# NOTE ON STATISTICAL ANALYSIS AND MICROSIMULATION FOR STUDYING LIVING ARRANGEMENTS AND INTERGENERATIONAL TRANSFERS

***Douglas A. Wolf****

## ON MODELS FOR THE ANALYSIS OF LIVING ARRANGEMENTS AND INTRA-FAMILY TRANSFERS

In the area of multivariate modelling of living arrangements and family transfers, we are inevitably led towards a desire, or need, for complexity in model specification and hence difficulty in estimation. This complexity arises because we generally wish to represent the situations of multiple actors (decision makers), for example, an older person or couple and their several children, and, possibly, the children's parents-in-law as well. Furthermore, each actor may engage in one or more of a set of multiple activities of interest, including co-residence, financial transfers, or the provision of personal-care services. The spatial proximity of members of a kin group may be a further outcome of interest. Thus, we are faced with jointly modelling what may be an array of discrete and continuous (or truncated) outcomes of varying numbers over varying numbers of individual decision makers (when we look across observations in, say, a cross-sectional data file).

Yet, the econometrics literature provides a number of tools with which to formulate estimable models, and provides estimation algorithms, for most or all of the situations that might arise in the preceding substantive context. Software with which to estimate all the potential variants of these models may not be readily available, which can be a major problem. Nonetheless, the binding constraint at the margins of applied work in this area is more likely to come from data limitations than from a lack of suitable econometric machinery. Data sets may not include information on all the actors of interest, or may do so but provide too little information on each to permit interesting analysis. Furthermore, not all resource flows may be represented.

Paradoxically, while available data typically have numerous shortcomings with respect to their ability to support the estimation of complex resource flows in multiple domains in multi-actor family networks, available data have generally not been much exploited with respect to the analyses they could potentially support. As an example, Aykan and Wolf (2000), use Turkish Demographic and Health Surveys (DHS) data to model the competition between a married woman's parents and parents-in-law for co-residence with the woman and her husband; the DHS standard design was intended to inform research on fertility, family planning, child development and maternal and child health, yet the Turkish survey adopted optional questions on the survivorship of respondents' parents and parents-in-law, permitting our analysis with its "ageing" focus.

*Syracuse University, United States of America.

There are no doubt numerous other instances of available data that could support unexpected analyses in domains far from their originally intended range of topics.

A difficult issue that arises in the specification of models that depict outcomes in multiple domains is that of endogeneity, or simultaneity. But, these issues can easily be misunderstood. If two outcomes are both viewed as choice variables under the control of a single actor, then it does not make sense to think of them as reciprocally causally related (i.e., that a change in A causes a change in B, while B similarly produces its own distinctive causal response in A). Instead, they should be treated as "jointly determined". This will give the statistical specification of the model the appearance of a reduced form. Variables A and B are still jointly endogenous, and presumably depend, in part, on common unmeasured variables (i.e., exhibit correlated disturbances) but do not appear as each other's regressors. For example, an area of considerable research activity at present is the question whether women's hours of paid employment and their hours of familial caregiving activity are negatively related. If a woman is viewed as the sole decision maker, and her decision is made conditional on exogenously given prices (e.g., market wages, household productivity, and costs of market substitutes for her own caregiving time), a fixed "care production technology", and fixed preferences, then the two time-use outcomes are jointly determined.

However, there are many cases of outcome variables that are simultaneously determined, for example a woman's parent-care hours and her sister's parent-care hours, or, an older person's health and co-residence status, or, an older person's co-residence status and co-resident child's labour supply. In these cases, outcomes reflect the decisions of multiple actors, and one can imagine an exogenous change that would lead one actor to adjust, for example, her care hours, while that change would lead to a reaction on the part of some other actor's care hours.

Estimation of these reciprocal causal relationships demands identifying restrictions, that is, suitable instruments with which to identify the effects of endogenous variables. A great challenge -- two challenges, in fact -- is to decide which variables can be used as instruments in this type of model, and to ensure the presence in data sets of such variables. The theories on which choice models are based often offer little help in pointing out such instrumental variables; theories are more likely to suggest what factors do belong in a relationship than to suggest that some other factor definitely does not belong in that relationship. Ethnographic, or other qualitative and intensive, field efforts might help to produce the sort of empirical evidence that would support decisions about how to identify simultaneous-equations models.

Two final points about endogeneity deserve mention. First, longitudinal data, which might provide a temporal sequence of values for both independent and dependent variables, does not automatically provide a way out of the problem of establishing causality. Yet, many analysts seem to assume that the observed

temporal sequence is the same as the causal sequence. However, actors make plans, they have expectations about the future, and they take steps today that reflect their plans about the future. Thus, there is a sense in which events in the future "cause" events in the present. Secondly, "contextual" variables are not necessarily exogenous. Multilevel modelling is, at present, a popular and rapidly developing analytic tool, but as context is virtually always location-specific, one must recognize that the inclusion of contextual variables introduces possible endogeneity bias, as actors are to some extent free to choose their location. They may choose their location so as to achieve a favourable context, for example, older persons may migrate to a service-rich area (or to their child's neighbourhood) in anticipation of future care needs. If so, the contextual variables are not exogenous. This criticism is often made; solutions to the problem are far more rare.

## ON MICROSIMULATION, IN THE CONTEXT OF FAMILY/KIN NETWORKS AND INTRA-FAMILY TRANSFERS

It makes sense to turn from a discussion of model specification to microsimulation, since microsimulation must be preceded by model specification and estimation.

What is microsimulation? The essential ingredients are the use of computer-based sampling, and an analysis that is conducted at the maximally disaggregated level, that is, that of the individual (which might be a person, a couple, a firm or organization -- whatever is the fundamental analytic unit at hand). The "sampling" is, in fact, a process of making stochastic assignments of values to variables. These remarks pertain to a situation in which the "model" is a set of relationships among observed and unobserved factors (in the demographic domain, primarily); the unobserved factors are assumed to come from particular distributions; the model produces a distribution of possible values for the outcome of interest, and the computer program—the sampling algorithm—selects a particular value from that distribution. The sampling process may be repeated many times for a particular individual, and there may be many individuals (e.g., a sample, and even, perhaps, everyone in some population) to which the sampling algorithm is applied.

An interesting question is the following: is microsimulation a complement to, or an alternative to, "macro" simulation? Before addressing this question, it should be noted that the distinguishing features of "macro" versus "micro" simulation is not deterministic versus stochastic (stochastic aggregate population forecasting techniques, for example, are currently gaining increased attention), nor "expected value" versus "frequency distribution" (although advocates of microsimulation often give as a rationale for doing microsimulation the fact that it can produce an estimate of the population frequency distribution of some outcome, while macro models generally produce just the expectation), but, instead, aggregated versus disaggregated. To make progress towards answering the question posed above, there are areas where microsimulation has been shown to be useful, and others where macrosimulation has little or nothing to say. The best-known area in which microsimulation has proved useful is in depicting the details of kinship

networks, and the best-known work in this area is by Kenneth Wachter and his colleagues (Hammel, Wachter and McDaniel, 1981; Wachter, 1997).

Microsimulation could, in addition, be used to conduct a conventional population projection, but it is hard to imagine that anyone would seriously want to. Situations in which microsimulation does reveal its value include those characterized by (a) complex models—for example, multiple-equation models in which multiple actors make decisions about multiple interrelated domains of behaviour, such as the intra-familial transfer situations described above; (b) situations involving interactions between individual members of a population, for example the workings of mating markets; (c) models that explicitly represent "unmeasured heterogeneity", such as the "frailty" models of human mortality developed by Vaupel and colleagues (Vaupel and Yashin, 1985; Vaupel, Manton and Stallard, 1979) or the random-mixture models of Hutterite fertility developed by Heckman and Walker (1987); (d) the analyst's wishes to quantify the various sources of uncertainty, or forecast variance, in a model. Microsimulation can also be a way to extend the range of lessons that can be learned from some types of models. For example, in a conventional linear single-equation regression setting, most of what one might want to learn from the estimated model can be learned from the coefficients themselves, or simple transformations of them. Forecasts are also easy to carry out. In contrast, a Markov renewal model of, say, labour market transitions may incorporate a set of age- and duration-dependent hazard functions for transitions among states "never worked", "working", "unemployed", and "retired". Further complexity can be introduced by distinguishing between different jobs held over the worklife. Having estimated all the parameters of such a model (even a simple one, with only a few time-invariant covariates), the analyst can draw only a limited set of conclusions about the overall life-course process from the parameters of the hazard functions themselves. But with microsimulation, the analyst is free to compute numbers that answer questions as detailed as "what are the chances that someone who entered the labour market at age 24 is in his seventh job at age 47?" and so on.

Since microsimulation is fundamentally an exercise in sampling, it is crucial that the simulator pay attention to the issue of sampling error. A run of a microsimulation computer program produces, typically, a microdata file full of randomly assigned variable values. The values might purport to represent the situation at some future date, starting from an observed starting point for some well-defined population. If a sample of equivalent size could be drawn from the actual future population, it would be possible to proceed to compute estimated standard errors for any summary statistics based on that sample data. The same should be done if the data are simulated.

It is also important to remember, as noted above, that for each individual whose future is being simulated, the "model" (embedded in the computer program) generates a probability distribution over possible values of each variable in the future, while one run of the simulation program produces one draw from this distribution

for each person. These draws do not represent, on their own, the expected value of that person's variable, but rather a randomly-selected particular value of that variable. The value assigned may be far from the expected value but can still be "correct" (in a probabilistic sense). The expected value (for the person) may, in fact, not be an admissible value in the random-assignment algorithm. For example, the expected value of survivorship in some future year is a survival probability, whereas a run of a microsimulation program will assign to each person either a "zero" or a "one", corresponding to survivorship and non-survivorship respectively. Nonetheless, when everyone's simulated values are averaged, it is possible, in principle, to treat the resulting summary statistic as an estimate of the expected value of the variable in the population (just as it would be in a sample from a real population).

There are many sources of variability, or uncertainty, in the projections that come from a microsimulation exercise. Many people appear to think that the "Monte Carlo" variation is the primary such source of uncertainty, but it is probably not, and may even be so small as to be disregarded. The Monte Carlo variation refers to the fact that different runs of the computer program will assign different values to a given outcome for a given individual, because different random numbers (corresponding to different points on the support of the distribution of possible values of that variable) were used in different runs. Each of the different values is equally valid (conditional on the appropriateness of the model structure overall). If the exercise were repeated often enough, the set of values assigned for an individual would gradually converge to the theoretical probability distribution of the variable for that individual. The average of all those values is probably a good estimate of the expected value of that person's value. And, the sample average of those expected values is probably a good estimate of the population mean. But, it may not be necessary to run the computer program many times, or, possibly, more than once, since the "sampling error" present in one person's stochastic assignment is balanced by an offsetting error in some other person's assignment, and so on. This is the same reasoning that is used to develop an intuitive understanding of the central limit theorem. A more likely limiting factor than the Monte Carlo variation in determining the sampling error in summary statistics based on simulated data is classical sampling error in the data file that represents the starting point for the microsimulation. Other important sources of variation in simulated microdata are item and unit imputation error in the starting population, and sampling errors associated with the parameters of equations that make up the projection model. Efforts to quantify this uncertainty is an area of current research.

Despite the relatively undeveloped nature of our knowledge about dealing with microsimulation uncertainty, the preceding considerations lead me to two views that are sharply at odds with some current practice. First, the practice of "calibrating" microsimulation programs is dubious. This practice tends to consist of imposing ad hoc ex post adjustments (such as adding or subtracting constants from the intercepts of regression equations) in order to ensure that summary statistics from microsimulations match some "target", which is typically a number from someone else's projection. There is no reason to expect even a "perfect"

model to generate a sample whose sample mean is exactly equal to its expected value in the population. Simple random sampling produces summary statistics to which are attached standard errors. Attempts should be made to compute the standard error of any sample statistic that is based on simulated microdata, as well as a confidence interval (chosen to represent a pre-specified confidence level), in order to determine whether the simulated summary statistic's confidence interval covers the target value. Its failure to do so signals an inadequacy in the model, and may suggest the need for respecification or re-estimation of the model.

Secondly, "variance reduction" techniques are inappropriate in the context of stochastic simulation models. Some practitioners of demographic microsimulation (e.g., van Imhoff and Post, 1998) advocate the use of variance-reduction techniques. The author questions this view, for two reasons. First of all, it appears to imply that for each person we want to ensure that the value randomly assigned to him or her is close to its expected value. But as noted above, one run of the simulation program produces not an estimate of a person's expected value but a draw from the full relative frequency distribution for that variable. Limiting the range of values potentially assigned to an individual might not bring the assigned value closer to its expectation. More importantly, if it is agreed that the computation of standard errors is important, then bias should not be introduced into the estimate of the population variance of the variable in question. If and asterisk (*****) is used to denote a simulated value, then by analogy to usual sampling theory, the standard error of a simulated sample mean is

$$\mathrm{SE}\left(\overline{Y}^{*}\right) = \frac{\sigma^{*}_{Y}}{\sqrt{N}},$$

in which the unknown population standard deviation for the variable has been replaced by its (simulated) sample counterpart, and N is the number of individuals simulated (not the number of repetitions of the simulation algorithm). Variance reduction techniques, applied to individual values of Y, will inappropriately shrink the estimate of the population variance, and lead to downward-biased estimates of standard errors.

## REFERENCES


Aykan, Hakan, and Douglas A. Wolf (2000). Traditionality, modernity, and household composition: parent-child co-residence in contemporary Turkey. *Research on Aging*, vol. 22, No. 4 (July), pp. 395-421.

Hammel, E. A., K. W. Wachter, and C. K. McDaniel (1981). The kin of the aged in AD 2000. In *Aging: Social Change*, S. Kiesler, J. Morgan and V. Oppenheimer, eds. New York: Academic Press, pp. 11-40.

Heckman, James J., and James. R. Walker (1987). Using goodness of fit and other criteria to choose among competing duration models: a case study of Hutterite data. In *Sociological Methodology 1987*, Clifford C. Clogg, ed. Washington, D. C.: American Sociological Association.

Van Imhoff, Evert, and Wendy Post (1998). Microsimulation methods for population projection. *Population: An English Selection*, vol. 10, No. 1, pp. 97-138.

Vaupel, J. W., K. Manton and E. Stallard (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, vol. 16 (August), pp. 439-454.

Vaupel, James W., and Anatoli I. Yashin (1985). The deviant dynamics of death in heterogeneous populations. In *Sociological Methodology 1985*, Nancy Brandon Tuma, ed. San Francisco: Jossey Bass Publishers, pp. 179-211.

Wachter, Kenneth W. (1997). Kinship resources for elderly. *Philosophical Transactions of the Royal Society of London* (series B), vol. 352, pp. 1811-1817.