2012-11-14

SESIM

III

- A Swedish dynamic micro simulation model

By

Lennart Flood, Fredrik Jansson, Thomas Pettersson, Tomas Pettersson, Olle Sundberg & Anna Westerberg

CONTENTS

1. Introduction

Micro simulation modelling (MSM) as an analysis tool in social science was introduced already in the 1950s, although, as most computer and data intensive methods, its usage became widespread much later.[1] The main purpose of micro-simulation is to model and simulate the distribution and not only the mean values of the variable of interest. One of its main advantages is that it permits heterogeneous behaviour: every individual or household is not assumed to behave as the average economic agent. According to Klevmarken [1997] this "widens the scope of micro-simulation beyond that of conventional econometric modelling. When economic relations are highly nonlinear, when tax laws and rules of transfer programs introduce censoring and truncation and when sub-populations differ in behaviour, then models of average behaviour become inadequate to evaluate the average impact of policy changes, while a micro-simulation model can be used also for this purpose."

Since MSM are designed to study questions about distributional effects of changes in tax and benefit systems, they obviously have a potential as a tool for policy analysis. If the models are used for forecasts, they can generate profiles or life cycle paths of individuals. This is particularly important for the full distributional impact of some long-term policies, such as the pension or educational systems, whose full effects take a considerable amount of time to become visible.

MSM can be classified according to a large number of characteristics, see Merz [1991], from completely static to fully dynamic life-cycle models. FASIT is one example of a static tax-benefit model developed by Statistics Sweden (SCB) and used at the Ministry of Finance in Sweden. Such models do not attempt to incorporate behavioural change, and are used to calculate the immediate impact of institutional changes in the tax and benefit system.[2] In principle this class of models is only a detailed description of the tax and benefit system, although some behavioural effects can be integrated. One version of FAST for instance has been extended to include labour supply responses in order to generate more realistic income predictions.

Classic static models (without labour supply modules) assume that changes in taxes, transfer payments etc. do not result in changes in hours of work. The realism of this assumption depends on the economic reform we are evaluating. The possible behavioural effects of tax reform for instance have been discussed at length without consensus. After all, the assumption of no changes in labour supply may be realistic. For other kinds of reforms this assumption seems less reasonable. A dynamic micro simulation model with behavioural relations allows the individuals to adjust as a result of changes in the economic environment, which seems more realistic. The really difficult question when incorporating behavioural response is the size and perhaps also the direction of these changes.

The main reason for focusing on dynamic models is that they allow for an evaluation of the long run effects of a policy change. This is the reason for developing SESIM, as a useful complement to the static FASIT-model. Dynamic models are designed to incorporate behavioural response as

---

[1] For recent surveys of dynamic MSM in economics se e.g. Merz [1991], Klevmarken [1997], Zaidi & Rake Rake & Zaidi [2001] and O´Donoghue [2001].

[2] For a survey of static micro simulation models in Europe, see Sutherland [1995]

well as simulating the policy environment in the long-run. Adding a dynamic element into a MSM requires modelling changes in characteristics or behaviour at the individual unit level. These changes are commonly referred to as the 'ageing' of an individual unit. There are, in fact, two approaches to such ageing – static and dynamic. Static implies re-weighting of the micro database in such a way that the characteristics of the model individuals are aligned with an external data source (without modelling any behavioural change). Many static models include some form of static aging. Dynamic aging, on the other hand, simulates the attributes of each person at time t+1 using the attributes at time t. This is typically accomplished using a behavioural equation and a Monte Carlo process. Thus, to model for instance participation in the labour force, we will first estimate an econometric model (using e.g. a logit model) and calculate the probability of labour force participation rate. Next, we draw a uniformly distributed (0, 1) random number. If this number is smaller than the estimated probability of labour force participation, we assign labour force participation to that individual; otherwise he or she is assigned to be out of the labour force.

Most dynamic MSM use discrete ageing, i.e. the relevant relations are updated once a year. The choice of discrete or continuous ageing is dependent on the data sources available for estimation and calibration. Many of the processes that have been modelled are probably best described in continuous time, e.g. length of an unemployment spell, and a discrete one-year event must be considered as a (crude) approximation. For a discussion about choice of frequency in a dynamic MSM, see Galler [1997].

Given a choice of frequency, the next choice is to decide a sequence of events. A recursive structure is chosen such that events that happen during the end of the year only can use information of events that have happened earlier during the same year. For example only women who survive can give birth to a child. Thus, mortality should come before fertility. A typical structure in the sequence is to have basic demographic events as the first events. It is important that the statistical models have been estimated having this structure in mind, i.e. that they condition on the same risk-population.

An additional distinction needs to be drawn between deterministic and stochastic processes. In a deterministic model, relationships are fully determined by the parameters defined within the model. A stochastic model, on the other hand, incorporates random processes, either to reflect the random nature of the underlying relationship or to account for random influences due to incomplete model specification. Most dynamic MSM in the social policy make use of a combination of stochastic and deterministic simulation processes.

In a dynamic MSM exogenous inputs are typically used in order to characterize the macroeconomic development. A MSM is not an equilibrium model, where markets are modelled. Normally the supply side is modelled, e.g. labour supply, but not the demand side, e.g. the demand for labour. An implicit assumption is that the supply side is the important one in the long run. To align MSM results to external forecasts, for instance to the aggregated level of employment, calibration is then used.

A related issue is whether the dynamic MSM should be used in *steady state*, or not. Steady state means that the processes that are simulated are assumed to continue in the same way as the preceding period. Thus, the relations that persist during the estimation period are assumed to persist forever. For models not simulated in a steady state different exogenous sources of information are used, for instance information about future growth rate, size of the labour force and population growth. However the behavioural estimations used are always assumed to

express steady state relations between the underlying variables.

This documentation is planned as follows. First we give a general description of the structure of SESIM. Next the data sources used are presented and important adjustments of these data are discussed. Next stochastic simulation is discussed; special attention is given on stochastic modules or models for regional mobility, retirement decision, generation of earnings, financial and real wealth, non-cash benefits and health and care and of elderly. The final part presents the technical platform and the user interface. In the Appendix, a list of all stochastic models in SESIM is presented.

2 SESIM structure

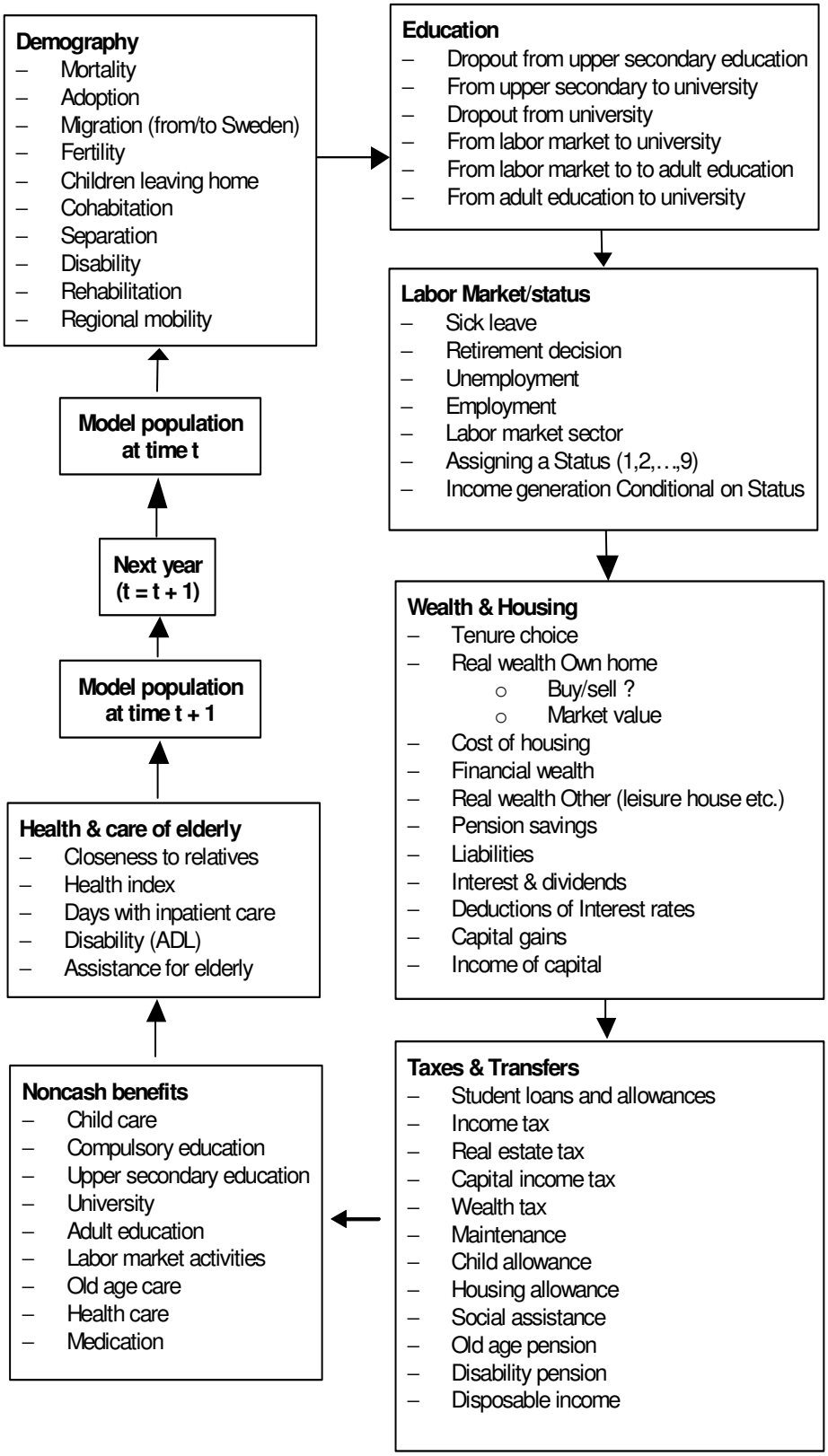
The development of SESIM started in 1997 as a tool to assess the Swedish education financing system. Part of that work has been documented in Ericson and Hussénius, [2000]. We refer to this as version I of SESIM. Since year 2000 the focus has shifted from education to pensions. A major purpose of SESIM has been to evaluate the financial sustainability of the new Swedish pension system. This new focus has also implied that SESIM has been developed into a general MSM that can be used for a broad set of analyses. This version is the second version of SESIM and is documented in Flood et.al [2003]. The present version, SESIM III, is used for several purposes, the most important of which is still pensions, but it extends to the analysis of health issues amongst elderly.

SESIM is a mainstream dynamic MSM in the sense that the variables (events) are updated sequentially, and the space in time between the updating processes is a year. The start year is 1999 and every individual included in the initial sample (≈300 000) then goes through a large number of events, reflecting real life phenomena, like education, marriage, having children, working, retiring etc. Every year individuals are assigned a status, reflecting their main occupation during the year. Every status is related to a source of income: working gives earnings, retirement's gives pensions etc. The tax and benefit system is then applied and after-tax income is thus calculated. If this simulation is repeated over a long time individual life-cycle income can be generated.

The sequential structure in SESIM is presented in Figure 2.1. The first part consists of a sequence of demographic modules (mortality, adoption, migration, household formation and dissolution, disability pension, rehabilitation and regional mobility. After that comes a module for education (compulsory school, high School (Gymnasium), municipal adult education (Komvux) and university. Next module deals with the labour market including the retirement decision. The date of retirement can be decided according to a retirement model, see Eklöf & Hallberg [2004], but it is also possible to choose a specific age (it is also possible to allow for some variation around this age).

The labour market module also includes a model for sick leave, Bolin & Lindgren [2005], unemployment, employment and a model for imputation of labour market sector. The sector is required for calculations of occupational pensions. In SESIM, we have implemented the rules for occupational pensions as well as the choice of labour market sector. We also allow for change of sector and the occupational pension is then adjusted in accordance to the new rules for occupational pensions in that sector.

Figure 2.1. Structure of SESIM



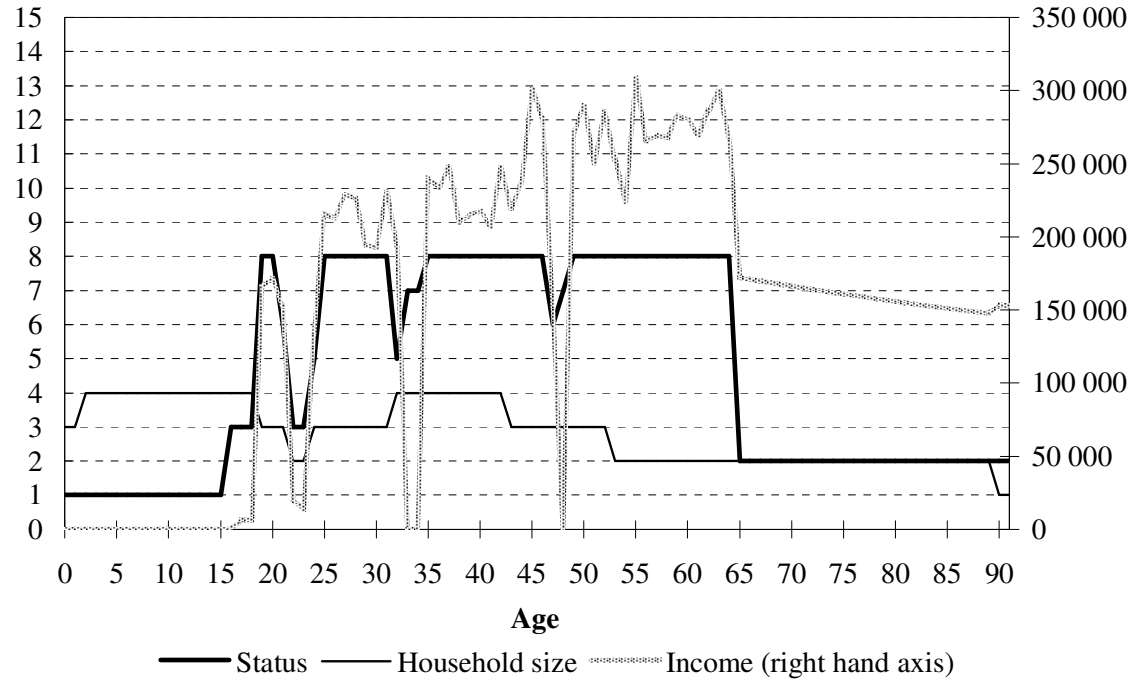
The first step is to decide a status for each individual. There are nine different statuses that reflect individuals' main occupation during the year. Note that each individual can only have one status each year (the status emigrated is an exception). The different statuses are given below:

1. Child (0-15 years old)
2. Old age pensioner: individuals with income from old age pension
3. Student: individuals who study at gymnasium, adult education or university
4. Disabled: individuals with income from disability/sickness benefit
5. Parental leave: women who give birth during the year
6. Unemployed: individuals with income from unemployment insurance or from labour market training
7. Miscellaneous
8. Employed: market work
9. Emigrated: individuals living abroad with Swedish pension rights. Note that this classification is not unique since they also can have income from early retirement or old age retirement.

Given the status the next step is to generate the individual's income. For status 8 (employed) an earnings equation is used to determine income. For other kind of statuses, e.g. unemployment, different rules can be applied to impute an income. After calculation of income, a module for wealth capital income and housing is entered. Since it is rather unusual to include formation of wealth in a MSM, we give a more detailed description below. After wealth/housing a large module describes all relevant tax, transfer and pension rules. For the old age pension system, the rules for public and occupational pension have been implemented in all relevant details. Given all information above, the household disposable income can be defined. Next, a module for public consumption is entered; the details are discussed in Pettersson & Pettersson [2003]. The final module reflects an important update in SESIM III, the health module. In this module the need for care is imputed. In order to assess the importance of relatives as a resource we impute the geographical distance to relatives. The health status is calculated next and then days with inpatient care followed by severe disability. Finally assistance for elderly is imputed.

An important factor in most analyses is the household composition. In SESIM the model population lives in households. Like in reality, the household composition can change; new households are formed and they can split. In the base population households are based on real observed households (with some modification). Information about the household composition is important since many stochastic models use household information Moreover many benefit system are also based on household information, e.g. social assistance, housing- and child allowance. Figure 2.2, display, as an illustration, a SESIM created biography. Size of household, income and status are displayed each year for a woman.

Figure 2.2. A SESIM created life path from the cradle to the grave



After having been classified as a child up to 15 years of age, she studies between 16 and 24 years of age. There is a short interruption in studies during this period while she is working. At age 22 she moves from her parents and form a new household together with a husband. At 24 she gives birth to her first child, therefore she is on parental leave, but after this she returns to work. Her second child is born when she is 32 years old. Two years thereafter she is classified as ”other”, which can be interpreted as a housewife. After that she returns to work, with some exceptions, until she retires at 65 years of age. Also displayed in the figure is her earnings profile. There is an increasing trend, but a considerable variation, for instance in those years when she is a student and a ”housewife” her income is quite low. After retirement her income is given by the pension rules, and for this reason there is no volatility anymore. At the age of 43 the first child leaves the household and ten years later the other. Her husband dies when she is 90, two years before herself.

During a lifetime a large amount of information is generated for each individual in the model population. Totally about 300 variables are stored. In Figure 2.2 three of them were displayed.

3 Data sources

This section gives a brief description of the main data sources for estimation and construction of the model population. We also discuss some corrections or adjustments that have been done.

3.1 LINDA – a panel data base[3]

LINDA is the main source of information used in SESIM. This data covers about 3.5 percent of the Swedish population. For year 1999 this implies 308 000 randomly selected individuals. To these selected individuals all household members have been matched. In total the sample size is 786 000 individuals in 1999. This is the primary database for SESIM, used for estimation of statistical models as well as construction of the base population.

The selected individuals are followed backward and forward and all relevant information is collected. Some information, for instance pension rights, can be traced back as long as to 1960. Thus this is a panel data since the same individuals are followed over time. Selected individuals, who in a certain year disappear from the data, by death or emigration, are replaced by newly selected individuals in such a way that statistical representativeness is maintained.

Note that the database is completely created from administrative registers. Thus no interview is needed and therefore a major advantage is that there are no problems of attrition bias. The database has been created by merging information from a large number of registers: Income- and wealth, earnings, pension rights, sickness- and unemployment benefit and schooling.

The base population used in SESIM is formed by a random draw of 104 000 individuals from LINDA. To this sample 8 000 individuals have been added from the National Social Insurance Board register for pensions rights. This additional sample includes individuals living outside Sweden, but with Swedish pension rights.

In construction of the base population in SESIM two main adjustments have been done in order to obtain a model population consistent with the definitions used in SESIM. This is described below.

3.2 Other data sources

Apart from LINDA, there are some additional data used for estimation or imputation: HEK, GEOSWEDE, Kungsholmen study and ULF. The HEK[4] is a survey based on interviews, merged with register information, and therefore can better identify a household according to an economic definition. This is in contrast to LINDA, which essentially is based on a tax definition of the household. Apart from being used for correcting the household composition, HEK has also been used in the estimation of public consumption as well as housing location and cost of housing.

Regional mobility and tenure choice is based on GEOSWEDE.[5] This database is constructed from Louise and RTB: GEOSWEDE have been used in models for regional mobility and health.

---

[3] Longitudinal Individual Data for Sweden. For a documentation see Edin and Fredriksson [2000].

[4] SCB income distribution survey, a yearly survey including a sample of about 30 000 individuals merged with administrative data.

[5] A database constructed from Louise, RTB, geography database and tax assessed values.

HUT[6] is used for calculation of indirect taxation. The Health and care module is based on data from Kungsholmen study[7] and ULF[8].

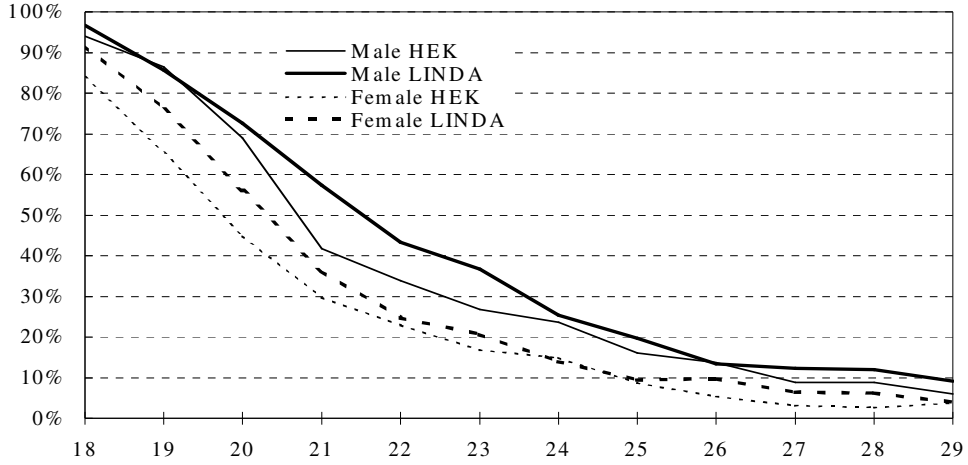
3.3 Adjustment of household definition

The household definition in LINDA is based on information from the total population register[9] and the ”generational register”. These households do not always match a meaningful economic definition. First of all individuals are assumed to live in the municipality where they are recorded according to the national tax registry, secondly adults living together without being legally married or having common children are considered as separate households. A comparison of other data sources that have better information on household composition shows that there are two problems in LINDA.[10] First, the numbers of youngsters between 18 and 29 who still live in their parent’s household are over reported. Many children who have moved are still tax registered in their old household. Secondly, the number of cohabitants without children is underestimated, especially for younger households.

Older children living with parents

The Figure below, compare the share of young individuals, 18-29, living with their parents according to LINDA and HEK in 1999.

Figure 3.1. Share of youngsters living with their parents in LINDA and HEK



Source: SCB, HEK, LINDA, Ministry of Finance.

Figure 3.1 show that the share of youngsters living with parents is overestimated in LINDA. The difference is largest for males 21-year of age, where the share from LINDA is 57 percent compared to only 42 percent in HEK. In order to correct for this bias a model has been estimated and used for prediction. The probability to move away from parents is estimated on HEK data

---

[6] SCB household expenditure survey.

[7] Mårten Lagergren [Source?]

[8] Statistics Sweden, Level of Living Survey.

[9] Total population census (RTB).

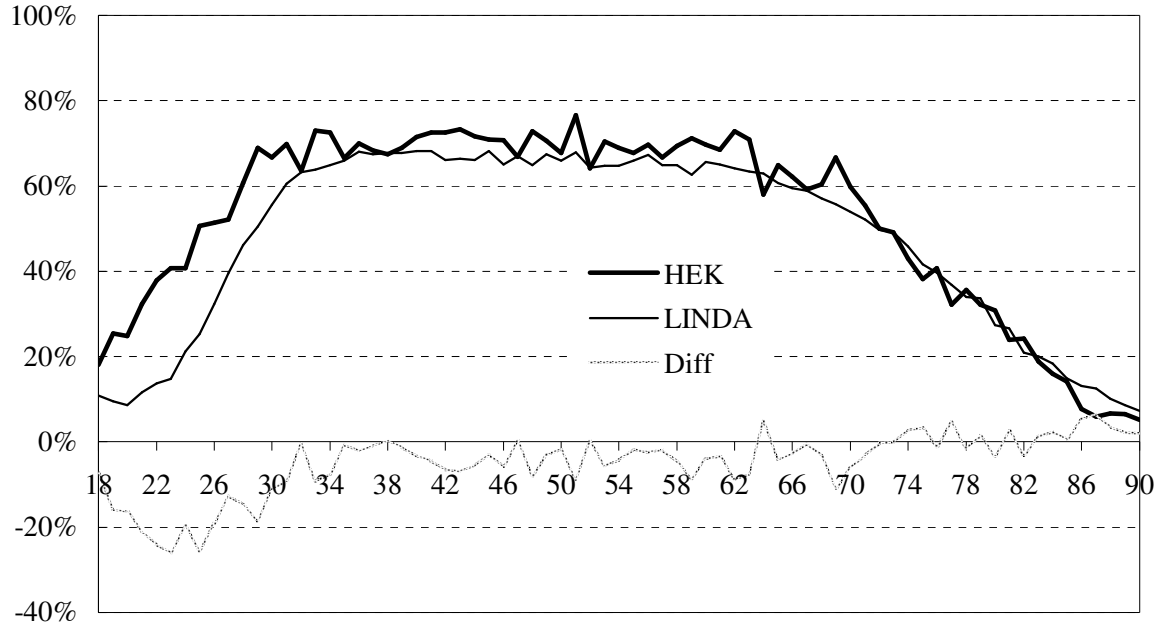
[10] LINDA is compared with the income distribution survey HEK. In HEK the definition of a household comes from interview data.

for individuals age 18-29.[11] Based on this model the probability of moving away from the parents is imputed for those 18-29 years old in LINDA. Then the individuals living together with their parents are ranked according to the predicted probability of moving. Finally the number of individuals with highest probabilities is chosen such that the HEK frequencies are matched.

Cohabitation

In figure 3.2 below the share of married or cohabiting women per age in LINDA is compared to the shares in HEK.

Figure 3.2. Comparison of married/cohabiting women in LINDA and HEK 1999.

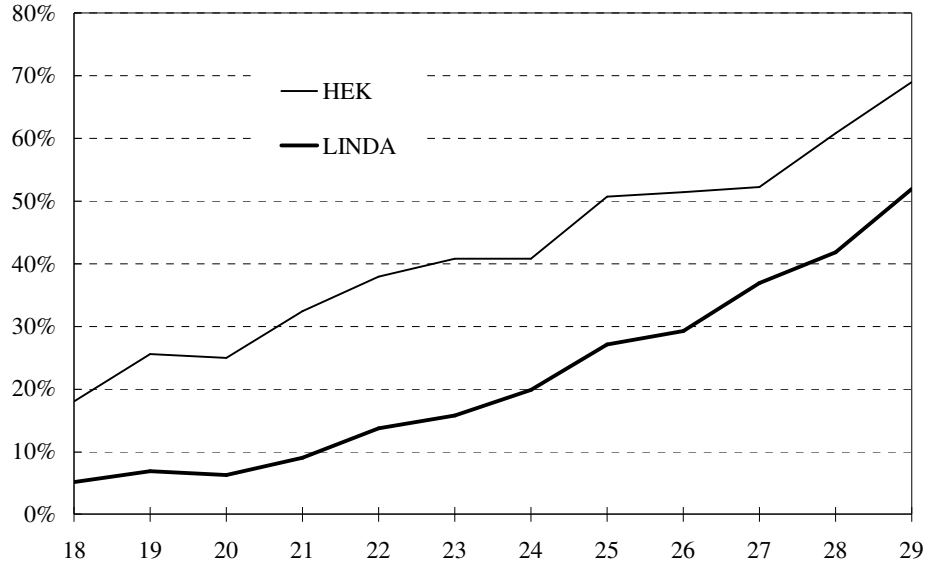


Source: SCB, HEK, LINDA, Ministry of Finance.

As expected the largest difference is found among young individuals. For this reason a correction is done only in the age group 18 to 29. When children move away from their parents, they are classified, by default, as single. This results in an even larger underestimation of non-single households. The figure below shows the share of married/cohabiting women after correcting for moving from parents.

[11] The probabilities are estimated by a logit regression. The explanatory variables are: Gender, age, square of age, born in Sweden (dummy), Study allowance / 1000, earnings / 1000, highest education (dummies compulsory school and upper secondary school), interaction age and education.

Figure 3.3. Share of married/cohabiting women in LINDA and HEK (after correction of youngsters living with parents).



Source: SCB, HEK, LINDA, Ministry of Finance.

In order to correct this bias a model for the probability of being married/cohabiting is estimated for women in the age group 18 to 29 without children in HEK.[12] Based on the estimated model the probability of cohab/marriage is calculated for 18-29 years old females without children in LINDA. The individuals are again ranked after the predicted probabilities. The X number of women with highest probabilities for a marriage/cohab is matched with a single male two years older and the same level of education. X is chosen so that the imputed frequencies coincide with these from HEK-data.

3.4 Adding emigrants with pension rights.

Individuals with Swedish pension rights will keep these entitlements regardless of in which country they live, and after retirement their pension is paid out regardless of whether they live in or outside Sweden. Thus in order not to underestimate the cost of yearly pensions all individuals with pension rights must be considered.

Since LINDA does not include individuals living outside Sweden, other data sources have been used. Information comes from a sample from the National Social Insurance Board pension point's database. These data include all individuals with Swedish pension rights regardless of where they live. However, in order to merge these data some adjustments have been made. First, the additional sample is from 1995 and the LINDA sample from 1999. In order to correct for this the pension rights for the emigrants are moved forward four years. Second, the emigration sample is individual based and no household information is available. In order to correct for this a household composition is imputed from HEK. Since large shares of individuals who have emigrated also are born outside Sweden, the imputation of household composition is based on

[12] The probabilities are estimated by a logit regression. The explanatory variables are: Gender, age, square of age, born in Sweden (dummy), Study allowance / 1000, earnings / 1000, highest education (dummies for compulsory school and upper secondary school), interaction age and education.

those households in HEK where at least one member is born outside Sweden.

After the adjustment the emigrants are merged with the base population. Note, however that the emigrants are treated somewhat different in the simulation. For instance for these individuals no change in household composition is modelled.

4. SESIM – a stochastic simulation model

SESIM is a *stochastic simulation model,* which means that the statistical models also include a random component. In the simulation a *Monte Carlo* technique is used to generate a stochastic process. Consider the typical case in SESIM, and in dynamic MSM, where the dependent variable is binary. This variable then have a Bernoulli distribution, i.e., $Y_i \sim bernoulli(\pi_i)$, where $\Pr[Y_i = 1] = \pi_i$ and $\Pr[Y_i = 0] = 1 - \pi_i$.

As an illustration, let Y_i denote unemployment for individual *i* during the period of interest. Let $Y_i = 1$ denote unemployment and $Y_i = 0$ work, π_i denote the probability that the individual is unemployed during the year. This event is simulated by comparing π_i with a uniform random number. If $u_i < \pi_i$ the event is realized and individual *i* become unemployed.

The propensity of becoming unemployed is determined by π_i, by allowing π_i to be determined by individual or household attributes which can explain unemployment. This is typically accomplished by a logit or probit regression. The logit model is given as $\pi_i = [1 + \exp(-X_i\beta)]^{-1}$, where X_i is a vector of individual or household characteristics (or any other characteristic relevant for explaining unemployment, i.e. rate of regional unemployment) and β is a vector of parameters.

Due to the Monte Carlo simulation, the number of generated events when repeating the simulation does not have to be always the same.[13] Let *T* denote the total number of individuals in a population size *N* that experience the simulated event, that is $T = \sum_{i=1}^{N} Y_i$. If the individuals are simulated independent of each other the expected number of events is $E(T) = \sum_{i=1}^{N} \pi_i$ and the variance $Var(T) = \sum_{i=1}^{N} \pi_i(1 - \pi_i)$. If *N* is large enough, and π_i not too close to zero or one, *T* is approximately normally distributed. Assume an event with a 10 percent probability (and for simplicity that all individuals face the same risk). For a population size of 10 000 individuals and a large number of repeated simulations, the number of individuals that experience the event in 95% of the cases[14] is between 941 and 1 059.

The Monte Carlo variation can be problematic in evaluating the results from an experiment. If, for instance a change in a tax rate is evaluated then, due to the Monte Carlo variation, it is difficult to isolate the pure tax effect from the stochastic Monte Carlo effect. One approach could be to repeat the simulations a number of times and use the average result, since this reduces the effect of the stochastic simulation.

An alternative approach is to use methods that are related to the reduction of the Monte Carlo variance. In SESIM a method is used which is directly related to calibration. Calibration is a technique used in order to predict stocks which meet an a priori defined exogenous target. In the binary model this implies that the expected number of predicted events have to be adjusted in

---

[13] Given that the seed used for generating random numbers is changed in each simulation.

[14] If number of events is approximately normally distributed a 95 % interval is defined as: $10\,000 * 0{,}1 \pm 1{,}96\sqrt{0{,}1 * 0{,}9 * 10000}$

order to coincide with a given target. In SESIM this is accomplished by adjusting the predicted probabilities rather than the predicted stocks. A simple, and quite common, technique is a proportional adjustment, $\pi^*_i = \alpha\pi_i$, where π^*_i is the adjusted probability and α is the factor of adjustment. A problem with this technique is that it does not restrict π^*_i to be in the [0,1]-interval. If instead $\pi^*_i = \min(1, \alpha\pi_i)$ is used, this implies that individuals where $\pi^*_i = 1$ with certainty will experience the event. This can produce unrealistic results, for instance all individuals with the same set of attributes will die. Alternatively, the adjustments can be made using a different scale. In SESIM an additive adjustment on the logit scale is used. This is equivalent to adjusting the intercept term in the estimated logit model. Thus, SESIM uses $\text{logit}(\pi^*_i) = \alpha + X_i\beta$, where $\text{logit}(x) = \log[x/(1-x)]$, or $\pi^*_i = [1+\exp(-\alpha - X_i\beta)]^{-1}$ which implies $\pi^*_i \in [0,1]$.

Regardless of the approach, the adjustment factor α must be calculated. In the first approach this is simple (given that the frequency of truncated probabilities are low) but in the second it is a bit more complicated. Even if the techniques discussed above ensure that the expected number of events corresponds to the desired, the Monte Carlo variance can still produce discrepancies. The method for variance reduction that is used in SESIM, eliminates these discrepancies by using the α that, given the random values of u_i, generates the exact number of events n. The problem of calculating α is solved by the fact that this is equivalent to sorting the variable $v_i = \text{logit}(u_i) - \text{logit}(\pi_i)$ in an ascending order and letting individuals with the lowest rank obtain a positive event.

Even if calibration is a common method in dynamic MSM there have still been some critique, see Klevmarken [1998]. If the discrepancy from the expected result is due to an incorrect specified model, then this should be corrected for, and not solved by calibration. However, despite this it is difficult to accomplish credibility if for instance the model does not track a given a priori demographic forecast. Alternatively calibration can be viewed as a method of implementing different scenarios in a simulation, for instance, the effect of two different assumptions on fertility. To adjust the estimated parameters in accordance to these forecasts is difficult due to the complexity of the model.

The different models/processes in SESIM that use calibration or variance reduction are mentioned in section 4.3 below.

The Monte Carlo variance is not the only source of random variability in a dynamic MSM. Since the sample used for simulation is drawn from some population, this introduces another source of randomness. Furthermore since the estimated parameters are random this also introduces a source of randomness[15]. Thus, in order to conduct inference that incorporates all the possible random sources all this randomness should be accounted for. However, due to the complexity of a dynamic MSM it is quite difficult to derive analytical results for the true variance in the variables of interest. However, methods based on simulation could be used, for instance *bootstrap*, se Davison and Hinkley [1997]).

Below we give a more in-depth description of some of the stochastic models in SESIM.

---

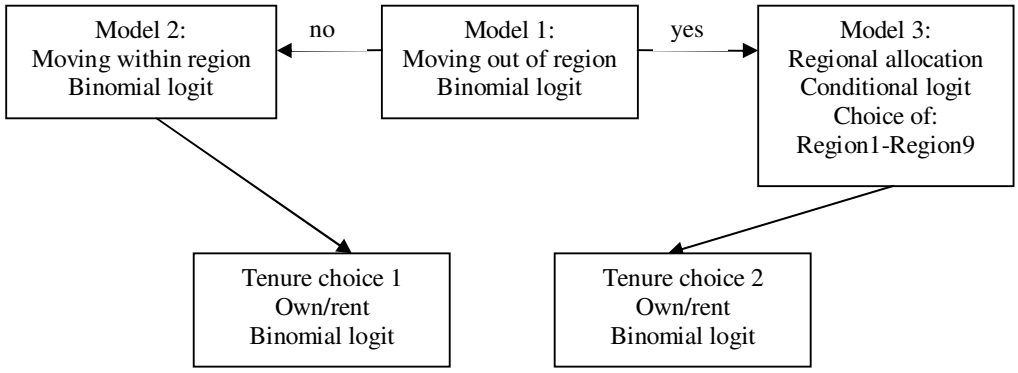
[15] However, due to the large sample size that is used for estimation in SESIM this source of error is presumably rather small.

4.1 Regional mobility and tenure choice

An important extension in SESIM III is the inclusion of a model for regional mobility, see Eklöf [2005]. This module is completely based on the GEOSWEDE-database, covering all individuals that have lived in Sweden during the period 1990-2000. Regional mobility is partitioned into four parts. The first part models the probability that an individual will move outside of his region (so called LA-regions[16]). The second part models the probability of moving inside a region. In the third part, conditional on moving outside the region, the probability to move into any other region is modelled. Finally, conditional on a move, the tenure choice (owning or renting) is modelled.

The first model, i.e. moving outside of one's region, focuses on labour market motives including the individual's history of unemployment spells, education and age. In contrast, the intra-regional model considers changes in family composition and demand for housing. The model for the choice of region focuses on unemployment, local tax rates and price of housing.

Figure 4.1 Model structure for regional mobility and tenure choice



An important change compared to SESIM II is that mobility also implies selling or buying a house/apartment. The tenure transitions considered here are: rent to rent, rent to owned, owned to rent and owned to owned. Depending on the transition, market values of dwellings, capital gains etc are calculated.

4.2 Retirement decision[17]

The retirement model in SESIM focuses on early retirement, that is, retirement before age 65. Hence, in its present implementation, all individuals are forced to retire at 65 at the latest. Note that the model is estimated on 1990's data. After this period, the national old age pension system as well as the occupational pension systems has been reformed, and thus the estimations are obsolete and need to be updated. However, this is of minor importance currently, as the model normally is used with this module exogenised.

---

[16] 1=Stockholm, 2=Göteborg, 3=Malmö, 4=Urban southern parts, 5=Urban central parts, 6=Urban northern parts, 7=Rural southern parts, 8=Rural central parts, 9=Rural northern parts.

[17] See Eklöf & Hallberg [2004]

Consider an individual who faces two alternatives; retire early (R) or remain at work (W). The economic incentives are measured by two variables; the net present value (NPV) of the future pension benefits and the accrual value of NPV. Formally, let Y_t denote the income in period *t*, and let *B(t,r)* denote the pension benefits received in period t if retiring in period *r*. Then NPV if working until 65 is

$$NPV_t^W = \sum_{s=t}^{64} \delta^{s-t} Y_s + \sum_{s=65}^{T} \delta^{s-65} B(s,65)$$

where T is maximum life length and δ a discount factor.

And the NPV of retirement in period t is

$$NPV_t^R = \sum_{s=t}^{T} \delta^{s-t} B(s,t)$$

This measure is usually referred to as social security wealth (SSW).

All benefits from the public, private, and occupational pension schemes are included in the NPV measures. In addition to NPV^W and NPV^R, we also consider the level of income next year and the accrual as determinants of early retirement. The income the following year is denoted

$$INC_t^W = Y_t$$

$$INC_t^R = B(t,t)$$

for earnings and pension benefits, respectively. Finally, the accrual of NPV^R if postponing retirement one year is defined as

$$ACC_t^R = \delta NPV_{t+1}^R - NPV_t^r$$

The accrual reflects the increase in NPV^R given one additional year at work. If the compensation in future benefits for postponing claims is less (more) than actuarially fair, the accrual value is negative (positive). We expect a higher accrual to reduce the probability of retiring. For completeness, let

$$ACC_t^W = \delta NPV_{t+1}^W - NPV_t^W = \delta Y_{t+1}$$

Both NPV and the accrual are measured in terms of qualifying wage (annual taxable incomes).

The utility associated with alternatives j=R,W is

$$U_j = \alpha_j + \beta_1 NPV^j + \beta_2 INC^j + \beta_3 ACC^j + \varepsilon^j$$

Where ε is a zero-mean iid error component.

The individual is assumed to choose the alternative with the highest utility. Thus, the probability that the individual retire before 65 is,

$$P(R) = P(U_R > U_W)$$

$$= P(\varepsilon_W - \varepsilon_R < \alpha_R - \alpha_W + \beta_1 (NPV^R - NPV^W) + \beta_2 (INC^R - INC^W) + \beta_3 (ACC^R - \delta Y_{t+1})$$

Apart from the economic incentives other covariates in the model include dummies for sex, educational level, household composition, and labour force status of the spouse. We have deliberately omitted age dummies as we want the predictions to be based on economic incentives rather than historic age structures of retirement.

The net present values of the future pension benefits are calculated using the same index as it is used for indexing the different benefits. In most cases this is the basic amount index.

Furthermore, the NPV calculations include a time preference factor equal to 3 percent and survival probabilities collected from exogenous survival tables in SESIM. When calculating the NPV of future pension benefits, the procedure iterates over time periods for each individual separately and calls the SESIM functions relevant for calculating the pension benefits. That is, for each individual that is eligible for early retirement, the procedure iterates over age and calls pension benefits procedures. This produces a sequence of future benefits that are discounted to a present value and summed up. Hence, future benefits calculated by the NPV algorithm do not necessarily equal the actual benefits once SESIM is stepped to next year as benefits depend on random elements that are not realized at the time of the NPV calculations, e.g. random health shocks or shocks to earnings. However, the NPV calculations should be close to realized values as SESIM iterates over model years.

As discussed in Eklöf & Hallberg [2004], early retired individuals frequently claim occupational pensions only, while the national old age pension is not claimed until their 65'th birthday. In order to capture this feature, we need to make some modifications to the original SESIM pension rules. First, for white collars and central government employees, the replacement ratio before age 65 is much higher than the replacement ratios after 65. Secondly, in the reformed pension scheme for central government employees, the benefit level prior to the 65 birthday is based on the "capital value" of the future benefits (this applies to both the defined benefit and the defined contribution part). In the old pension scheme, the high replacement ratios before 65 were not financed, whereas in the reformed scheme, the temporary pension before 65 directly affects the benefit levels after 65.

4.3 Simulation of Earnings

Due to the importance of earnings a detailed description of this process is provided. It is well known that using information from a cross section only in general produces incorrect predictions of individual earning profiles. As a consequence it also produces incorrect predictions for a given cohort. For this reasons the estimated earnings model in SESIM is a random parameter model estimated on panel data, i.e. the same individual is observed repeatedly in the data. The model is given as:

$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \gamma_i + \varepsilon_{ij}$, where
$\gamma_i \sim N(0, \tau^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$.

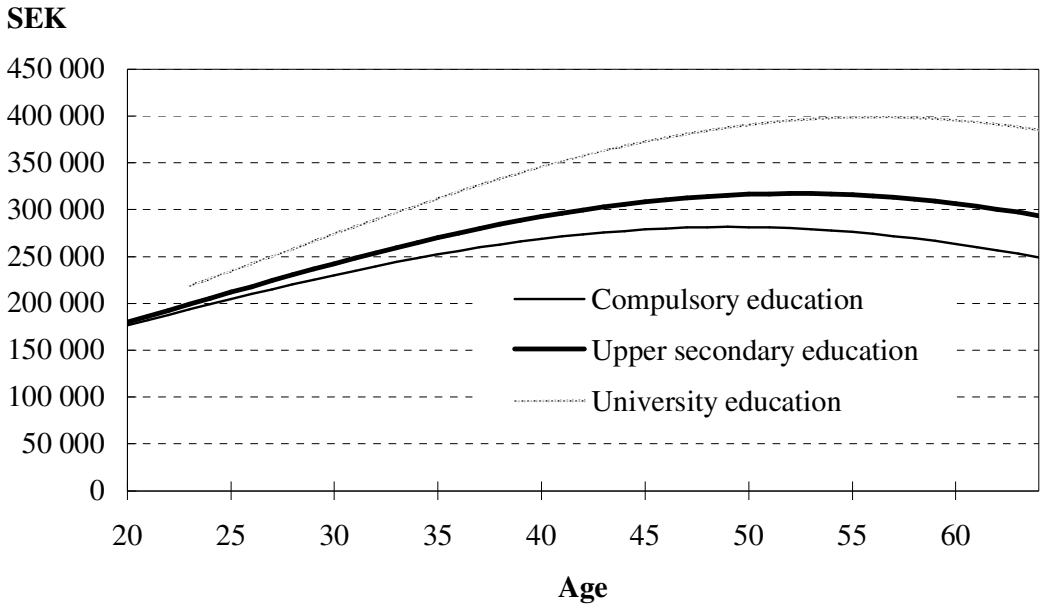
The error components γ_i and ε_{ij} are assumed to be independent. The random intercept γ_i is designed to represent unobserved heterogeneity (typically interpreted as ability). It allows for the fact that given identical X-variables the predicted wage does not need to be the same. The implication is that earnings for a given individual are not independent over time, but independent across individuals. The covariance matrix for $\mathbf{Y} = [Y_{11}, Y_{12}, \mathrm{K}, Y_{1J}, Y_{21}, \mathrm{K}, Y_{IJ}]^T$ is block-diagonal. Each block, that represents data from one individual, has as its diagonal element $\tau^2 + \sigma^2$ and off-diagonal elements τ^2. This implies that the correlation, for each individual, over time is $\rho = \tau^2 / (\tau^2 + \sigma^2)$. Thus, the correlation is high if the variance for the random intercept is high in relation to the residual variance. For a presentation of statistical models for panel data see for example Diggle, Liang and Zeger [1994] or Baltagi [2001].

The earnings equation in SESIM includes in the X-vector variables such as; experience, highest level of education, occupational sector, marital status and nationality. Separate models are

estimated for males and females and separate estimations of τ^2 and σ^2 are done for each occupational sector. The dependent variable is the logarithm of earnings.

Figure 4.2 shows expected earnings for Swedish born, married/cohab males working in private white-collar sector for different levels of highest education. Since the model includes working experience, but not age, it is assumed that working life starts at 16, 19 and 23 years of age for the different levels of education and continue until 64.

Figure 4.2. Estimated earnings profiles (1999 years prices).



Source: SESIM

As expected there is a clear effect of education, as the profiles indicates that a higher level of education implies higher earnings and a steeper curve.

Table 4.1. Effects of independent variables in the SESIM earnings equation.

Variable	Comment	Male	Female
Education	Compulsory	0,71	0,66
	Upper secondary	0,80	0,75
	University	1,00	1,00
Sector	Private blue collar	1,60	1,43
	Private white collar	2,02	1,80
	Governmental	1,66	1,49
	Local governmental	1,51	1,26
	Own employed	1,00	1,00
Nationality	Sweden	1,00	0,97
	Abroad	1,00	1,00
Marital status	Single	1,01	1,06
	Non-single	1,00	1,00

Source: SESIM

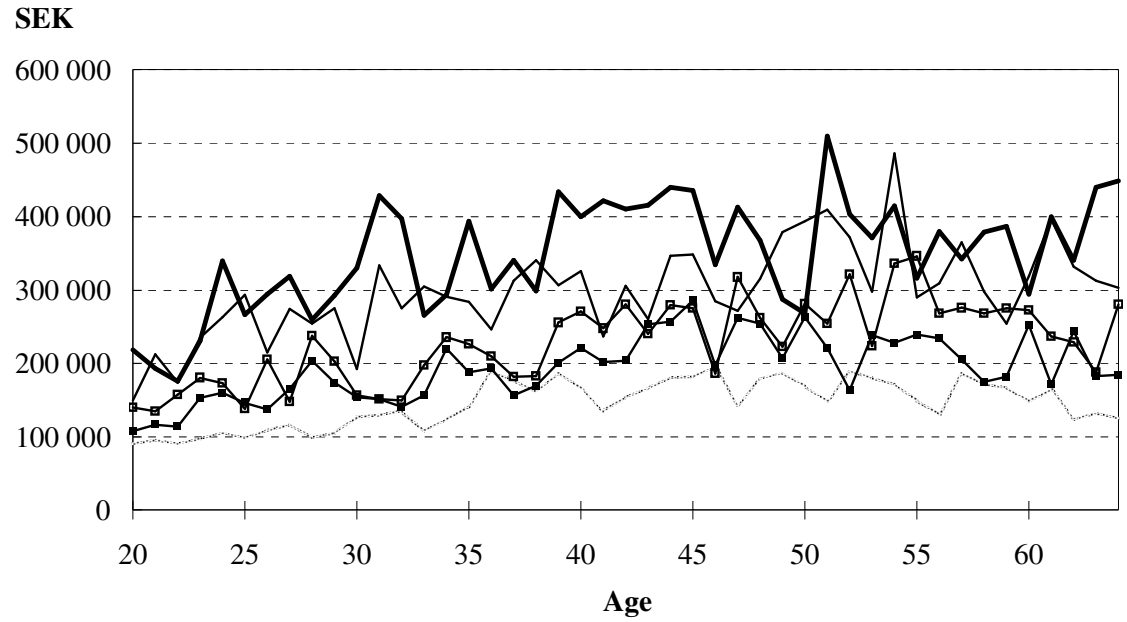
Table 4.1 shows the effects of all of the X-variables (except experience) on expected earning[18]. The educational premium is somewhat higher for women, the premium for highest versus lowest is 34 % for females and 29 % for males. The highest occupational sector is private white collar. Nationality and marital status have only minor effects.

The simulations of the earnings equation is based on the individual attributes in X_{ij}, the estimated parameters $\hat{\boldsymbol{\beta}}$ and the random numbers η_i and ε_{ij}. The random numbers are drawn from two independent normal distributions with variance $\hat{\tau}^2$ and $\hat{\sigma}^2$ respectively. The simulated earnings is calculated as $Y_{ij} = \mathbf{X}_{ij}\hat{\boldsymbol{\beta}} + \eta_i + \varepsilon_{ij}$,. Since η_i is specific for each individual and constant over time, only one draw at the start of the simulation is need, but draws for ε_{ij} have to be repeated for each year (and new individual).

Figure 4.3 shows five independent earnings simulations for Swedish born, private sector white-collar men with gymnasium as the highest education. Note that the profile in the middle corresponds to the expected values. Thus, there is an individual variation in earnings over time. It is also possible to see that the levels are different; this follows from the random intercept.

Note, that the model described only generates income from market work, i.e. individuals with status= 8. For other individuals the incomes are generated conditional on their status. However, the predicted earnings in SESIM are not only determined by the earnings equation, since this equation is conditioned on the status. All models for estimating a status play a role in determining final earnings.

Figure 4.3. Five simulated earnings profiles (1999 prices).



Source: SESIM

[18] These values are defined as the expected value of earnings evaluated at the value of the relevant X-variable divided by the expected value evaluated at the reference value, *ceteris paribus*.

4.4 Modelling financial and real wealth

Information on wealth from the LINDA as well as complementary information on housing characteristics from HEK is used to estimate the portfolio allocation (e.g. between real and financial wealth) as well as the cost of housing of individuals or households.

Data on income taxes and benefits comes from administrative records. How reliable are the data? One problem is that some assets like car, boats and other durables as well as some assets abroad is underreported. Another problem is related to household wealth, since the definition of a household in administrative data like LINDA is problematic. Finally there is the lack of long time series - here we only have access to wealth data from 1999 to 2002 - the implication being that we are not able to identify time or cohort effects.[19] Furthermore, both 1999 and 2000 represent a period of unprecedented high returns on the Stockholm Stock Exchange.

A special effort has been spent on the construction of accumulated tax-deferred pension savings. To the best of our knowledge this is the first time in Swedish data that the value of the stock of pension savings has been imputed at the individual level. This has been achieved by summing up yearly tax-deferred pension savings. In order to minimize the starting value problem we have used data from1980 and followed individuals up to year 2000. Details are described later on in this chapter.

The wealth and pension savings module includes a large number of variables. Four separate assets are considered in the household portfolio; financial wealth, owned home, other real wealth (e.g. second home) and private pension savings. The calculations are carried out sequentially; the order is given in Figure 4.4. It is instructive to start with financial wealth, other real wealth and pension savings.

The flow chart in Figure 4.4 starts the process in year 2000 (the first simulated year) by the diamond-shaped box in the upper left-hand side. This is a check whether the household had financial wealth in the start data (year 1999). If no the probability of having financial wealth in that year is imputed using model (1). If a positive wealth is predicted the next model (2) is applied in order to calculate how much. If instead the household had already financial wealth in 1999, the value year 2000 is updated using a dynamic random effect model, model (3).

Note that new financial wealth is modelled as a two-part model. That is, the probability of financial wealth is estimated independent of the value. The reason for using the two-part model compared to, for instance a generalized tobit or two-stage methods (heckit), is that we are not interested in explaining selectivity. Here, the purpose is to obtain good predictions: it is demonstrated in Manning et. al. [1987] that the two-part model performs at least as good as the tobit type 2. In Flood & Gråsjö [2001] the sensitivity of the generalized tobit model to specification is demonstrated (i.e. errors in the specification of the selection equation produce bias in all the estimated parameters). Here we are much more concerned in robustness compared to a potential increase in efficiency.

The parameters of the estimated logit model and the robust regression model (M-estimation) are presented in Flood [2005]. The estimated age profile for probability of financial wealth has an inverted U-shape, with a maximum at about 60 years. However, the estimated age profile conditional on a positive wealth is almost flat over the whole age interval. There is a strong

[19] Andersson, Berg & Klevmarken [2001] reports important cohort effects.

effect of income; higher income implies a higher probability of wealth as well as a higher level.

The dynamic random effect model (3), used for uprating existing financial wealth, has as RHS-variables the lag of financial wealth, interaction with lag financial wealth and age and income and finally a variable for the value of the general index on the Stockholm stock exchange in that year. There is a strong effect of wealth in the previous year (the coefficient is 0.92); income and age as well as the stock index also have an effect.

Subsequently the model estimates household real wealth. Household real wealth is decomposed into two sub-components: own home and other wealth. Since the probability of owning a home is modelled in the regional mobility module, only other real wealth is discussed here. The steps for other real wealth are quite similar to those for financial wealth. First, a check is made of whether the value of real wealth was positive in the previous year. If not, a logit model for the probability of getting real wealth during the year and a robust regression for the level are run. The covariates are age, income and marital status. If there was a positive value year *t-1*, the value for year *t* is instead updated by a simple random walk model (5).

Since the probability of buying/selling a house/apartment is directly related to the mobility decision this is modelled in the mobility module. Conditional on the decision to move, the tenure choice (own/rent) is determined by a logit model. Given that the model predict owned, the market value as well as the area of the new home is predicted. Housing area is needed in order to impute cost of housing as well as housing allowances.

Next, individual pension savings are imputed in model (6) and (7). These refer to yearly tax-deferred pension savings Model (6) and (7) apply to first time savers, then a simplifying assumption is made, namely that this amount grows with consumer price index each year until retirement.

The estimation of models (6) and (7) of course depends on the known stock of pension savings. This offers a challenge since the stock of accumulated tax-deferred pension savings, at the individual level, is not known in Linda (or any other data). This is because these savings are not taxed until after retirement, while the returns on these savings are added to pension income and taxed as ordinary income. The only information available from data is the yearly tax deductible savings. For an analysis and descriptive statistics of these yearly savings, see Johannesson [2001] and Konsumentverket [1999].

The simple idea used here is to construct accumulated savings by using repeated Linda panels. Individual savings are summed up over years and the resulting stock is increased each year by applying the average return from life insurance companies. In order to reduce the starting value problem, we start as early as 1980, a time when private tax-deferred pension savings was rather unusual.

Table 4.2, below summarizes the main characteristics of pension savings during the period 1980-2000. Column (2) gives the share of all individuals with pension savings; note this is the share of the whole population, regardless of age. During this period there has been an increase from about 4 to 21%. The share with positive accumulated savings, i.e. private pension wealth, is given in column (6). In year 2000, more than 30% have positive accumulated savings, their mean value is 110 863 SEK, see column (7), and the corresponding mean of yearly savings is 6 591 SEK, see column (3). Even if the share of pension savers has increased over time the yearly amounts have not. The yearly savings reached the highest values in 1989 and since then they have decreased.

The reason for this is that changes in the rules after 1989 have done savings less generous; also in recent years the return on these savings has been quite low.

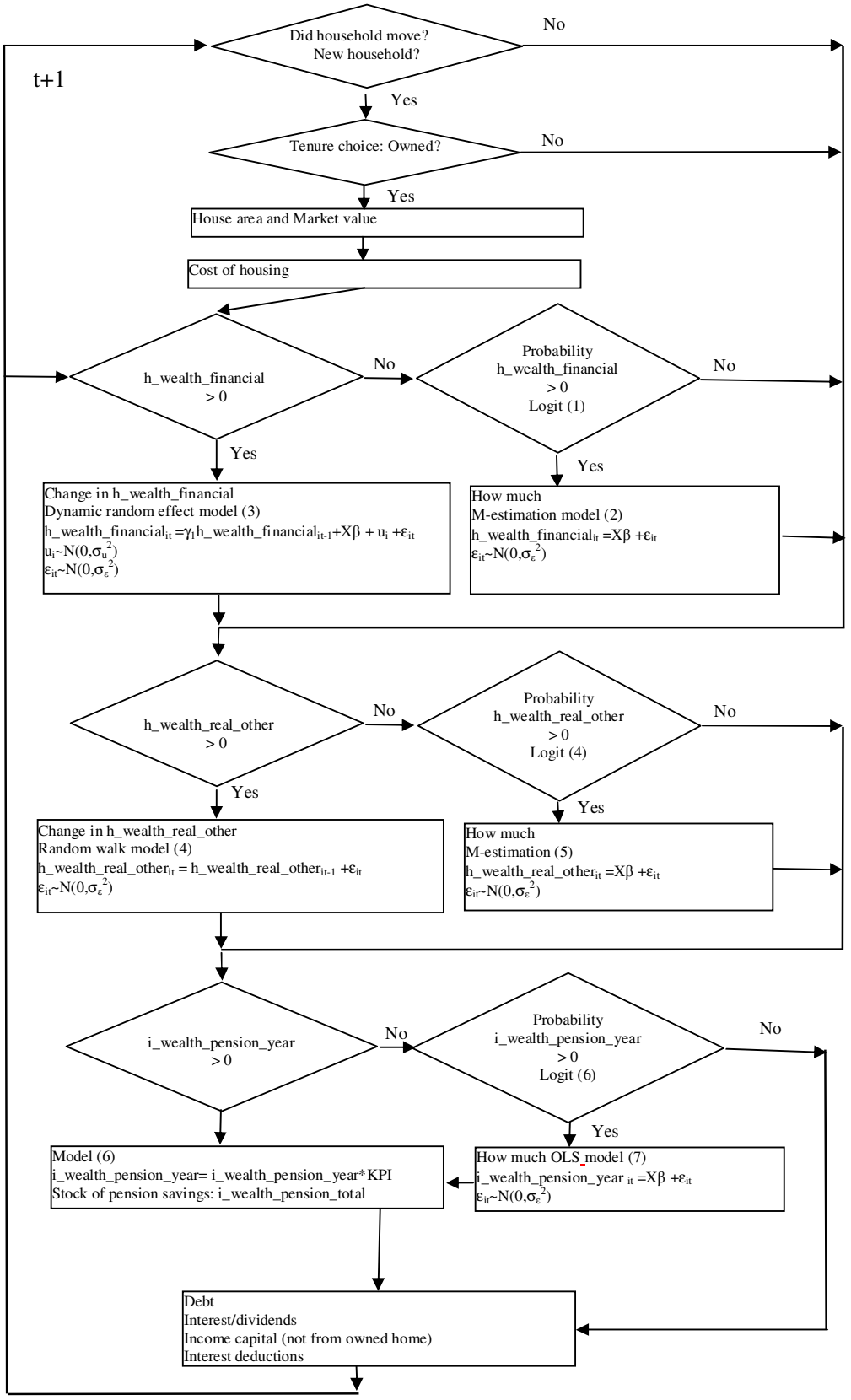
Table 4.2 also includes information about the share of individuals with income from private pension savings, column (4) and the mean values, given an income, column (5). The incomes from pension savings are relatively small, the reason for this is that this saving is a new phenomenon and the generated stock is still relatively small. However, the average amount for those 4.8% who had an income in year 2000 was 32 598 SEK.

Table 4.2. Pension savings 1980-2000

	Share with pension savings	Mean value given savings	Share with income from pension savings	Mean value given income	Share with pension wealth	Mean value given pension Wealth	Sum of pension wealth	Assumed return on savings
	(%)	(tkr)	(%)	(tkr)	(%)	(tkr)	(mkr)	(%)
1980	4.60	3 529	0.00	936	4.10	3 882	1 396	10
1981	4.70	3 962	0.00	1 416	4.40	8 086	3 137	10
1982	4.90	4 748	0.00	981	4.70	13 136	5 449	10
1983	3.80	6 968	0.00	2 127	4.80	19 728	8 411	12
1984	4.40	7 846	0.00	1 469	5.00	28 427	12 550	13
1985	8.20	8 321	0.00	1 427	5.40	39 080	18 828	15
1986	8.50	9 229	0.70	11 621	6.90	43 074	26 553	14
1987	9.70	9 969	0.80	14 074	8.70	46 748	36 056	12
1988	11.90	11 170	0.70	7 676	11.00	51 523	50 285	14
1989	14.40	12 955	0.70	8 027	13.60	62 291	74 903	21
1990	14.50	8 138	0.70	8 319	15.50	69 798	95 710	16
1991	12.50	9 656	2.60	21 013	17.20	73 414	111 944	10
1992	12.80	8 339	2.90	22 175	18.50	76 117	125 012	7
1993	13.30	8 465	3.30	23 476	19.50	79 001	136 367	5
1994	15.00	8 762	3.50	23 572	21.40	80 551	152 702	7
1995	16.20	6 861	4.10	22 528	23.00	82 478	168 393	7
1996	17.30	6 764	4.10	23 608	24.60	85 822	187 482	8
1997	18.20	6 705	4.20	25 272	26.10	92 546	214 326	11
1998	19.20	6 659	4.30	27 870	27.80	100 973	248 870	13
1999	20.50	6 785	4.50	30 540	29.70	104 530	275 265	8
2000	21.90	6 591	4.80	32 598	32.00	110 863	315 101	12

Source: own calculations based on the Linda panel 1980-2000. Information on average returns, in column (9), comes from The Swedish Insurance Federation (www.forsakringsforbundet.com). Note, these returns are returns before tax and administrative costs.

Figure 4.4. Financial and real wealth and cost of housing in SESIM



The accumulated pension savings are given in column (8). The low value in 1980 indicates that the starting value problem is quite small; pension savings were unusual before 1980. The total pension wealth has increased to 315 billion SEK in 2000.

Given this accumulated stock of pension savings, we have information for each individual on their savings starting from 1999. For individuals who made any deductions for pension savings in 1999, we assume that they will continue saving this amount (adjusted for CPI) every year until the age of retirement. For individuals who did not have any savings in 1999, a two-part model for new pension saving in 2000 has been estimated. Populations at risk are all individuals 18-64 year in 2000 who did not have pension savings in 1999. In forecasting accumulated pension savings, we have to estimate the probability and the amount saved the first time. Then we assume that the individual save the same amount (adjusted by CPI) each year until age 64.

Yearly savings thus estimated are then added to the stock and an assumption on a yearly return is used.

Regarding the private and occupational pension payments stream, many different options are possible. In the current version of SESIM, a five year period is the norm, but we also allows for some variation in order to match the observed profiles in year 1999.

The next step is to impute debt. Several models and subgroups are used to this end. First, we look at households without debt in year t-1. For these households a model for probability of contracting debt in year t is applied, and then conditional on having a debt a robust regression model is used for predicting the level. Next, household with a debt in year t-1 are divided into three cases, depending on changes to the market value of their own home:(1) a decrease larger than 100 000 SEK(2) no large change and (3) an increase larger than 100 000 SEK. For each case a two level model (probability and level of debt) have been estimated.

The next step is to impute, for each household, income from interest and dividends. Interests and dividends are simulated as a rate that, multiplied by the household's financial wealth, returns the amount of interests/dividends. Due to difficulties in finding a suitable statistical model the rates are simulated from the empirical distribution function of all household rates.

Capital gains are divided into two categories: gains on own home and other capital gains. Capital gains on own home are calculated by comparing purchased value and market value of the sold house/apartment. Note that the purchase value is known for all home/apartments bought during a simulation. However, the initial purchase value is not known and for this reason these values are imputed using statistics about sold properties during 1999-2002. The obtained relation between purchased value and market value for different age groups is used to impute initial purchased value at the start year 1999. Other capital gains are imputed using a probit model.

Finally, in order to impute a level of deductions on interest rates, the population is divided into two parts; households with and without deductions in year t-1. For both these groups two-level regression models are used for the imputations.

4.5 Non-cash benefits

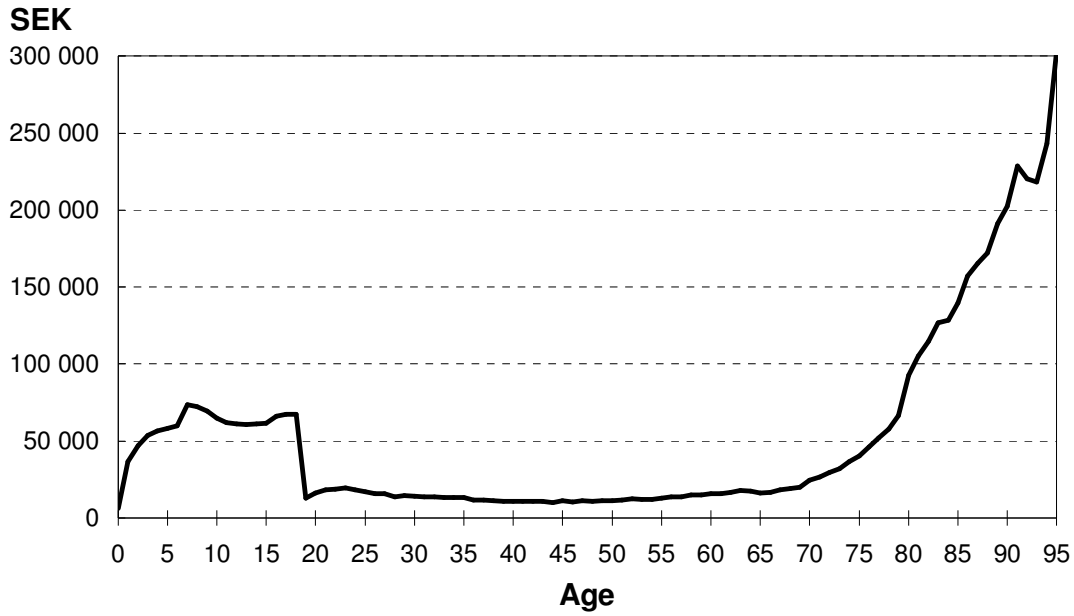
SESIM allows for a rich definition of income since, apart from all other income components, also the value of various non-cash benefits is included. The impact of non-cash benefits on the cross-sectional income distribution is studied in Ministry of Finance [1999, 2002]. These studies used the income distribution surveys of 1997 and 1999, respectively. For all individuals in the survey, actual use of various non-cash benefits was multiplied by the subsidy value of the respective benefits.

The value of the subsidy is assumed to equal production costs net of fees. Only benefits that can be attributed to a specific individual are included. The following is included:

- Compulsory education
- Upper secondary education
- University
- Adult education
- Child care
- Old age care
- Labour market activities
- Health care
- Medications

LINDA does not contain any information concerning the use of non-cash benefits; instead this information is imputed from the income distribution survey of 1999. Logistic regression models are used to predict participation to a given service and linear regression models are used predict subsidy values for those participating. For some of the subsidies, it is not realistic to assume that the value, or utility, is equal to the net production cost. Following Smeeding et. al. [1993] imputation of health and old-age care subsidies is based on a risk-related insurance premium approach. That is, old-age and health care are regarded as an insurance benefit received by all covered independently of their actual use. Benefit levels (insurance premium) differ between age and gender according to differences in need (based on actual usage within each group). In this way the insurance premium is considered to be actuarially adjusted to account for differences in need-related values of being covered by the insurance. There is a clear age pattern in the imputed subsidies shown in Figure 4.5 below.

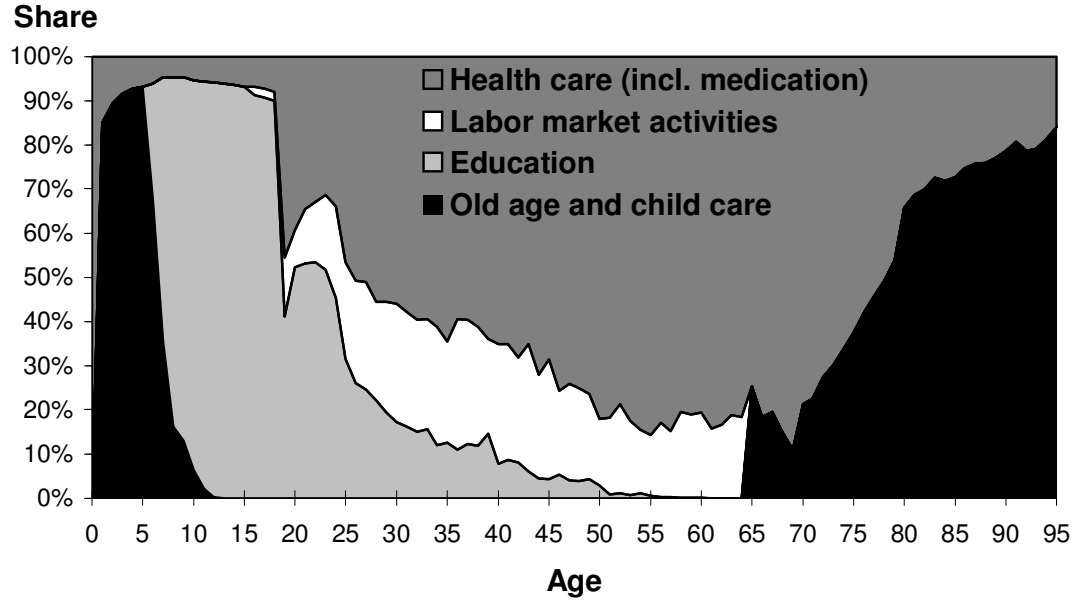
Figure 4.5. Average non-cash benefits per age



Source: SESIM

The average individual, with a life span of 80 years, receives 74 percent of total subsidies before the age of 20 and after the age of 64. The corresponding share for an individual with a 95-year life span is 88 percent. Figure 3.5 shows the average composition of non-cash benefits at different ages. At preschool ages, and from 80 years and on, childcare and old-age care dominates, somewhat more pronounced for children than for the elderly. For the oldest, health care subsidies also comprise a large share of the total. Health care dominates between ages of 25 and 78 as well, particularly between 50 and 70 years of age. Education naturally dominates during school age years and continues up to age 24.

Figure 4.6. Composition of non-cash benefits



Source: SESIM

4.6 Health and care of elderly.

The main building blocks in the health and care modules are models for:

- Closeness to relatives (the distance to a relative)
- Health index
- In patient care
- ADL (dependence in activities for daily living)
- Assistance elderly

Closeness to relatives is coded as a binary variable, where one denotes a relative that is living in the same LA-region. The importance of this measure is as an indicator of access to informal care. Logit models are used to simulate probability of a close relative and separate models are estimated for single male households, single female households and cohabiting/married households. The population at risk is households where the oldest individual is 65 years of age. Imputation models are used for households that enter the population at risk during the current year and dynamic models are used for all other households. Explanatory variables are: Age, income, education, marital status, and region. In the dynamic model the lagged value of the closeness measure is also used as a RHS-variable.

Health index is modelled as an ordered probit with, 0=unknown, 1=severe illness, 2=some illness, 3=not full health and 4=full health. The individuals are divided into two groups depending on if the individual is below or above 50 years of age. For each group two models are estimated, an imputation model and year-to-year model. The imputation model is used for the period 1999-2006 and the year-to-year model is used from 2007. Explanatory variables are: Age, income, education, marital status, number of children, region, gender and nationality. In the dynamic model the lagged value of the health index is also used as a RHS-variable.

Number of days in patient care is modelled as a zero inflated negative binomial [?]. The individuals are divided into two groups depending on age below or above 50. For each group two models are estimated, an imputation model and a dynamic model. Explanatory variables are: health index, age, income, education, marital status, number of children, region, gender and nationality. In the dynamic model the lagged value of days inpatient care is also used as a RHS-variable.

Dependence in activities for daily living (ADL) is measured by an index where, 0=unknown, 1=non-disabled, 2=slightly disabled, 3=moderately disabled and 4=severely disabled. The data comes from the so called Kungsholmen study (a region in Stockholm city). An ordered logit is estimated using the health index, age and gender as explanatory variables. The population at risk is all individuals 65 year or older.

Assistance to the elderly is measured by three values, 0=no assistance, 1=assistance at home and 2=special accommodations. Again, the data from the Kungsholmen study is used; a multinomial logit model is estimated and used for simulating the transition probabilities. The population at risk is all individuals 75 year or older. The explanatory

variables are ADL, previous level of assistance, number of years since the individual became 75, closeness to relative. The initial level of assistance is imputed based on observed frequencies per age, sex and ADL-level.

5. Technical Platform

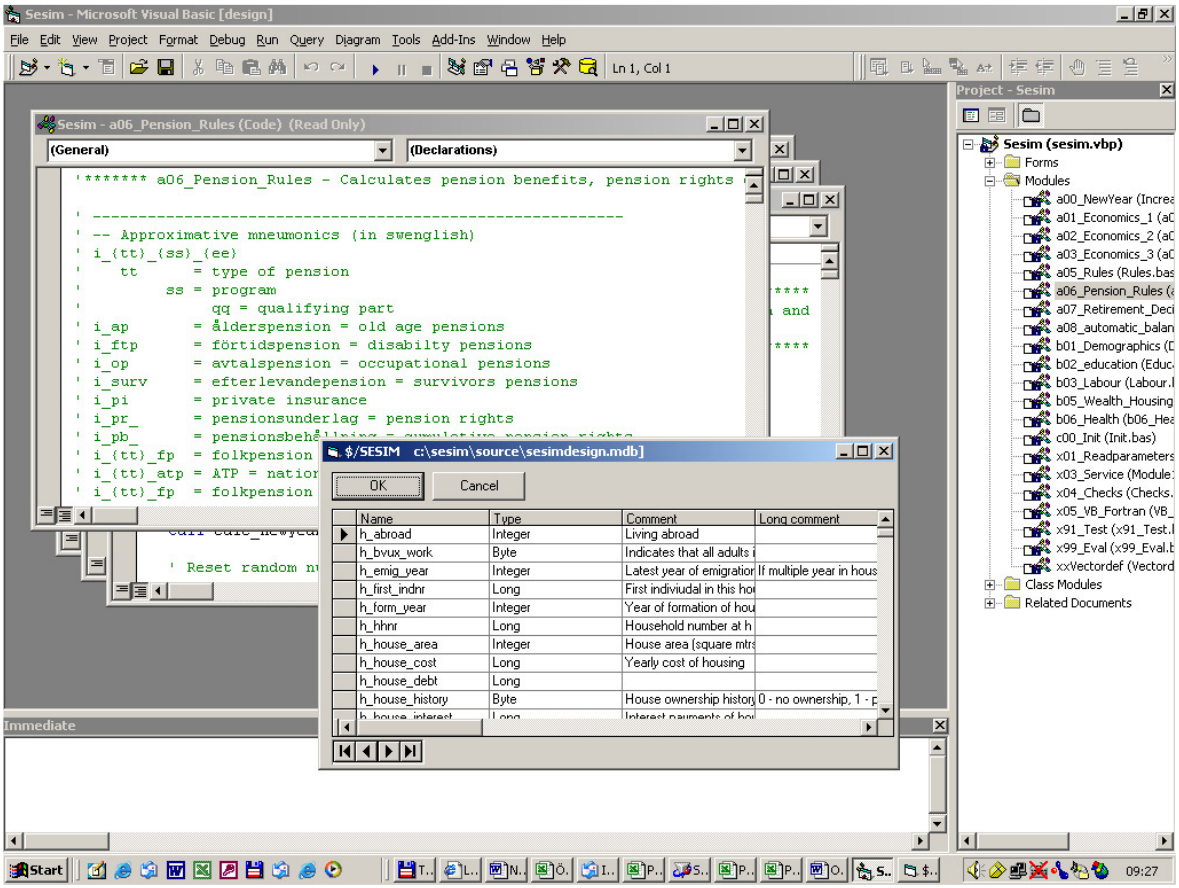
5.1 Introduction

SESIM can be operational in three different environments, or user levels.

1. Source code programming
2. The SESIM model interface
3. Report generator in Excel

In level 1 the user are programming in Visual Basic. The environment in Microsoft's Visual Basic 6 is shown in Figure 6.1. The programmer has direct access to the source code and can compile and run the program from the workbench. This level is used to develop the model and implement changes.

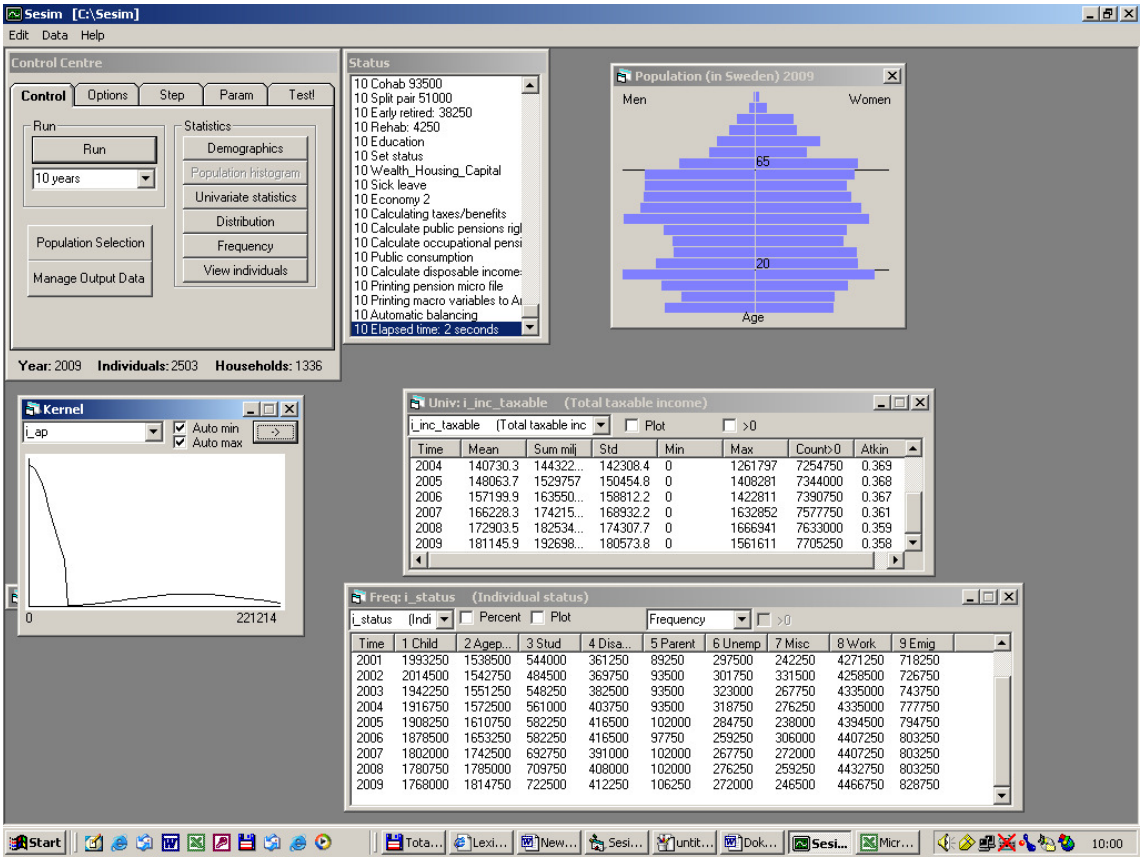
Figure 5.1. Microsoft's Visual Basic 6



Level 2, the model interface, is an environment created with Visual Basic 6, which gives an easy access to simulation results. Descriptive statistics, kernel estimators of densities and frequencies are available for all variables in the model. Also it is possible to investigate subgroups of

individuals. An example is shown in Figure 6.2. In the upper right corner a population diagram is drawn. Each segment corresponds to the number of individuals born in the same cohort spanning over five years. An upper and lower limit indicates 65 and 20 years of age, respectively. Below the population diagram a window with statistics for an arbitrarily selected variable is shown. In the right bottom corner the frequency of the individual status is followed over time. In the bottom left corner public pension density is estimated by a kernel estimator. The remaining two boxes appear by default: A control centre, where simulations are ordered, and a log window. Up to now this level has been used extensively for debugging.

Figure 5.2. SESIM user interface - The debug environment



Many applied economists use Microsoft's Excel as a tool for creating reports and to perform statistical analysis. The third level in SESIM is developed for these users. By utilizing the opportunity of writing Visual Basic code in Excel its possible to order a simulation from a prepared spreadsheet. See section 5.6 "The Excel report generator" for a more extensive presentation of this option.

5.2 Status of this version

As most dynamic MSM, SESIM is a moving target and will be updated frequently. The model core and infrastructure is remarkably stable as it has been tested and evaluated extensively. Even the user interface has rarely been modified. Regarding the estimated models and the programmed rules, changes and extensions are instead much more frequent. To keep track of different versions and revisions Microsoft Source Safe, or other revision control systems, is used.

5.3 System requirements

To run the model a computer with at least 256 MB of RAM memory is recommended. To edit the model Visual Basic version 6.0 SP5 is required. The system requirements increase with the sample size and the simulation period.

5.4 Installation

For instructions about how to install SESIM we refer to www.sesim.org.

5.5 How it works

When you first press the"Run" button, SESIM will read the data from the input files in the *microdata* directory. All old files in the *microdata* directory except the start data ii.bin and hh.bin are deleted. Parameters from two Excel files with exogenous data about demographic and macro economic assumptions are also read.

Next time the"Run" button is pressed and "1 year" is selected, the model will be dynamically updated:[20]

1. *Macro variables* for the actual year are initialized.
2. Some individuals in the model will *die* and their data disappear.
3. A module for *migration*; cloning of some households.
4. New individuals are *born* and added to the data vectors.
5. New *couples* are formed.
6. *Divorce* for some couples.
7. Disability retirement.
8. *Rehabilitation* of disabled.
9. *Education* module.
10. The *status* variable is updated.
11. *Incomes* are calculated.
12. Wealth/housing variables are calculated.
13. Non-cash benefits are calculated.
14. Tax and benefits *rules* are applied, and disposable income can be defined.

The program flow can be followed in the status window. In the file *trace_doc.txt,* which is generated in every run, module details can be studied.

The program can be controlled in two ways:

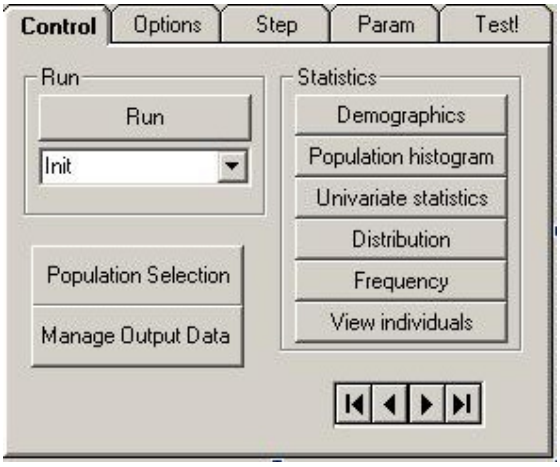
- From the SESIM model interface
- From the "Excel report generator".

Normally we use the model interface in the stage of development, and the "Excel report generator" to produce standard outputs as graphs and tables. More extensive analysis of the results is performed in standard statistical packages like SAS or R.

[20] It's also possible to immediately choose the wanted period, without the init-step.

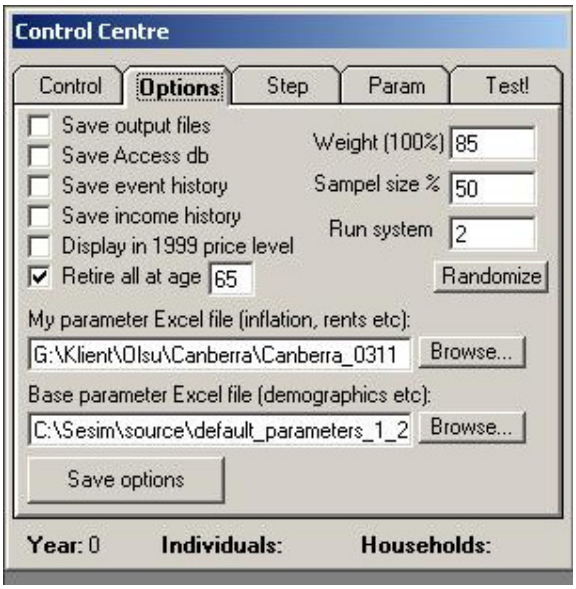
The SESIM model interface, the ”Control Centre” has 5 tabs:

The Control-tab



Run - runs the model the number of years chosen in the box.
Demographics - the current demographic statistics.
Population histogram - shows a population histogram diagram.
Univariate statistics - lets you select a variable for which the sum, mean etc are calculated.
Distribution - simple kernel estimator of the selected variable.
Frequency - calculates counts of categorical variables as sex etc. It is also possible to select a second variable for which means in each group can be computed. A special group may be selected, which force all other statistics windows to show computations only for that group, for example men.
Population selection – Optional global selection for all active viewers, e.g. for a certain status or age group.
Manage out put data – Optional output of microdata for selected variables to a text or binary file.

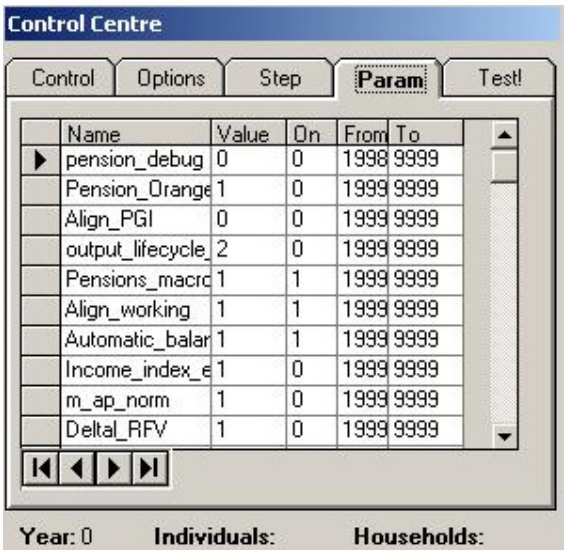
The Options-tab



Save output files - writes all the data to binary files. This feature makes it possible to step back to a specific year later.
Save access db - writes all individual data to a MS-Access file. The file can be viewed from inside SESIM in the tab ”More”, or of course from MS-Access.
Save event history - saves some events (up to the programmer) to an Access file. Mainly used for debugging. Can be viewed in tab ”More”.
Save income history - saves individual income records to an Access file. The file can be viewed in tab ”More”. Not currently used, but may be useful for pension calculations. In the source code there is an example on how to retrieve one individuals complete income history from this file.
Display in 1999 price level - when this box is checked, all *new* statistics windows will show the figures in the 1999 price level.
My parameter Excel file - filename for macro economic assumptions.
Base parameter Excel file - filename demographic assumptions.
Weight - multiplier to get population size.
Sample size – selection of sample size
Randomize button - generates a random seed. By default the same seed is used.
Retire all at age – selection of exogenous retirement age at chosen age. If the check box not marked the endogenous retirement module will be active.
Run system – Obsolete. Shall always be set to 2.

Step-tab
Obsolete

The Param-tab



This grid is for specifying ”runtime parameters” or switches. It makes it possible to make selections of what code to execute and to change some parameters. The implemented switches can be found under the help menu in the model interface. If for example the parameter ”fertility” is set to 1.02 and enabled (On), and used from time 0 to 9999, the fertility rate will increase by 2%. It may be useful for ”interactive calibrating”. It is up to the programmer to actually use these parameters in the code. Look in the fertility module for an example. It is also possible to use runtime parameters to select what code to execute, e.g. to switch off and on the automatic balancing mechanism in the pension system, or to select whether the labour market should be aligned.

Test-tab
Obsolete.

Results
Results in the user interface can be transported or printed in the following way:

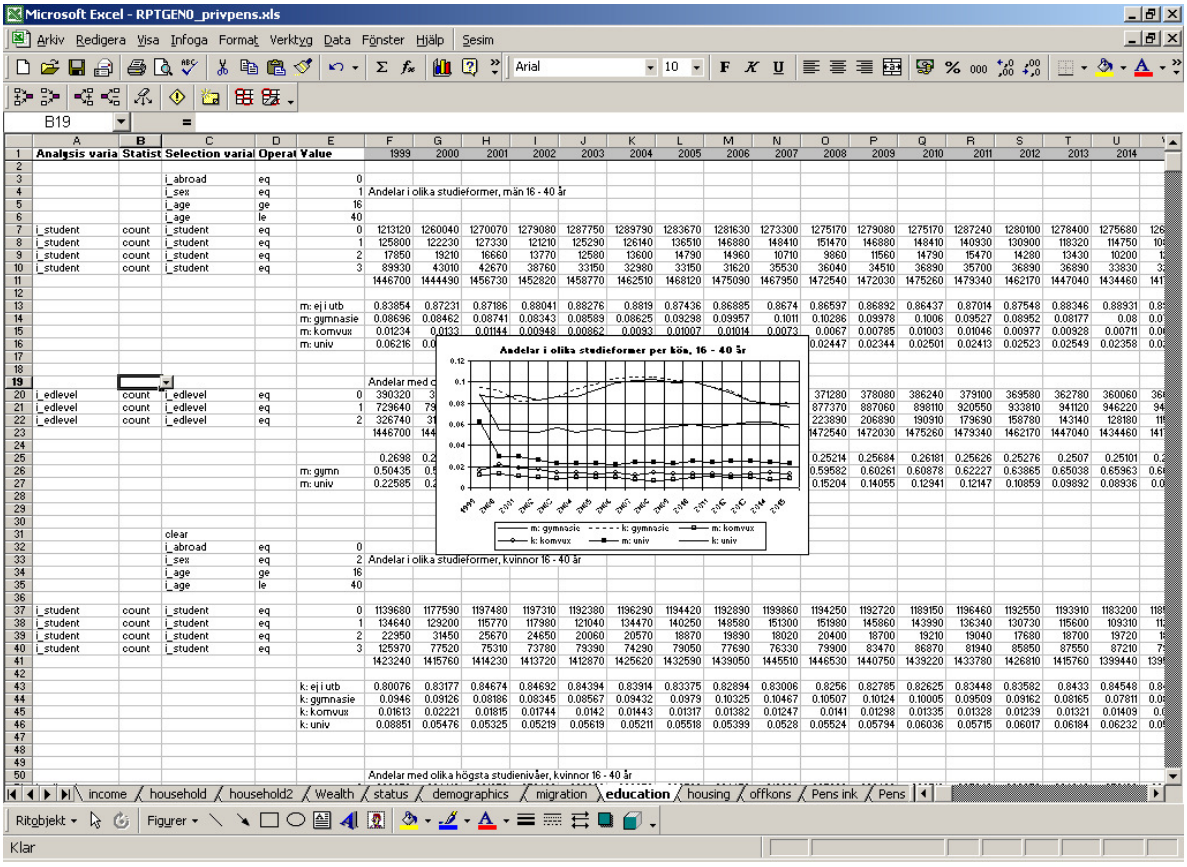
1. Press for example the ”Univariate statistics” button.
2. Select variable ”i_age”.
3. Run the model some years.
4. When the statistics window is active, select ”Copy to editor” on the main menu.
5. Edit, print or copy the results in the editor.

It is also possible to print micro data to a text file. Press “Manage output data” in the control tab and select wanted variables. More detailed information about the different methods for extracting data can be found in the document “Rutiner för datauttag”, currently available in Swedish only.

5.6 The Excel report generator[21]

To use the Excel report generator, first start SESIM, and then open the RPTGEN.XLS file. When the report generator is opened an extra menu will be added to the menu bar in the top “Sesim”. By selecting ”Sesim”-“Execute”, DDE requests are sent to SESIM, which in turn computes and returns the results to Excel.

Figure 7. ”The Excel report generator”



[21] The report generator doesn’t work under Windows 7 or higher currently.

As an example of the syntax, look at the figure above. At line 7 the number of individuals with student-status 0 is requested. The result is returned on the same line for all years chosen on line 1. The possible statistics are: Sum, mean, count, min, max, std. It is also possible to add selections in columns C to E. Actual operators are eq, lt, le, ne, gt and ge.

At line 3 a global condition is stated ”i_abroad eq 0”. A global condition applies to all the following lines until it is cancelled by a ”clear” in column C. It is possible to use several global conditions at the same time.

5.7 Data structure

(Note: To edit the model you need the source code and Visual Basic 6.0 SP 5)

There exist three types of ”system” variables in SESIM, individual, household and macro variables. Individual variables begin with ”i_”, household variables with ”h_” and macro variables with ”m_”. These variables are defined in the Visual Basic environment with an add-in. When a new variable is defined, source code is automatically generated to handle removal, adding etc. of individuals and households. That is, the variable names are ”hard coded” in the source code. Another and more traditional approach would be to store the information in a large matrix where variables were numbered columns. However, we think our solution is more convenient to non-programmers.

Individuals and households are linked together in a special way. We explain it below by using an example of a few individuals.

Below is an example of six individuals living in three households. Individual #1 lives together with #3 and #10, they have the same household number i_hhnr=1. As it can be seen, there is a link from the first individual to the next individual in the household, i_next_indnr=3 for individual #1. When a person dies, the corresponding row is deleted from all defined individual vectors. The vectors are then re-packed, i.e. the row below the dead person is shifted upward. When an individual is born it is added to all vectors.

Example of individual vectors.

Index	Individual nr i_indnr	Household nr i_hhnr	Next individual nr i_next_indnr	Age i_age	Sex i_sex	More variables
1	1	1	3	36	1	
2	3	1	10	32	2	
3	4	5	8	45	2	
4	8	5	0	47	1	
5	10	1	0	2	1	
6	12	3	0	60	1	

To make it possible to quickly find an individual by *i_indnr,* without searching, there is a pointer vector named *indnr2index* translating from individual number to the actual index number.

As an example, to get the age of individual #4: i_age(indnr2index(4))
indnr2index(4) returns 3, the row number corresponding to individual #4.

”0” in indnr2index denotes a dead person. The individual vectors consist of exactly as many rows as the current number of individuals. The pointer vector on the other hand increases every year (not a practical problem).

Example of pointer vector (indnr2index)

Vector index = individual nr	Vector index in individual vector
1	1
2	0
3	2
4	3
5	0
6	0
7	0
8	4
9	0
10	5
11	0
12	6

The same structure applies for household vectors. There is one vector *h_first_indnr* pointing to the first individual number.

Example of household vectors

Index	Household nr h_hhnr	First individual nr h_first_indnr	Household variable 1 h_var1	More variables
1	1	1	1	
2	3	3	12	
3	5	5	4	

Example of pointer vector (hhnr2index)

Vector index = individual nr	Vector index in individual vector
1	1
2	0
3	2
4	0
5	3

Some examples:

Value of household variable 1 for household #3:
h_var1(hhnr2index(3))

Age of the first individual in household #5:
i_age(indnr2index(h_first_indnr(hhnr2index(5)))

Individual number of the second person in household #5 (if the returned value is 0, there is no next person):
i_next_indnr(indnr2index(h_first_indnr(hhnr2index(5)))

5.8 Parameters

As discussed earlier, parameter values, or assumptions, are read from the two Excel files. To enter new parameters you have to create a new Excel sheet, fill it with values and add source code for reading the Excel table into some matrix or vector. The module for reading parameters is named ”Readparameters”. For a template, study for instance how ”fertility” rates are read.

5.9 Editing the source code

To edit the source code, from your VB program, open the project file sesim.vbp in the source catalogue.

Objects under Forms belong to the user interface and objects under Modules are modelling material. ”MDIForm1” is the starting form. The modules begin with a letter to guide non-programmers:

a = easy programming, rules etc.
b = easy programming, but you have to understand the structure a bit.
c = a bit harder than b.
x = not so easy, you have to understand everything, data structure etc.
xx = don’t edit, automatically generated by add-in.

5.10 Add-in

Installation

If you want to add, delete or rename SESIM system variables, the add-in has to be installed:

1. Open the DOS-prompt and change directory to *addin* below the main directory.

2. Register the add-in with the following command:
regsvr32 addinjh.dll

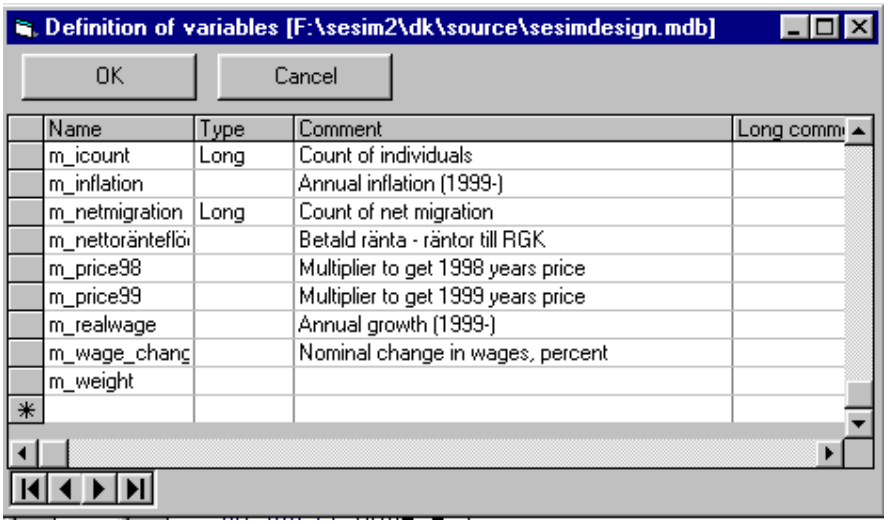
3. Add the following line to a file named Vbaddin.INI in your Windows directory:
Addinjh.AddInConnect=1

4. In Visual Basic, select from the menu, ”Add-Ins”, ”Add-in manager” and select the checkbox in front of addinjh.AddinConnect.

Running the add-in

It should now be possible to start the add-in from the menu ”Add-Ins”, when selecting ”Define

SESIM variables" the window below should appear. New variables can be added on the last row. Existing variables may be deleted if a row is selected and the delete key is pressed. When OK is pressed the source code in the xxVectordef module is regenerated with the new variable definitions.



Appendix: Summary of stochastic models in SESIM

Below a short summary is provided of all stochastic models in SESIM, we concentrate on the following characteristics:

Event/outcome: the event or the outcome variable.
At risk: individuals or household at risk, i.e. they are affected of the outcome of the model.
Model: a crude classification of the statistical model used. In SESIM the typical ones are linear regression for continuous dependent variables and logit/probit models for binary dependent variables.
Covariates: The independent variables used to explain the variation in the dependent variable. We only list the variables not their functional form.
Comment: general comments, for instance about calibration.
Algorithm: sometimes the process is a combination of stochastic models and rules or algorithms.

Mortality

At risk: individuals in age 0 to 29
Model: yearly death risks in accordance to the SCB population forecast.
Comment: due to the very low death risks in these ages, it is difficult to specify a statistical model. For this reason the average risks for each age/sex are used. These are constructed based on observed risks for the total population, in the relevant age range.

At risk: individuals in age 30 to 64
Model: logistic regression
Covariates: sex, age, indicator for early retirement, pensionable income (quintile), marital status

At risk: individuals older than 64
Model: logistic regression
Covariates: sex, age, indicator for early retirement at 64 years of age, marital status, highest level of education.

Comment: for mortality a calibration is done for gender and age group in accordance to SCB:s long-term population forecast[22].

Adoption (orphans)

Event/outcome: Household adopting a child is drawn at random from all households with an age between 24-51.
At risk:
1. children (younger than 18) that have become orphans during the year
2. adopting household, where the lowest age among the spouses is 24 and the highest 51.
Model: regression
Covariates: indicator for household type (single female, single male, non-single), female age (male age for single male).
Algorithm: Number of children are simulated for adopting households (using the estimated

[22] See SCB demographic forecast [BE18SM0201], Swedish population until year 2050.

model) and compared with the actual number of children. Orphans are adopted by these households with the largest difference between simulated and actual children.

Emigration

At risk: household that has immigrated
Model: logistic regression
Covariates: householder's highest age, number of children, number of adults, household's highest education, time since immigration

At risk: Swedish born household
Model: logistic regression
Covariates: household's highest age, number of children, number of adults, households highest education

Comment: the total number of simulated emigrants is calibrated to total number emigrants according to SCB:s demographic forecast.

Immigration

At risk: household/individuals that have emigrated
Model: hazard model
Covariates: time since emigration

At risk: first time immigrants
Algorithm: existing households in the model population are "cloned" and considered as first time immigrants. The selection of household for the matching is done such that the immigrated household achieves a reasonable composition according to the highest age and size of the household.
Comment: total immigration from the model is calibrated to total immigration according to SCB:s forecast. This is done by adjusting the number of first time immigrants within the sample of total immigration.

Fertility

At risk: women 18 to 49 year of age, who moved from her parents and have not given birth to a child
Model: logistic regression
Covariates: age, marital status, pensionable income (quartiles), indicator for market work, highest education.

At risk: women 18 to 49 year of age, who moved from her parents and have given birth to at least one child
Model: logistic regression
Covariates: age, marital status, pensionable income (quartiles), indicator for market work, highest education, age of youngest child
Comment: separate models have been estimated for women who have given birth to one, two and three children respectively, more than four births are not modeled.

Comment: The total numbers of births are calibrated against total number of children born each

year, according to the SCB forecast.

Moving from parents

At risk: individuals older than 17 living with their parents.
Model: logistic regression
Covariates: age, highest education, indicator for ongoing studies, indicator for taxable income greater than zero, nationality[23]
Comment: separate models estimated for male and women. The Model is calibrated against the observed share of movers per sex and age according to HEK 1999.

Creation of new households

At risk: adults (not living in the parent's household) women and male who are single.
Model: logistic regression
Covariates: age
Algorithm: The women who will form the new household during the year are drawn at random from the estimated probabilities. For every female, from the sample of "selectable" males the first male three years older and in the same region as the female is picked. If no match is found the process is repeated but now the male is four years older and if necessary the process can be repeated until a match is found. There is no assortative matching e.g. by education level.

Dissolution of households

At risk: non-single male
Model: logistic regression
Covariates: age
Algorithm: based on estimated probabilities of leaving the household, males are drawn at random

Disability pension

At risk: individuals between 16 and 29 years of age who are not disability pensioners
Model: logistic regression
Covariates: age, highest education, indicator for age 16, pensionable income (quartiles)

At risk: individuals between 30 and 60 years of age who are not disability pensioners
Model: logistic regression
Covariates: age, sex, pensionable income (quartiles), highest education, marital status, nationality

At risk: individuals between 61 and 64 years of age who are not disability pensioners
Model: logistic regression
Covariates: age, sex, pensionable income (quartiles), highest education

Comment: during the first year the inflow to disability retirement is calibrated to the actual number of disability retirement.

[23] Nationality is set to one if the individual is born in Sweden, else zero.

Rehabilitation from Disability pension

At risk: Individuals who are disability pensioners
Model: logistic regression
Covariates: age, time as disability pensioner, highest education

Regional mobility

Event/outcome: household moving
At risk: All household
Model: logistic regression
Covariates: age, region, income, unemployment

Event/outcome: choice of region
At risk: households who move
Model: conditional logit
Covariates: age, region, income, unemployment

Event/outcome: tenure choice (owned/rented)
At risk: households who move
Model: conditional logit
Covariates: age, region, income, unemployment
Comment: two models

Education

Event/outcome: starting gymnasium (high school)
At risk: 16-year old not disability pensioner
Algorithm: all individuals start gymnasium at 16 years of age.
Comment: at most three years of studies are required in order to obtain a gymnasium degree.

Event/outcome: exit from gymnasium
At risk: students at gymnasium
Model: logistic regression
Covariates: sex, parent's highest education, parent's highest age, nationality, indicator for divorced parents, household's income (quartiles), number of children in household
Comment: for those individuals who are predicted to exit, a time for this exit is also predicted by a random draw.

Event/outcome: starting university studies directly after gymnasium
At risk: 19 years old who finished their gymnasium degree the preceding year and who are not disability pensioners or on parental leave.
Model: logistic regression
Covariates: sex, nationality, indicator for living with parents, indicator for own children, indicator for living in big city, rate of unemployment

Event/outcome: exit from university studies before a degree
At risk: students at the university
Model: logistic regression

Covariates: sex, nationality, marital status, age, number of children, indicator for living in big city
Comment: university studies are assumed to continue three or four years. Students that exit after the third year obtain a degree, after the forth year they obtain a degree by default.

Event/outcome: transition from labour force to university studies
At risk: individuals with a degree from gymnasium who do not have a student status
Model: logistic regression
Covariates: sex, nationality, GNP-growth, time since the gymnasium degree, age, marital status, number of children

Event: transition from labour force to ”komvux” (gymnasium for adults)
At risk: individuals between 20 and 64 years of age without a degree from gymnasium, who are not disability pensioners or on parental leave
Model: logistic regression
Covariates: sex, nationality, GNP-growth, marital status, time since the gymnasium degree, number of children, age
Comment: studies at ”komvux” are assumed to continue for three years, this gives a gymnasium degree.

Event/outcome: university studies directly after ”komvux”
At risk: individuals who finished ”komvux” the preceding year and who are not disability pensioners or on parental leave.
Model: logistic regression
Covariates: sex, nationality, age, marital status, number of children, indicator for living in big city, GNP-growth

Market work

Event/outcome: market work/no market work during the year
At risk: individuals older than 15 not old age pensioners, disability pensioners, students, on parental leave or unemployed
Model: logistic regression
Covariates: indicator for market work the preceding year, age, highest education, indicator for current studies
Comment: separate models estimated for male and women.

Unemployed

Event/outcome: unemployed/not unemployed during the year
At risk: individuals older than 15 who are not old age pensioner, disability pensioner, student or on parental leave
Model: logistic regression
Covariates: indicator for unemployment the preceding year, age, age group, highest education
Comment: separate models estimated for male and women.

Occupational sector[24]

At risk: individuals in labour force without an occupational sector
Model: multinomial logit
Covariates: age, sex, highest education, nationality

At risk: individuals in labour force with an occupational sector
Model: multinomial logit
Covariates: age, sex, highest education, nationality
Comment: Separate models estimated for each occupational sector

Housing/wealth

Event/outcome: probability of financial wealth
At risk: households without financial wealth, who did not change real wealth >= 100 000 SEK between 1999 and 2000.
Model: logistic regression
Covariates: age group (5-year intervals), quartiles and percentiles for taxable income (p25, p50, p75, p90, p95, >p95)

Event/outcome: financial wealth
At risk: households with a positive financial wealth predicted from the preceding model
Model: robust regression (M-estimation), logarithm of response variable
Covariates: age group (5-year intervals), quartiles and percentiles for taxable income (p25, p50, p75, p90, p95, >p95)
Comment: no Monte Carlo simulation.

Event/outcome: financial wealth
At risk: Households with a positive financial wealth (the previous year), no change in household composition, not sold or bought a home.
Model: A dynamic random effect model
Covariates: lag of financial wealth, interaction with lag financial wealth and age and income and finally a variable for general index on Stockholm stock exchange.

Event/outcome: probability of other real wealth
At risk: households without other real wealth, who did not change real wealth >= 100 000 SEK between 1999 and 2000.
Model: logistic regression
Covariates: age group (5-year intervals), marital status, quartiles and percentiles for taxable income (p25, p50, p75, p90, p95, >p95)

Event/outcome: other real wealth
At risk: households with a positive other real wealth predicted from the preceding model
Model: robust regression (M-estimation), logarithm of response variable
Covariates: age, quartiles and percentiles for taxable income (p25, p50, p75, p90, p95, >p95)
Comment: no Monte Carlo simulation.

[24] The sectors are blue-collar workers in the private sector, white-collar workers in the private sector, central government employees, local government employees and own employed.

Event/outcome: other real wealth
At risk: Households with a positive other real wealth (the previous year), no change in household composition, not sold or bought a home.
Model: A random walk

Event/outcome: probability of private pension savings
At risk: individuals between 18-64 year without private pension savings.
Model: logistic regression
Covariates: age, age-square, sex, education, marital status, nationality, quartiles and percentiles for taxable income (p25, p50, p75, p90, p95, >p95)

Event/outcome: private pension savings
At risk: households with a positive private pension savings predicted from the preceding model
Model: regression private pension savings/1000
Covariates: Covariates: age, sex, education, nationality, quartiles and percentiles for taxable income (p25, p50, p75, p90, p95, >p95)
Comment: In forecasting accumulated pension savings we have to estimate the probability and the amount saved first time. Then we assume that the individual save the same amount (adjusted by CPI) each year until age 64.

Event/outcome: house area
At risk: household who bought a house during the year
Model: regression
Covariates: age group for oldest spouse, indicator for married/cohab, number of children below 18, quartiles for taxable income, indicator for living in Stockholm, Göteborg or Malmö dense areas (H-region)
Comment: no Monte Carlo simulation, i.e. the variance is assumed equal to zero.

Event/outcome: market value of house/apartment
At risk: household who bought house/apartment during the year
Model: regression, log transformed response variable
Covariates: age group for oldest spouse, indicator for married/cohab, number of children below 18, quartiles for taxable income, indicator for living in dense areas (H-region), quartiles for households financial wealth

Event/outcome: probability of debt
At risk: households without debt.
Model: random effect probit
Covariates: age group, household taxable income / median, household gross wealth / median and household gross wealth / median squared.

Event/outcome: debt
At risk: households with a positive debt predicted from the preceding model
Model: robust regression (M-estimation), logarithm of debt
Covariates: age group, household taxable income / median (lag of household debt)

Event/outcome: debt
At risk: Households with a debt (the previous year) and a reduction in market value of home more than 100 000 SEK.

Model 1: probit for probability of debt
Model 2: robust regression for level
Covariates: age group, marital status, household taxable income / median, household gross wealth / median and household gross wealth / median squared.

Event/outcome: debt
At risk: Households with a debt (the previous year) and no large changes in market value of home.
Model 1: probit for probability of debt
Model 2: robust regression for level
Covariates: age group, marital status, household taxable income / median, household gross wealth / median and household gross wealth / median squared.

Event/outcome: debt
At risk: Households with a debt (the previous year) and an increase in market value of home more than 100 000 SEK.
Model 1: probit for probability of debt
Model 2: robust regression for level
Covariates: age group, marital status, household taxable income / median, household gross wealth / median and household gross wealth / median squared.

Event/outcome: interest/dividends
At risk: all households
Comment: Interests and dividends are simulated as a rate that, multiplied to the households financial wealth, returns the amount of obtained interests and dividends. Due to difficulties in finding a suitable statistic model the rates are simulated from the empirical distribution

Event/outcome: probability of income of capital (excluding capital gain on own home)
At risk: all households
Model: probit
Covariates: ?

Event/outcome: income of capital (excluding capital gain on own home)
At risk: households with a positive value predicted from the preceding model
Model: robust regression (M-estimation), logarithm of income of capital
Covariates: ?

Event/outcome: probability of capital loss (excluding capital loss on own home)
At risk: all households
Model: probit
Covariates: ?

Event/outcome: capital loss (excluding capital loss on own home)
At risk: households with a positive value predicted from the preceding model
Model: robust regression (M-estimation), logarithm of income of capital
Covariates: ?

Event/outcome: probability of interest deductions
At risk: households without interest deductions t-1.
Model: random effects probit

Covariates: age group, education, debt ratio, dummy if debts > total wealth, market value own home / median value, household taxable income / median

Event/outcome: interest deductions
At risk: households with a positive debt predicted from the preceding model
Model: random effects GLS
Covariates: age group, education, debt ratio, dummy if debts > total wealth, market value own home / median value, household taxable income / median, household financial wealth/median, change in market value of home/median

Event/outcome: probability of interest deductions
At risk: households with interest deductions t-1.
Model: random effects probit
Covariates: age group, education, debt ratio, dummy if debts > total wealth, market value own home / median value, household taxable income / median

Event/outcome: interest deductions
At risk: households with a positive value predicted from the preceding model
Model: random effects GLS
Covariates: age group, education, debt ratio, dummy if debts > total wealth, market value own home / median value, household taxable income / median, household financial wealth/median, change in market value of home/median

Earnings[25]

Event/outcome: earnings
At risk: individuals in labour force
Model: regression (mixed regression), log transformed response variable
Covariates: working experience, highest education, occupational sector, nationality, marital status, random intercept
Comment: a separate model for each gender and separate estimations of the variance components for each occupational sector

Event/outcome: probability of earnings
At risk: students
Model: probit regression
Covariates: age, sex, nationality

Event/outcome: earnings
At risk: students predicted to have earnings from previous model
Model: regression, log transformed response variable
Covariates: age, sex, nationality

Event/outcome: probability of earnings
At risk: individuals classified as other (status=7)
Model: probit regression
Covariates: age, sex, nationality, highest education

[25] The earnings equation was discussed in section 4.1.

Event/outcome: earnings
At risk: individuals classified as other and predicted to have earnings from previous model.
Model: regression, log transformed response variable
Covariates: age, highest education

Compensated sickness days

Event/outcome: probability of compensated sickness days
At risk: individuals between 16 and 64 years of age who are not disability- or old age pensioners or on parental leave.
Model: probit regression (mixed regression)
Covariates: indicator for positive number of sickness days preceding year, age group, highest education, sex, rate of unemployment, indicator for living in big city, marital status, nationality, random intercept

Event/outcome: number of compensated sickness days
At risk: individuals predicted from preceding model
Model: regression, log transformed response variable
Covariates: age group, sex, indicator for children in household, highest education, indicator for living in big city, marital status, nationality, working experience, indicator for sector (governmental/municipal), indicator for own employed, indicator for ongoing studies, indicator for unemployment, GNP-growth

Take-up social assistance[26]

At risk: household with an adult whose income is below the norm for social assistance
Model: logistic regression
Covariates: age, sex, number of children, number of children under 7 years of age, highest education, indicator for unemployment, the difference between disposable income and the norm for social assistance, working experience, indicator for ongoing studies, nationality, indicator for living in big city, indicator for any earlier divorce

At risk: household with two adults and a disposable income below the norm for social assistance
Model: logistic regression
Covariates: male age, female age, number of children, number of children below 7 years of age, male highest education, female highest education, indicator for male unemployment, indicator for female unemployment, the difference between disposable income and the norm for social assistance, male working experience, female working, male nationality, female nationality, indicator for living in big city

Non-cash benefits

Event/outcome: subsidy for school (day care, primary school)
At risk: all individuals 6 to 15 years of age
Model: regression (expected value[27])

[26] Take-up denotes the propensity to apply for a benefit given that the individual/household is entitled to it. In Social assistance it is quite common that a large portion of eligible household do not apply.

Covariates: age, age group

Event/outcome: subsidy for gymnasium
At risk: students at gymnasium
Model: regression (expected value)
Covariates: sex, nationality

Event/outcome: subsidy for komvux
At risk: students at komvux
Model: regression (expected value)
Covariates: sex, nationality, age

Event/outcome: probability of subsidy for adult studies (including särvux, statens skola för vuxna, svenska för invandrare, folkbildning/folkhögskola, kvalificerad yrkesutbildning)
At risk: individuals age 20 to 50 who do not participate in other studies
Model: logistic regression
Covariates: sex, age, highest education, nationality

Event/outcome: subsidy for adult studies
At risk: individuals predicted from previous model
Model: regression (expected value)
Covariates: sex, age, highest education, nationality

Event/outcome: subsidy for university.
At risk: students at university
Model: regression (expected value)
Covariates: sex, age

Event/outcome: probability of subsidy for childe care
At risk: individuals in age 1 to 12
Model: logistic regression
Covariates: age, nationality, household's highest education, household's degree of employment

Event/outcome: subsidy for child-care
At risk: individuals predicted from previous model
Model: regression (expected value)
Covariates: age, age group, nationality, household's highest education

Event/outcome: probability of subsidy for care of older (hemtjänst, särskilt boende)
At risk: individuals in age 65 and older
Model: logistic regression
Covariates: age, age group, sex, nationality, income from pension (quintiles), marital status, indicator for employment

Event/outcome: subsidy for care of older
At risk: individuals predicted from previous model
Model: regression (expected value)

---

[27] For these models the simulation is done based on the prediction of the expected value, conditional on the X-variables. The components of variance are ignored here, due to the difficulty to estimate their size.

Covariates: age, income from pension (quintiles), sex, nationality

Event/outcome: probability of subsidy for labour market activities
At risk: unemployed
Model: logistic regression
Covariates: sex, highest education, nationality

Event/outcome: subsidy for labour market activities
At risk: individuals predicted from previous model
Model: regression (expected value)
Covariates: age, highest education, nationality

Event/outcome: subsidy for health/care (primärvård, slutenvård, tandvård)
At risk: individuals 19 years and younger
Model: regression (expected value)
Covariates: age, sex, nationality, household's highest education
Comment: All individuals 19 years and younger are assumed to obtain this subsidy

Event/outcome: probability of health/care subsidy
At risk: individuals in age 20 to 64
Model: logistic regression
Covariates: age, sex, indicator for disability pension, indicator for employment, nationality, marital status, taxable income (quintile)

Event/outcome: subsidy for health/care
At risk: individuals age 20 to 64 predicted from previous model
Model: regression (expected value)
Covariates: age, age group, sex, nationality, highest education, taxable income (quintile)

Event/outcome: probability of health/care subsidy
At risk: individuals 65 or older
Model: logistic regression
Covariates: age, sex, nationality, marital status, old age pension (quintile)

Event/outcome: subsidy for health/care
At risk: individuals 65 or older predicted from previous model
Model: regression (expected value)
Covariates: age, sex, nationality, marital status, old age pension (quintile)

Event/outcome: subsidy for medicine
At risk: all individuals
Model: given as the average value for the whole population per sex and age group (0-4, 5-9,…,90-)

Health and care of elderly

Event/outcome: probability that a relative lives in the same labour market region
At risk: individuals 65 or older
Model: logistic regression
Covariates: region, marital status, own/rent, nationality, age group, income and education

Comment: the model is applied 1999 or then the individual enters the population at risk.

Event/outcome: probability that a relative lives in the same labour market region
At risk: individuals 65 or older
Model: dynamic logistic regression
Covariates: region, marital status, own/rent, nationality, age group, income, education and closeness to relative t-1
Comment: the model is applied after 1999.

Event/outcome: health index (0=unknown, 1=severe illness, 2=some illness, 3=not full health and 4=full health.)
At risk: all individuals
Model: ordered probit
Covariates: region, marital status, own/rent, nationality, age group, taxable income, education, number of children and sex
Comment: the model is applied 1999-2006 or if the individual enters the population at risk. Different model depending on if the individual is older or younger than 50.

Event/outcome: health index (0=unknown, 1=severe illness, 2=some illness, 3=not full health and 4=full health.)
At risk: all individuals with a value on health index t-8
Model: ordered probit
Covariates: region, marital status, own/rent, nationality, age group, taxable income, education, number of children, sex and health index t-8.
Comment: the model is applied from 2007 for individuals with a value on health index t-8. Different model depending on if the individual is older or younger than 50.

Event/outcome: number of days in patient care
At risk: all individuals older than 60
Model: Zero inflated negative binomial
Covariates: health index, age, taxable income / mean, education, marital status, number of children, region, sex and nationality.
Comment: the model is applied 1999 and for new individuals. Different models depending on the individual between 16-49 or older than 50.

Event/outcome: number of days in patient care
At risk: all individuals older than 60
Model: Zero inflated negative binomial
Covariates: health index, age, taxable income / mean, education, marital status, number of children, region, sex and nationality, days in patient care t-1.
Comment: the model is applied after 1999 and for new individuals. Different model depending on if the individual is between 16-49 or older than 50.

Event/outcome: ADL (0=unknown, 1=non-disabled, 2=slightly disabled, 3=moderately disabled and 4=severely disabled)
At risk: all individuals 65 or older
Model: ordered logit
Covariates: health index, age group and sex

Event/outcome: Assistance elderly (0=no assistance, 1=assistance at home and 2=special

accommodations)
At risk: all individuals 75 or older
Model: multinomial logit
Covariates: ADL, previous level of assistance, number of years since the individual became 75, closeness to relative.
Comment: The initial level of assistance is imputed based on observed frequencies per age, sex and ADL-level.

Literature

Andersson, B., Berg, L., and Klevmarken, A., [2001] "Inkomst- och Förmögenhetsfördelningen för dagens och morgondagens äldre" www.nek.uu.se/faculty/klevmark/reswork.html.

Baltagai, B.H.,[2001] "Econometric Analysis of Panel Data", John Wiley and Sons Ltd.

Davison, A. C. & Hinkley, D. V. [1997], Bootstrap Methods and Their Application, Cambridge University Press.

Diggle, P. J., Liang, K. Y. & Zeger, S. L. [1994], "Analysis of Longitudinal" Data, Oxford University Press.

Eklöf, M. & Hallberg, D. [2004] "Private Alternatives and Early Retirement, Working paper Uppsala

Ericson, P. & Hussénius, J. [2000] "Studiebidragen i det långa loppet", Rapport till Expertgruppen för studier i offentlig ekonomi (ESO). Ds 2000:19,

Flood, L., and Gråsjö, U., [2001] " A Monte Carlo simulation study of Tobit models", Applied Economics Letters, 8, 581-584.

Flood, L., [2003] "Formation of Wealth, income of capital and cost of housing in SESIM", SESIM working paper, www.sesim.org.

Galler, H. P. [1997], "Discrete-Time and Continuous-Time Approaches to Dynamic Microsimulation reconsidered", NATSEM Technical Paper No. 13, October, National Center for Social and Economic Modeling, University of Canberra.

Edin, P. A. and Fredriksson, P., [2000], "LINDA – Longitudinal Individual DAta for Sweden", Working Paper 2000:19, Uppsala Universitet, Nationalekonomiska institutionen.

Jansson, F., [2003] "Modeling the Retirement Decision in Sweden". Paper presented at International Microsimulation Conference on Population Ageing and Health in Canberra.

Johannesson, I., [2001] "The Impact of Wealth on Tax-deferred Pension Saving- Cross Section Evidence from Sweden" University of Gothenburg.

Klevmarken, A. [1997], "Behavioral Modeling in Micro Simulation Models. A survey", Working Paper 1997:31, Nationalekonomiska institutionen, Uppsala Universitet.

Klevmarken, A. [1998], "Statistical Inference in Micro Simulation Models: Incorporating External Information", Working Paper 1998:20, Nationalekonomiska institutionen, Uppsala Universitet.

Manning, W. G., N Duan, and Rogers., [1987] "Monte Carlo evidence on the choice between sample selection and two-part models". Journal of Econometrics 35: 59-82.

Merz, J., [1991], ”Micro-simulation - A survey of principles, developments and applications”, *International Journal of Forecasting* 7pp 77-104

Ministry of Finance, [1999], “Fördelningspolitisk redogörelse”, appendix 4 to The Govenrment Budget Bill 2000, Regeringens proposition (1999/00:1).

Ministry of Finance, [2002], “Fördelningspolitisk redogörelse”, appendix 3 to The Govenrment Spring Fiscal Policy Bill 2002, Regeringens proposition (2001/02:100).

O´Donoghue, C., [2001] “Dynamic microsimulation: A Methodological survey”, Brazilian electronic journal of economics, Vol 4.

Pettersson, T., and Pettersson, T., [2003a] “Lifetime Redistribution Through Taxes, Transfers and Non-cash Benefits”, Paper presented at International Microsimulation Conference on Population Ageing and Health in Canberra.

Pettersson, T., and Pettersson, T., [2003b] “Fördelning ur ett livscykelperpektiv. Appendix 9, Long Term Survey, Swedish Ministry of Finance.

Smeeding, T., Saunders, P., Coder, J., Jenkins, S., Fritzell, J., Hagenaars, A., Hauser, R. and Wolfson, M., [1993], “Poverty, Inequality, and Family Living Standards Impacts Across Seven Nations: The Effect of Non-cash Subsidies for Health, Education and Housing”, The Review of Income and Wealth, Series 39, No 3.

Statistics Sweden, [2003] “Inkomstfördelningsundersökningen 2001”, HE 21 SM 0301.

Sutherland, H., [1995] “Static Micro-simulations Models in Europe: A survey”, DAE Working Paper No 9523, University of Cambridge.

The Swedish Consumer Agency, ”Slutrapportering av regeringsuppdrag rörande hushållens pensionssparande”, Dnr 1999/2339 (1999).

www.sesim.org

Zaidi, A & Rake, K., [2001] “Dynamic Microsimulation Models: A Review and Some Lessons for SAGE”, SAGE Discussion Paper no. 2 SAGEDP/02, www.lse.ac.uk/depts/sage