Statistical inference in micro simulation models: Incorporating external information**

N. Anders Klevmarken, Department of Economics, Uppsala University,
P.O. Box 513, S-75120 UPPSALA, Sweden. Fax +46 18 471 1478,
email anders.klevmarken@nek.uu.se

Abstract In practical applications of micro simulation models very little is usually known about the properties of the simulated values. This paper argues that we need to apply the same rigorous standards for inference in micro simulation work as in scientific work generally. If not, then micro simulation models will loose in credibility. Differences between inference in static and dynamic models are noted and then the paper focuses on the estimation of behavioral parameters. There are three themes: calibration viewed as estimation subject to external constraints, piece meal vs. system-wide estimation, and simulation based estimation.

1. STATISTICAL INFERENCE IN MICRO SIMULATION MODELS: INCORPORATING EXTERNAL INFORMATION

1.1 Introduction

Inference in micro simulation models (MSM) is in principle no different from statistical inference generally, but in current practice the inference aspects have been neglected. One has been satisfied if the model runs and approximately tracks observed data. The large size of a typical MSM and the difficulties to get coherent data has made many researchers and practitioners accept ad hoc methods.

There are, however, practical problems with inference in MSM related to the large number of relations and conditions, the frequent use of nonstandard functional forms often including discontinuities, and the fact that data typically are obtained from many different sources.

Micro simulation aims at statements about the distribution of some endogenous variables (for instance, the distribution of incomes) defined on a population (for instance, the population of Swedes in a particular year), given certain policy assumptions (for instance assumptions about tax rates) and initial conditions. These initial conditions are usually given by a sample of individuals on which the MSM operates. In the simulation sample values are changed or updated, and the new sample values are used to estimate properties of the distribution of interest (for instance a total, a mean or a Gini coefficient).

A proper inference usually involves several random experiments. One is drawing the sample of initial conditions, another is the random experiment or process assumed to generate population data, and a third is the generation of random numbers in the simulation experiment. The choice of methods is also determined by the mode of inference, whether there is an inference to a finite population or a "super population".

Because the random experiments involved and the mode of inference in static micro simulation in general is different from that of dynamic micro simulation it is useful first to discuss inference in static models and then turn to dynamic models. Then follows a section on the estimation of behavioral models. Although primarily based on the "super population" thinking of dynamic models much of that to be said about incorporating external information and simulation estimation also applies to static models. The paper ends with a few concluding remarks.

1.2 Inference in static MSM

The simplest case of a static model is one without behavioral response relations. It only includes a set of deterministic rules, for instance tax and benefit rules translated into computer code. The FASIT model of Statistics Sweden is an example. Given a sample of pretax incomes it computes taxes, benefits and disposable incomes for each individual in the sample. In this case there is no model-based inference but only an inference from the sample of initial conditions (pretax incomes) to the population from which this sample was drawn. In this case an inference to the finite population is meaningful and usually also desired.

** This is a shortened version of Klevmarken[1998]

If the sample of initial conditions is a probability sample this would seem to be a standard application of sampling theory. But usually the sample was drawn from a population dated a few years ago while an inference is desired to a population, which is present today, and in general these two populations differ. This problem is usually handled by reweighting. The sample weights are adjusted such that a standard inference will reproduce the observed distribution of certain variables in the present population. One might, for instance, know that the age distribution and the distribution of schooling have changed and then seek to adjust the sampling weights accordingly. A technical approach to achieve this is calibration, see Merz[1993, 1994] and Lindström[1997]. The idea is to obtain new weights, which are so close to the old ones as possible, but make the simulated values aggregate to know totals. Closeness is defined by some measure of distance. The choice of distance measure is rather arbitrary, but is has been shown that certain distance functions give estimators which are well-known in the sampling literature [Deville, J.C. And Särndal, C.E. ,1992, and Lundström, 1997]. Although calibration estimators aggregate to known totals this is no guarantee one obtains an inference to the desired population. The problem might remain if the knowledge of totals does not include the key variables of interest, or if not only the center of location but also the dispersion of key variables have changed. Without a thorough analysis of the causes to population changes any reweighting become ad hoc. Formulating a model, which captures causal relations, on the other hand leads into a dynamic MSM.

Static MSM:s can also include behavioral relations, for instance, labor supply as a function of the budget set (incomes, taxes and benefits). A static model (in the usual economic sense) has no time dimension, but in practice a micro simulation analyst wants to say something about a population in real time. It follows from the tax and benefit rules that a static tax-benefit model without behavioral relations gives the immediate, first-order effects of tax and benefit changes, but if behavioral relations are included there is an issue about their interpretation. Does a labor supply relation, for instance, give the behavioral response which materialize within a year, or does it give the total accumulated effect until some steady state is reached? Most economists probably think of static models in the latter sense. But, this raises new issues. To test and estimate such a model one needs a sample of individuals who have all reached a steady state. Is the adjustment process so quick that a random cross-section of individuals is suitable for inference?

Let us assume that this is the case. An inference would then have to account both for the random uncertainty, which arises because the model is simulated on a sample of initial conditions, and the uncertainty which is generated by the estimated behavioral model. The latter will include two components. The first arises because the unknown parameters are estimated. The properties of these estimates depend on the properties of the model, how data were obtained and on what estimation method was used. The second component arises because invoking a random number generator simulates the estimated model. The properties of simulations from a static model and how to estimate various variance components is discussed in some detail in Klevmarken[1998]. It is a problem that we do not know and cannot simulate the values of the exogenous variables of the observations not included in the sample. An inference has to be conditioned on the observed exogenous variables in the sample. If estimators can be written as sums of individual contributions then Horvitz-Thompson estimators are consistent, but for more general parameters there might not be any finite sample estimator. See also Pudney & Sutherland[1996].

1.3 Inference in dynamic MSM.

In a dynamic MSM there is no constant population to which an inference can be drawn, because the model defines how the population changes both in size and in composition. Only an inference to the super-population defined by the model would seem meaningful. Let's write the model in the following way,

$$y_t = g(y_{t-1},\dots y_{t-s} \mid y_0, \varepsilon_t, \theta); \qquad (11)$$

where y_0 is a vector of initial conditions, in practice set by the sample on which the simulations are done. Suppose we are interested in estimating

$$E_{\varepsilon_1,\dots\varepsilon_t,y_0}(\mu(y_t)) = \int\dots\int\int \mu f(\varepsilon_1)\dots f(\varepsilon_t) f(y_0) d\varepsilon_1 \dots d\varepsilon_t dy_0; \qquad (12)$$

It is assumed that $\varepsilon_s, \varepsilon_t$ and y_0 are independent for all s≠t and that f is known. $\mu(y_t)$ is a statistic of interest. The distribution of y_0 will in practice become estimated by the corresponding empirical distribution function obtained through the sample of initial conditions. One could either condition on the sample of initial conditions, if t is large and the model has some ergodic properties the influence of y_0 becomes small, or one could use the bootstrap technique to evaluate the random influence from the choice of initial conditions sample.

In general it will not be possible to evaluate expression (12) analytically, but by replicated drawings from the distribution of ε a number of replications of μ is obtained, and the mean of these μ-values is an unbiased estimate of expression (12). This procedure assumes that the parameters θ are known. In practice they are not and have to be replaced by some estimates $\hat{\theta}$. This implies that the simulated estimate of (12) is a random function of $\hat{\theta}$. If μ and g satisfy certain regularity conditions[1] and if $\hat{\theta}$ is a consistent estimate of θ, then the estimate of $E(\mu)$ is consistent too. But even if $\hat{\theta}$ is unbiased, the estimate of $E(\mu)$ is in general not unbiased, because μ and g are in general nonlinear functions. By replicating also over the domain of $\hat{\theta}$ we might thus like to estimate $E_{\hat{\theta}}E_{\varepsilon_1,\ldots\varepsilon_t,y_0}(\mu(y_t \mid \hat{\theta}))$. These replicated simulations will also give an estimate of the corresponding variance.

1.4 Estimation of parameters in behavioral models

1.4.1 Model alignment using external information – an estimation problem

Our limited capacity as model builders, the difficulties to get good comprehensive data from which the model parameters can be estimated, and the piece meal approach usually adopted in practice to estimate the model sub-model by sub-model, all contribute to deviations of simulated values and distributions from observed data. To make the model "stay on track" some model builders have aligned their models to external benchmark data. Population totals and means from official statistics or estimates from surveys not used to estimate the model are sometimes used as benchmarks. If a model is to gain credibility with users they often require that the model is able to reproduce the basic demographic structure of the population and predict well-known benchmarks like for instance, the labor force participation rate, the unemployment rate, the mean and dispersion of disposable income, etc. For this reason model builders have forced their models to predict these numbers without error. In the U.S. model CORESIM, for instance, adjusting the simulated values (and not the parameter estimates) does this alignment.

Alignment is usually done by simple proportional adjustments, but there are also more sophisticated procedures. The ADJUST procedure developed by Merz[1993, 1994b] and originally designed for reweighting in static models (cf. above) might also be used for alignment. However, in this context the whole approach appears even more ad hoc than when it is used for reweighting. It does not consider the stochastic properties of the model at all.

A natural way to incorporate this kind of externally given information is to look upon the estimation problem as one of constrained estimation. This approach is developed in some detail in Klevmarken[1998]. The following conclusions can be drawn: Alignment should in general not be done with simple proportional alignment factors, but each individual gets its own alignment factor. Also, in a model with more than one endogenous variable a constraint which applies to one variable will in general not only imply an alignment of that particular variable but also of all other variables. Furthermore, in nonlinear models there will in general not exist as simple alignment factors as in the linear case.

More or less explicitly the discussion above was based on the assumption that the sample used in the simulations had a size sufficiently large to justify the treatment of external data as exact constraints. If this is not the case one might not like the simulated total (mean) to equal the external total (mean) exactly but allow for the built in stochastic variation in the model. If the external data are estimates rather than population parameters then that is another reason not to enforce an exact equality. A natural approach to incorporate uncertain external information is that of mixed estimation, a technique, which is well developed for linear models in many textbooks, but less developed for nonlinear models.

1.4.2 Model-wide or piece meal estimation

Given the complexity and mixture of model types and functional forms in a large MSM its parameters are usually estimated in a piecemeal way, sub-model by sub-model. This is sometimes necessary because one does not always have access to one large sample including all variables, but have to use several samples

[1] MSM often include discontinuities which could imply that these regularity conditions do not hold, but models are more likely to be continuos in the behavioral parameters θ than in variables, the values of which are determined by legislation and government rules.

collected from different sources. None-the-less the piece meal approach may be inappropriate. It depends on the model structure. If the model has an hierarchical or a recursive structure and if the stochastic structure impose independence or lack of correlation between model blocks or sub-models, then a piece meal approach can be justified (cf. the discussion in Klevmarken,1997).

By way of an example consider the following simple two-equation model:

$$y_{1t} = \beta_1 x_t + \varepsilon_{1t}; \qquad E(\varepsilon_i \varepsilon_j) = \begin{cases} \sigma_1^2 & \text{if } i=j=1. \\ \sigma_2^2 & \text{if } i=j=2. \\ 0 & \text{if } i \neq j. \end{cases}$$
$$y_{2t} = \beta_2 y_{1t} + \varepsilon_{2t}; \tag{23}$$

This is a recursive model and it is well-known that OLS applied to each equation separately will give consistent estimates of β_1 and β_2. The estimate of β_1 gives the BLUP $\hat{y}_1 = \hat{\beta}_1 x_t$ while predictions of y_2 outside the sample range are $\hat{\beta}_2 \hat{y}_1$. However, this suggests the following model-wide criterion,

$$\frac{1}{\sigma_1^2} \sum_t (y_{1t} - \hat{y}_{1t})^2 + \frac{1}{\sigma_2^2} \sum_t (y_{2t} - \hat{\beta}_2 \hat{y}_{1t})^2; \tag{24}$$

Minimizing this criterion with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$ yields the OLS estimator for β_1 but the following estimator for β_2,

$$\hat{\beta}_2 = \sum_t y_{2t} x_t \,/ \sum_t y_{1t} x_t \tag{25}$$

In this case both the "piece meal" OLS estimator of β_2 and the "system-wide" instrumental variable estimator (25) are consistent but the OLS estimator does not minimize the prediction errors as defined by (24). In fact, under the additional assumption of normal errors the estimator (25) is a maximum likelihood estimator and thus asymptotically efficient.[2]

If we would add the assumption that ε_1 and ε_2 are correlated the recursive property of the model is lost and OLS is no longer a consistent estimator of β_2. The estimator (25) is, however, still consistent and under the assumption of normality a ML estimator. In this example we would thus prefer the "system-wide" estimator (25) whether the model is recursive or not.

In applied micro simulation work it might not always be necessary to insist on efficient estimators. Usually large micro-data sets are used for estimation and then also the variance of less efficient estimators might become acceptable. In the example above we could perhaps do with the piece meal OLS estimator if the sample is large, but only if the model is recursive.

Finally, a more general comment on the choice of estimation criterion is in place. The least-squares criteria commonly used assume that we seek parameters estimates such that the *mean* predictions give the smallest possible prediction errors, eq. (24) is an example. However, in micro-simulation we are not only interested in mean predictions, but we want to simulate well the whole distribution of the target variables. This difference in focus between micro simulation and a more conventional econometric analysis might suggest a different estimation criterion. We will return to this topic in section 1.4.3.

1.4.3 Simulation-based estimation[3]

The complexity and nonlinear character of MSM and the fact that they are designed to simulate suggest that simulation-based estimation is a feasible approach to obtain system-wide estimates. A discussion of simulation-based estimation does not only lead to new estimators but also highlight the need to change the conventional estimation criteria to one, which is compatible with the simulation context. Assume the following simple model,

$$y_t = g(x_t, \varepsilon_t, \theta); \tag{26}$$

x_t is an exogenous variable, ε_t a random variable with known p.d.f. and θ an unknown parameter.

$$E(y_t \mid x_t) = E(g(x_t, \varepsilon_t, \theta_0) \mid x_t) = k(x_t, \theta_0); \tag{27}$$

We assume that $k(x_t, \theta)$ does not have a closed form.

The basic idea of estimating θ is to obtain a distribution of simulated y-values, y_t^s, with properties which as closely as possible agree with those of the p.d.f. of y_t. It would appear to be a natural approach to choose $\hat{\theta}$ such that it minimizes

[2] The estimator (25) is a ML estimator because there is no additional x-regressor in the second relation. The reduced form becomes a SURE system with the same explanatory variable in both equations. In general the ML estimator will depend on the structure of the covariance matrix of the errors.

[3] This section relies to a large extent on Gourieroux & Monfort[1996]

$$\sum_{t=1}^{n}(y_t - y_t^S(\hat{\theta}))^2; \qquad (28)$$

However, as shown by Gourieroux & Monfort[1996] p. 20 this "path calibrated" estimator is not necessarily consistent. To see this consider an example which differs a little from the one used in Gourieroux & Monfort[1996]. Assume the following simple model,

$$y = \beta x + \sigma\varepsilon; \text{ where } \varepsilon \sim IID(0,1) \qquad (29)$$

We seek parameter estimates $\tilde{\beta}$ and $\tilde{\sigma}$ such that the model can be simulated,

$$y^S = \tilde{\beta}x + \tilde{\sigma}\varepsilon^S; \qquad (30)$$

where ε^S are draws independently of ε but from the same (known) distribution. Inserting (30) into (28) and solving the first order conditions gives the following estimators,

$$\tilde{\beta} = (\sum xy - \tilde{\sigma}\sum x\varepsilon^S) / \sum x^2; \qquad (31)$$

$$\tilde{\sigma} = \frac{\sum(\beta x + \sigma\varepsilon)x\sum x\varepsilon^S - \sum \beta x + \sigma\varepsilon)\varepsilon^S \sum x^2}{(\sum x\varepsilon^S)^2 - \sum(\varepsilon^S)^2 \sum x^2}; \qquad (32)$$

From the assumptions made and the additional assumption that $(1/n)\Sigma x^2$ converges to a finite limit when n tends towards infinity, it follows that

$$\operatorname*{plim}_{n \to \infty} \tilde{\sigma} = 0; \text{ and } \operatorname*{plim}_{n \to \infty} \tilde{\beta} = \beta; \qquad (33)$$

Using this criterion we thus get an inconsistent estimate of σ but a consistent estimate of β. Essentially this estimator tells us to ignore the random drawings of ε^s when we simulate, i.e. only to use mean predictions. As already noted, such a procedure does not agree with the objective of micro-simulation. In this particular model the estimate of β is consistent, but if there was a functional relation between β and σ then the slope would also become inconsistently estimated. It is perhaps possible to generalize this result and suggest that if there is any functional relation between the parameters, which determine the mean path and those determining the dispersion around this path in a micro simulation model, then one cannot use a path-calibrated estimator.

An alternative approach is to use a "moment calibrated" estimator, which minimizes the distance between observed and simulated moments.

Let θ be a vector of size p and x_t a vector of size r. Furthermore let $K(y_t, x_t)$ be a vector function of size q, and

$$E(K(y_t, x_t) \mid x_t, \theta_0) = k(x_t, \theta_0); \qquad (34)$$

K could, for instance be the identity function and the square of y_t. Also define a r x q matrix $Z_t = I_{q \bullet q} \otimes x_t$. From the exogeneity of x_t it follows that

$$E\left[Z_t(K(y_t, x_t) - k(x_t, \theta_0))\right] = 0; \qquad (35)$$

Because there is no closed form of $k(x_t, \theta)$, we will define an unbiased simulator of k,

$$\tilde{k}(x_t, \varepsilon^S, \theta) = \frac{1}{S}\sum_{s=1}^{S} K(g(x_t, \varepsilon_t^s, \theta), x_t); \qquad (36)$$

where ε^S is vector of S independent random errors ε_t^s drawn from the p.d.f. of ε_t.

A simulated GMM estimator is then obtained as,

$$\hat{\theta} = \arg\min_{\theta}\left(\sum_{t=1}^{n} Z_t\left[K(y_t, x_t) - \tilde{k}(x_t, \varepsilon^S, \theta)\right]\right)' \Omega\left(\sum_{t=1}^{n} Z_t\left[K(y_t, x_t) - \tilde{k}(x_t, \varepsilon^S, \theta)\right]\right); \qquad (37)$$

where Ω is a r x r symmetric positive semi-definite matrix. As shown in Gourieroux & Monfort[1996] this is a consistent estimator. The covariance matrix of the estimator has two components, one, which is the covariance matrix of the ordinary GMM estimator, and one, which depends on how well k is simulated. An optimal choice of Ω depends on the unknown distribution of y_t. Gourieroux & Monfort[1996] p. 32 gives a simulation estimator of the optimal Ω. Two observations are in place:

- The number of moment conditions (35) invoked must be no less than the number of unknown parameters, otherwise the model becomes unidentified.
- The quadratic expression in (37) can be minimized using the usual gradient based methods if first and second order derivatives with respect to θ exist. If the model includes discontinuities in θ one would have to rely on methods not using gradients. MSM which include tax and benefit legislation typically

have discontinuities in variables, which may or may not imply discontinuities with respect to behavioral parameters.

It should be possible to include constraints of the kind discussed in the previous section in the simulation-based approach. Suppose K is the identity function in y_t so the moment condition becomes,

$$E(y_t - E(g(x_t, \varepsilon_t, \theta_0)) = 0; \qquad (38)$$

The empirical correspondence to the expression to the left of the equality sign is

$$\bar{y} - \frac{1}{n}\sum_{t=1}^{n}\tilde{k}(x_t, \varepsilon^S, \theta_0); \qquad (39)$$

Suppose now that we know the finite sample mean $\bar{Y}$. How could we use this information? If we also knew the x_t values for all individuals in the finite sample, we could substitute $\bar{y}$ in (39) for $\bar{Y}$ and extend the summation in the second term of (39) to N, and thus get an empirical correspondence to (38) for the whole finite population. In practice this is of course not possible. One only knows the x-observations of the sample, but with know selection probabilities p_t they can be used to compute the following estimate,

$$\bar{Y} - \frac{1}{N}\sum\frac{1}{p_t}\tilde{k}(x_t, \varepsilon^S, \theta); \qquad (40)$$

The covariance matrix of the resulting estimate $\hat{\theta}$ should now have a third component, which reflects the sampling from the finite population. (A numerical example is provided in Klevmarken [1998].)

1.6 Concluding remarks

The credibility of micro simulation models with the research community as well as with users will in the long run depend on the application of sound principles of inference in the estimation, testing and validation of these models. This paper has reviewed a few issues of inference in static and dynamic micro simulation models.

The application of a model-wide estimation criterion will in general suggest an estimator which does not permit a piece meal estimation of sub-model by sub-model. Only if the model has a hierarchical or recursive structure it is possible to use a piece meal approach.

It was also suggested that the alignment procedures now used in practical work could be seen as part of the estimation procedure. Although not discussed above it is important to note that the constraints imposed by alignment must be tested if accepted by data. If they are not that is a clear indication that something is wrong with the model, and it should be reformulated rather than forced "on track" by alignment.

Finally it was also suggested that the simulation approach to estimation could be useful in micro simulation work. These models are designed to simulate and they also frequently include nonlinear and complex relations, which suggests that simulation-based estimation has a relative advantage. However, path-calibrated estimates are in general inconsistent and should be avoided, in particular in a micro simulation context which does not only focus on mean relations. A better alternative is moment-calibrated estimates.

1.7 References

Deville, J.C. and Särndal, C.E., Calibration estimators in survey sampling, *Journal of American Statistical Association* 87, 376-382, 1992.

Gouriéroux, C. And A. Monfort , *Simulation-Based Econometric Methods*, Oxford University Press, 1996.

Klevmarken, N.A. , Behavioral Modeling in Micro Simulation Models. A Survey. Working Paper 1997:31, Department of Economics, Uppsala University, 1997.

Klevmarken, N.A., Statistical inference in Micro Simulation Models: Incorporating External Information, Working Paper 1998:20, Department of Economics, Uppsala University, 1998.

Lindström, K., Utvärdering av framskrivnings-tekniken i FASIT-modellen (Evaluation of the calibration technique used in the FASIT model), memo Statistics Sweden, 1997.

Lundström, S., *Calibration as a standard method for treatment of nonresponse*, Department of Statistics, Stockholm University (diss.), 1997.

Merz, J. ADJUST – A program package for adjustment of micro data by minimum information loss principle. Program Manual, FFB-Documentation No 1, University of Luneburg, 1993.

Merz, J., Microdata adjustment using the minimum information loss principle, Discussion Paper no 10, Forschungsinstitut Freie Berufe, Universität Luneburg, 1994.

Pudney, S. And Sutherland, H., Statistical Reliability in Microsimulation Models With Econometrically-Estimated Behavioural Responses, Chapt. 21 in A. Harding (ed.), *Microsimulation and Public Policy*, North-Holland Elsevier, 1996.