^2 value suggests the fixed effects model describes just 2.9% of variance in the duration of eating and drinking, as opposed to 27% in the AEV model. Thus, there is evidence that, in this application, the AEV model provides the most convincing model fit. Finally, within the two random effects models the hybrid model improves parsimony, increasing the level of variance attributable to the individual, as opposed to the household, from 40% to 96%.

Applying an AEV measure does not merely improve model fit, but potentially alters the interpretation of the results obtained. Whilst most of the explanatory factors remain fairly constant, the AEV model implies that neither age, nor schoolchildren within the household are significant predictors of eating duration. Rather, living with people who eat slowly increases the duration of an individual's mealtimes, with no additional effects around age or children within the home. Those relationships observed when ignoring household composition appear to be spurious – perhaps with unmeasured household factors, such as eating together as a family, being more common in certain types of home. The fixed effects model similarly showed no significant effects for age within households, supporting the argument that whilst older households tend to spend longer on their meals it's not necessarily the case that older people do so. The random effects model suggested age was significant, highlighting the limitation of ignoring the 'within' effects. Whilst this distinction is, perhaps, a little trivial, and does not distract from the interpretation Cheng et al. (2007) made, this demonstrates that using an AEV measure can produce a meaningful improvement to statistical results and alter the substantive interpretation of results.

TABLE 3 ABOUT HERE

Model selection, however, should not be constrained to assessing parsimony and variance explanation (Angrist and Pischke, 2009). Tests for heteroskedasticity and problematic multicolinearity were undertaken on the models with and without parameters recognising household level structures. Across models, problematic multicollinearity did not seem to arise. However, regarding heteroscedasticity, a substantial difference was observed between results with and without adjustments recognising the household clustering. The model without household level controls suggested a Cook-Weisberg chi-square value of 7.29, suggesting significant but slight heteroscedasiticy that might in practice often be ignored (e.g. Kaufman 2013); however, including the AEV for the household increases the Cook-Wesiberg chi to 218.72, suggesting much more important heteroscedasticity problems, and that careful attention is required to determine the most appropriate model. There is no statistical rationale to assume that the AEV should increase issues of heteroscedasticity and, indeed, in many applications it is plausible such problems could be resolved.

Discussion

In this paper we have explored the influence of different household configurations on individual outcomes. This has extended recent similar work (Goldstein et al. 2000; Johnston et al. 2001; Lambert 2001; Chandola et al. 2005; Steele et al. 2013) in terms of the methods, the types of household, and the solutions recommended. Amongst methodologists it is reasonably well known that failing to control for household configurations could lead to results that are biased (choose inappropriate estimates) and/or inefficient (make inappropriate inferences). However, this awareness has rarely influenced applied research studies.

Our research shows that for a wide range of sociologically important outcomes, the use of statistical models that take into account information about other household members improves the quality of findings. Both the optimal statistical model, and the optimal configuration type, differs between outcomes, which makes a strong case for sensitivity analysis when using household level datasets across alternative models and configurations. We believe that the quantitative social sciences have often neglected the influence of alters, and that the utilisation of household structure has often been viewed as a methodological rather than substantive empirical issue. Indeed, we have found it more common for analyses to control for the effects of the others within the region than within their own household. Since it is often impossible, prior to analysis, to know if household factors play an important role in shaping individual-level outcomes or behaviours, it is hard to argue against further analyses of household effects when the data supports it. Chandola et al. (2005) advised that researchers using household data should estimate household level variance patterns initially to identify whether important patterns are likely to be related to the household, however they emphasis a random effects multilevel modelling framework to account for such similarities, whilst we demonstrate that other variables based upon the household, such as an 'AEV' which provides a characterisation of the circumstances of alter(s), can achieve similar contextualisations. Indeed, in some scenarios, where a random effects model is not feasible, the testing of an AEV or other household level measure may provide a simpler and more effective way to establish whether household clustering requires further consideration.

Many such evaluations are quite easily achieved when using datasets with household identifiers. Indeed, we would argue that survey compilers could routinely provide contextual information about households to enable researchers to quickly assess the likelihood of household-level clustering, for instance through providing household level ICC statistics for each variable within their standard documentary webpages. Similarly, we feel that the categorisation of individuals into different household configurations is something that could be automated by survey providers, which can be found in the online supplementary appendix. This would allow additional variables to be utilised by researchers to analyse different configurations of the 'household'.

Our analysis suggests that whilst social scientists have often overlooked within-household connections, this has not necessarily invalidated their findings. Across several models we demonstrated that we were able to improve model fit, but without substantially altering the coefficients or their standard errors, therefore leading to the same overall story. Critics might suggest this means we are recommending an extra step which brings further mathematical puzzles but that will often be only of limited consequence. However, there are two valid counter-arguments. Firstly, academic papers should seek to fit into a wider body of evidence. Therefore, even minor shortcomings in results, such as associational levels which are not optimised, might not substantively alter the interpretation of the results, but could cause difficulties if those figures are utilised in later research projects or meta-analyses. Secondly, it cannot usually be known in advance if analysing the within-household configurations will change the substantive sociological conclusions and, therefore, any analysis which does not try such approaches remains at risk of presenting incorrect results.

There are many surveys and datasets in the UK social sciences which provide information on all household members, which could readily be used to study 'the influence of others'. As it stands, most social scientists accept that ‘context’ or ‘composition’ matters to social processes, but our observation is that very few empirical studies are effective in building appropriate measures of the important context of the immediate household or family into their individual level analyses; if anything, they are much better as measuring and summarising wider contexts such as of the area or more distant ‘social capital’. Whilst it might be impractical for compilers of surveys for highly specific purposes to interview all household members, there may sometimes be a sensible intermediary position whereby they ask some questions about household sharers on key outcomes. For instance, we could envision a study focussing on levels of obesity within a specific workplace to ask for the respondent’s perceived BMI for each household members and their relationship with that person. This would enable such issues to be controlled for at relatively low cost. Although perceptions of measures may often be inaccurate, they may still serve as a parsimonious approximation (whilst minimising ethical and data privacy problems).

Bibliography

Allison, P. D. (2009). *Fixed Effects Regression Models*. Thousand Oaks, Ca.: Sage.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Berry, W. D. (1993). *Understanding Regression Assumptions*. Newbury Park, California: Sage.

Blackburn, R. M., Stewart, A., & Prandy, K. (1980). *Social Status in Great Britain, 1974 [computer file]*. Colchester, Essex: UK Data Archive [distributor], SN: 1369.

Bott, E. (1957) *Family and Social Network*. London: Tavistock.

Brashears, M.E. (2008) ‘Gender and homophily: Differences in male and female association in Blau space’, *Social Science Research*, 37, 400-415.

Brynin, M., Martinez Perez, A., & Longhi, S. (2008). How Close are Couples? In M. Brynin & J. Ermisch (Eds.), *Changing Relationships* (pp. 93-108). London: Routledge.

Casimir, G. J., & Tobi, H. (2011). Defining and using the concept of household: A systematic review. *International Journal of Consumer Studies, 35*, 498-506.

Chandola, T., Clarke, P., Wiggins, R.D., and Bartley, M. (2005) ‘Who you live with and where you live: setting the context for health using multiple membership multilevel models’, *Journal of Epidemiology and Community Health*, 59, 170-175

Cheng, S-L., Olsen, W., Southerton, D., and Warde, A. (2007) ‘The changing practice of eating: evidence from UK time diaries, 1975 and 2000’, *British Journal of Sociology*, 58(1), 39-61.

Christakis, N., and Fowler, J. (2009) *Connected: The amazing power of social networks and how they shape our lives*. London: Harper Press.

Cleveland, L. L., Davern, M., & Ruggles, S. (2011). *Drawing Statistical Inferences from International Census Data*. Minneapolis: IPUMS-International Working Paper, International Public Use Microdata Series, Minnesota Population Center, University of Minnesota.

Davern, M., Ruggles, S., Swenson, T., Alexander, J. T., & Oakes, J. M. (2009). Drawing Statistical Inferences from Historical Census Data, 1850–1950. *Demography, 46*(3), 589-603.

Davies, R., & Elias, P. (2010). The application of ESeC to three sources of comparative European data. In D. Rose & E. Harrison (Eds.), *Social Class in Europe: An introduction to the European Socio-economic Classification* (pp. 61-86). London: Routledge.

Gerber, E. R., Wellens, T. R., & Keeley, C. (1996). 'Who lives here?' The use of vignettes in household roster research. In *Proceedings of the section on survey research methods* (pp. 962-967). Alexandria, VA: American Statistical Association.

Gilbert, G. N. (2008). *Agent-Based Models*. Thousand Oaks, Ca.: Sage.

Goldstein, H. (2010). *Multilevel Statistical Models, 4th Edition*. New York: Wiley.

Goldstein, H., Rasbash, J., Browne, W., Woodhouse, G., & Poulain, M. (2000). Multilevel Models in the Study of Dynamic Household Structures. *European Journal of Population, 16*(4), 373-387.

Halsey, A. H., Goldthorpe, J. H., Payne, C., & Heath, A. F. (1978). *Social Mobility Inquiry, 1972 [computer file]*. Colchester, Essex: UK Data Archive [distributor], 1978. SN: 1097.

Harris, K.M., Halpern, C.T., Whitsel, E., Hussey, J., Tabor, J., Entzel, P., and Udry, J.R. (2009) *The National Longitudinal Study of Adolesecent Health: Research Design*. URL: http://www.cpc.unc.edu/projects/addhealth/design.

Haskey, J. (2010). Measuring Family and Household Variables. In M. Bulmer, J. Gibbs & L. Hyman (Eds.), *Social Measurement through Social Surveys: An Applied Approach* (pp. 9-27). Aldershot: Ashgate.

Hoffmeyer-Zlotnik, J. H. P., & Warner, U. (2008). *Private Household Concepts and their Operationalisation*. Mannheim: GESIS.

Hoffmeyer-Zlotnik, J. H. P., & Warner, U. (2014). *Harmonising Demographic and Socio-Economic Variables for Cross-National Comparative Survey Research*. Berlin: Springer.

Hox, J. (2010). *Multilevel Analysis, 2nd Edition*. London: Routledge.

Jenkins, S. P. (2011). *Changing Fortunes: Income Mobility and Poverty Dynamics in Britain*. Oxford: Oxford University Press.

Johnston, R., Jones, K., Propper, C., Sarker, R., Burgess, S., and Bolster, A. (2005) 'A missing level in the analysis of British voting behaviour: the household as context as shown by analyses of a 1992-1997 longitudinal survey', *Electoral Studies*, 24, 201-225.

Kaufman, R. L. (2013). *Heteroskedasticity in Regression*. London: Sage.

Kreft, I.G.G., and De Leeuw, J. (1995) 'The effect of different forms of centering in hierarchical linear models', *Multivariate Behavioral Research*, 30, 1-21.

Ipsos-RSL & Office for National Statistics (2003) *United Kingdom Time Use Survey, 2000* [computer file]. *3rd Edition.* Colchester, Essex: UK Data Archive [distributor], September 2003. SN: 4504.

Jenkins, S. P. (2011). *Changing Fortunes: Income Mobility and Poverty Dynamics in Britain*. Oxford: Oxford University Press.

Lambert, P. S. (2001). *Individuals in household panel surveys: dealing with person-group clustering in individual level statistical models using BHPS data*. Colchester, UK: Paper presented to the British Household Panel Survey Research Conference, Intstitute for Social and Economic Research, University of Essex, and http://www.iser.essex.ac.uk/files/conferences/bhps/2001/docs/pdf/papers/lambert.pdf.

Lareau, A. (2003) *Unequal Childhoods: Class, Race, and Family Life*. London: University of California Press.

Levitt, A., and Leonard, K.E. (2013) 'Relationship-specific alcohol expectancies and relationship-drinking contexts: Reciprocal influence and gender-specific effects over the first 9 years of marriage', *Psychology of Addictive Behaviours*, 27(4), 986-996.

Lynn, P., & Kaminska, O. (2010). *Weighting Strategy for Understanding Society*. Colchester: Understanding Society working paper series, 2010-05, Institute for Employment Research, University of Essex.

Mackinnon, S.P., Jordan, C.H., and Wilson, A.E. (2011) 'Birds of a Feather sit together: Physical Similarly Predicts Seating Choice', *Personality and Social Psychology*, 37(7), 879-892.

Marshall, G., Roberts, S., Burgoyne, C., Swift, A., & Routh, D. (1995). Class, gender and the asymmetry hypothesis. *European Sociological Review, 11*(1), 1-15.

McPherson, J.M., Smith-Lovin, L., & Cook, J.M. (2001) ‘Birds of a Feather: Homphily in social networks’, *Annual Review of Sociology*, 27, 415-444.

ONS. (2009). *Final Population Definitions for the 2011 Census*. London: Office for National Statistics, and www.ons.gov.uk

ONS. (2012). *Harmonised Concepts and Questions for Social Data Sources, Primary Standards: Demographic Information, Household Composition and Relationships*. London: Office for National Statistics (and http://www.ons.gov.uk/ons/guide-method/harmonisation/primary-set-of-harmonised-concepts-and-questions/, 1 April 2014)

ONS (2014) *What does the 2011 Census tell us about concealed families living in multi-family households in England and Wales?* London: ONS.

Park, A., Curtice, J., Thomson, K., Bromley, C., & Phillips, M. (Eds.). (2004). *British Social Attitudes: The 21st Report*. London: Sage.

Rafferty, A., & (2009). *Introduction to Complex Sample Design in UK Government Surveys*. Manchester: ESDS Government, Economic and Social Data Service, University of Manchester.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Steele, F., Clarke, P., & Washbrook, E. (2013) ‘Modelling Household Decisions Using Longitudinal Data from Household Panel Surveys with Applications to Residential Mobility’, *Sociological Methodology*, 43(1), 220-271.

UCLA: Statistical Consulting Group. (2014). Stata Library: Analyzing Correlated (Clustered) Data. Retrieved 1 October, 2014, from http://www.ats.ucla.edu/stat/stata/library/cpsu.htm

University of Essex, & Institute for Social and Economic Research. (2010). *British Household Panel Survey: Waves 1-18, 1991-2009 [computer file], 7th Edition*. Colchester, Essex: UK Data Archive [distributor], July 2010, SN: 5151.

University of Essex, Institute for Social and Economic Research and National Centre for Social Research (2013) *Understanding Society: Waves 1-3, 2009-2012* [computer file]. 5th edition. Colchester: UK Data Archive.

Upright, C.B. (2004) ‘Social Capital and Cultural Participation: Spousal Influences on Attendance at Arts Events’, *Poetics*, 32, 129-143.

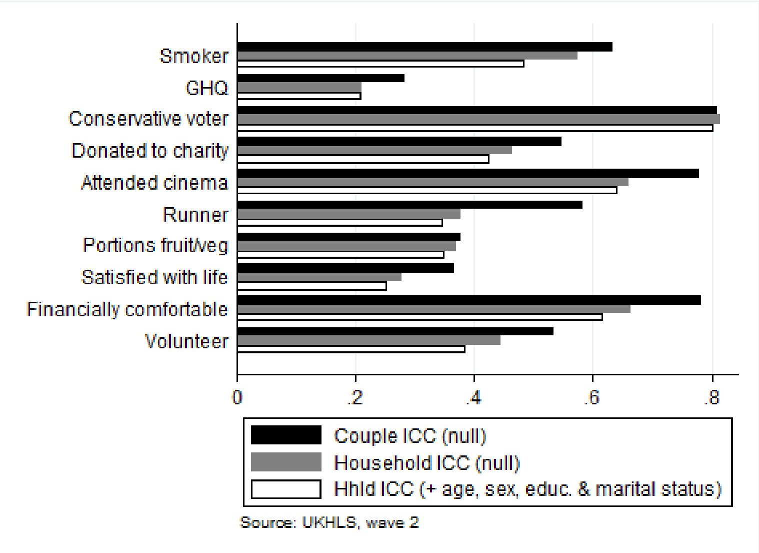
West, P., and Sweeting, H. (1995) *Background Rationale and Design of the West of Scotland 11-16 Study*. Working paper no. 52. Glasgow: MRC Medical Sociology Unit.

Wright, E. O. (1997). *Class Counts: Comparative Studies in Class Analysis*. Cambridge: Cambridge University Press.

Young, M., & Willmott, P. (1957). *Family and Kinship in East London*. Harmondsworth: Penguin.

Zuckerman, A.S., and Kotler-Berkowitz. (1998) 'Politics and society: political diversity and uniformity in households as a theoretical puzzle', *Comparative Political Studies*, 31, 464-497.

Tables and Figures



Source: UKHLS, wave 2

Figure 1: Inter-cluster correlations for selected outcomes.

Code	Category	Description	# Groups (2-level ICC for 'smoker') (6-level ICC for 'smoker')
PID	**Person**	Individual only	54,597 (n/a) (0.58)
CID	**Couple**	Married or cohabiting couples (16k pairs) or singles (22k)	38,726 (0.64) (0.06)
EID	**Economic family**	CIDs or singles plus dependent children (of either partner)	38,673 (0.63) (0.02)

IID	**Inner Family**	CIDs or singles plus unmarried & childless children (either parent); plus anyone they care for	38,496 (0.62) (0.00)
WID	**Wider Family**	Any family member linked by blood, marriage, guardianship, care	31,703 (0.56) (0.10)
HID	**Household**	All household sharers	30,476 (0.56) (0.24)

Table 1: Household configurations and number of groups in UKHLS eave 3 (2010/11)

	Couple	**Economic family**	**Inner family**	**Wider family**	**Household**
	*2-level models: Level 2 ICC with * to indicate significant deviance reduction from individual model (lowest deviance underlined)*				
GHQ	0.28*	0.28*	0.28*	0.21*	0.21*
Satisfaction with life	0.15*	0.15*	0.15*	0.13*	0.12*
Financial Anxiety	0.73*	0.73*	0.73*	0.55*	0.55*
Job satisfaction	0.11*	0.11*	0.11*	0.11*	0.11*
Tory voting	0.72*	0.72*	0.72*	0.70*	0.71*
Charity donations	0.55*	0.55*	0.55*	0.46*	0.47*
Cinema attendance	0.80*	0.80*	0.79*	0.66*	0.67*
Goes running	0.54*	0.54*	0.54*	0.36*	0.37*
Volunteers	0.52*	0.51*	0.51*	0.44*	0.44*
	Significant variance attributed to household (H), couple (C) and couple within household (HC) in 3-level & 4-level models				
GHQ	H	H	H	C	C
Satisfaction with life	H	H	H	C	C
Financial Anxiety	H	H C CH	H C	C	C
Job satisfaction					
Tory voting	H	H C	H C	H C	C
Charity donations	H	H	H	C	C
Cinema attendance	H	H C	H	C	C
Goes running	H	H C	H C	C	C
Volunteers	H	H C	H	C	C

Table 2: Model fit statistics for models with different configurations on selected outcomes (Source: UKHLS wave 2)

	Individual	AEV	Random effects	Hybrid model	Fixed effects
Part-time work	6.65 (1.54)**	5.16 (1.45)**	6.11 (1.47)**	1.83 (0.77)*	2.91 (1.87)
Unemployed	11.35 (4.01)*	10.31 (3.64)*	11.05 (3.35)**	4.91 (1.81)*	11.14 (4.45)*
Retired	20.03 (2.40)**	13.66 (2.22)**	18.67 (2.37)**	7.97 (1.51)**	14.31 (3.79)**

Student	12.16 (3.26)**	8.88 (2.97)*	11.68 (3.11)**	4.07 (1.60)*	5.31 (3.89)
Economic inactive - other	14.25 (2.29)**	9.25 (2.10)**	12.84 (2.25)**	5.37 (1.24)**	10.01 (3.06)*
Female	-2.05 (1.22)	-1.06 (1.12)	-1.30 (1.03)	-0.07 (0.50)	0.71 (1.22)
Couple	7.97 (1.42)	5.92 (1.35)**	7.41 (1.39)**	2.68 (1.22)*	0.97 (3.42)
Age in years	0.05 (0.20)	0.03 (0.19)	-0.06 (0.19)	-0.31 (0.13)*	0.10 (0.34)
Age squared	0.005 (0.002)*	0.004 (0.002)	0.005 (0.002)*	0.007 (0.001)**	0.003 (0.004)
Child >5 in hhld	-7.28 (1.60)**	-3.95 (1.44)*	-8.18 (1.99)**	-20.3 (3.3)**	
Child 5-16 in hhld	-5.28 (1.22)**	-1.63 (1.13)	-5.74 (1.46)**	-20.3 (2.4)**	
Further Education	0.34 (1.78)	1.20 (1.60)	1.03 (1.68)	-0.12 (0.88)	1.16 (2.15)
Higher education	8.23 (1.34)**	5.68 (1.24)**	6.43 (1.29)**	1.20 (0.77)	0.93 (1.91)
‘AEV’		0.47 (0.02)**		-1.01 (0.01)**	
Constant	64.03 (4.24)**	68.88 (3.93)**	68.05 (3.99)**	90.70 (2.80)**	69.31 (6.03)**
ll	-39,537	-38,866	-39,254	-38,930	-34,156
BIC	79,200	77,866	78,651	78,012	68,420
R^2	0.12	0.26			0.03
ICC			.40	.96	
Cases	7,520	7,520	7,520	7,520	5,513
Households	n/a	4,460	4,460	4,460	2,453

Table 3: Daily time consuming food and drink (Source: National Time-Use Survey, 2001) */** (p<.05/.001)