Working papers 44/2013

Meng Zhou

Microsimulation

Statistical methodology and assessment of uncertainty

Kela, Research Department

Author
Meng Zhou
University of Helsinki
firstname.lastname@gmx.com

www.kela.fi/research

ISSN 2323-9239

Helsinki 2013

Acknowledgements

From October 2011 to February 2012, the theoretical part of my thesis was first carried out in University of Helsinki. From February to August 2012, the empirical part of my thesis was done and improved in Finnish Social Insurance Institution. The successful completion of this thesis within these 10 months is not only a result of one person's effort, whereas it also concentrates others' support and encouragement. I am sincerely grateful to my supervisor, Prof. Risto Lehtonen. Thanks him a lot for the professional guidance, encouragement, and inspiration given to me during my whole work. I also would like to express my deep gratitude to other supervisors Dr. Pertti Honkanen and Researcher Maria Valaste, who gave me the most confidence and the expertise insight on my thesis. My sincere and deep gratitude is also extended to Prof. Anita Haataja, from whom I obtained a strong support and was granted to getting to know the general concept of the studied social benefits in Microsimulation project in Finnish Social Insurance Institution. And my thanks are meanwhile extended to all friendly staff, with whom I have enjoyed one of my happiest period in Finland.

Additionally, I would like to express my acknowledgement and respect to all those tutors who ever taught me before.

Meng Zhou

Helsinki, April 2013

Abstract

Nowadays, microsimulation method has been introduced to different fields, such as Social Science, Medicine research and Economic study. This method evaluates the effects of the proposed interventions or policies before they are implemented in the real world. In this article, I concentrate on microsimulation method used in Social Science by firstly explaining two main streams in microsimulation world, Static approach and Dynamic approach, in them, how statistical models are used are carefully explained by giving examples in Dynamic approach. In the following section, a Norwegian case is studied, this case gives the typical example of how the dynamic microsimulation used in the labor force and child care research, the effects of four different reform options are measured in this study. In the last section, the empirical study of a Finnish static microsimulation model-JUTTA is carried out. The uncertainty of JUTTA is assessed and one of its sub model called Toimtuki (income-related supplementary benefit) is detected to have space to be more accurate. In order to do so, two statistical models- Linear Regression model and Two-Stage Least Squares (2SLS) model are applied to it. From their results and diagnostics, we could conclude that both the Linear Regression and 2SLS successfully improves the accuracy of TOIMTUKI to some extent.

Key words: static microsimulation, dynamic microsimulation, JUTTA, assessment, 2SLS

Contents

1 Introduction

1.1 The aim of the study

The theoretical part of the thesis is beginning with the introducing two main types of microsimulation-Static microsimulation and Dynamic microsimulation, and then going deeper into the Dynamic microsimulation, followed by one application: Norwegian special case with mother labor supply and child care. Next, a practical framework is carried out by assessing the accuracy of the JUTTA model and model improvement.

1.2 Why microsimulation?

Briefly speaking, the purpose of the microsimulation is mainly to evaluate the effects of the proposed interventions or policies before they are implemented in the real world. By using microsimulation, people can easily estimate the impacts of a new scheme by producing outputs on a wide range of measures of effectiveness.

1.3 What is microsimulation?

A microsimulation model differs from other types of models in that it operates on individual units rather than on aggregate information (TRIM3 2012a). Typically, in social sciences, those units are individual substantial or economic units. The database used as input to a microsimulation model contains records describing persons, households or business. And the simulation model applies a set of rules to each individual record. The result of the computations might be the amount of taxes owed by the unit to which the unit is entitled under certain government legislation. Also, if we are interested in the total tax, each individual result should be multiplied by whatever weight is associated with the unit in the microdata file, then the weighted individual results are added together to obtain the aggregate result. Thus, different policies could apply to the same microdata file, and the report of the comparisons among the different results could be a good helper to the wise government.

2 State of the art

Generally, there are two main streams in this field: static microsimulation and dynamic microsimulation.

2.1 The static microsimulation approach

This kind of model suits for performing detailed simulations of past, the present, and the near future. It typically use static aging techniques, changing certain variables on the original microdata file to produce a file with the demographic and economic characteristics expected in the future year. Person weights are modified to change the total population and the weighted characteristics of the population; labor force status may be changed to alter the unemployment rate; and incomes are adjusted for price changes. Simulations can then be run on the aged microdata file to estimate the impact of a change to be implemented in the future year (TRIM3 2012b). Compared with the dynamic microsimulation, the static microsimulation does not take account individual behavior of people, which means that it is assumed the rules are obeyed 100% without any variation.

Currently popular used models and their main properties are presented in Table 1 (p. 8).

2.2 The dynamic microsimulation approach

Dynamic microsimulation models age each person in the microdata file from one year to the next by probabilistically deciding whether or not that person will get married, get divorced, have a child, drop out of school, get a job, change jobs, become unemployed, retire, or die, then the same procedure is repeated as many times as the user wants to achieve the final simulation year. Simulations of government legislations can be run in the current year, the final year of the aging process, or any interim year. The simulation of the government program in one year may affect a person's characteristics in the subsequent year (TRIM3 2012b). For example, whether or not someone will drop out of school could be programmed to depend partly on family income, which could, in turn, be affected by government transfer payments. This kind of models could create the synthetic database for a future year, which is capable of performing simulations into the distant future, but it couldn't capture as much details as static models do.

Table 1. Summary of static models.

Country	Model Name	Aim of use	Initial Database	Programming	Sample size	Organization
Finland	JUTTA	Taxation, Social transfers	Registered data	Access, C++, Excel	10000 households	KELA
	SOMA	Taxation, Social transfers	Registered data	SAS		Finnish Ministry of Social Affairs and Health
	TUJA	Taxation, Social transfers	Registered data	APL		Finnish Ministry of Finance
	ASUMISTUEN MALLI	General house benefit	Registered data	SAS	165000 households	KELA
	HVS	Personal taxation	VTL database	Oracle/Forms 10g	120000, 80000 households	Finnish Tax Administration
	UUSI MALLI	Taxation, Social transfers		SAS		Statistics Finland
European	EUROMOD	Income and taxation	Survey and registered data	Excel, Access, C++	Varied	University of Essex
Canada	Mu.Sim	Income and taxation	SLID survey data	APL, Dyalog APL 10		Department of Finance
	SPSD/M	Income, direct and indirect tax	Survey and registered data	C++	200000, 80000 households	Statistics Canada
United States	ITEP	Direct,indirec, property taxation	Survey and registered data		365000	Institution on Taxation and Economic Policy
	MATH	Income	Survey data		69000, 26000 households	Mathematica Policy Research
	TRIM3	Income and taxation		Perl, MySQL, C++	206000, 76000 households	Urban Institute
Country	Model Name	Aim of use	Initial Database	Programming	Sample size	
Australia	STINMOD	Income, direct and indirect tax	Survey data	SAS	23263 households	NATSEM
Sweden	FASIT	Income, direct and indirect tax	Registered data	SAS	38000, 17000 households	Statistics Center
Germany	FiFoSim	Income, tax, labor market	Survey data		3000000 households	University of Cologne
	STSM/Pace-L	Income, tax, labor market	Survey data	STATA	12000 households	The Centre for European Economic Research(ZEW)
Great Britain	POLIMOD	Income, taxation	Survey data	C, SAS, Excel	62400, 26900 households	University of Essex
	TAXBEN	Income, direct, indirect tax, labor supply	Survey data	Delphi, Stata	63300	Institute for Fiscal Studies
	Virtual Economy	Income, taxation	TAXBN(Survey data)			Institute for Fiscal Studies
Spain	GLADHISPANIA	Income, taxation	Survey data	Visual Basic, JAVA	6420 households	University of Balearic Island
Denmark	Lovmodellen	Income, taxation	Registered data	SAS	179000	Ministry of Finance
France	INES	Income, direct and indirect taxation	Survey data	SAS	80000, 28000 households	INSEE
Irland	SWITCH	Income, taxation				Economic and Social Research Institute

Source: Honkanen Pertti. Katsaus malleihin. Helsinki: Kela, unpublished document.

In Dynamic microsimulation models the transition probabilities play the important role, because they are used to create the synthetic database about the individuals' life paths on the demographic events, personal events and so on.

Table 2 (p. 10) shows the famous dynamic models in the world.

2.3 Small Area Estimation through spatial microsimulation model

The term spatial microsimulation refers to a set of techniques that allow the characteristics of individuals living in a particular area to be approximated, based on a set of "constraint variables" such as auxiliary variables that are known about the area. Small area estimation (SAE) is one technique in the spatial microsimulation. The spatial microsimulation could be either static or dynamic. (Ballas et al. 2005.)

Small area estimation* could be divided into two big groups, design based method and model based method. Direct small area estimation is based on both of these two methods and it includes two common estimators, HT estimator (Horvitz-Thompson Estimator) (Lehtonen 2004), GREG(Calibration), whereas in indirect estimation, the "borrowing strength" is applied, and the common used models are synthetic estimators, E-BLUE(Lehtonen and Veijanen 2009) estimators and so on. However, all the models mention above are statistical models. In contrast to statistical approaches, there is also a so called geographic approach based on the microsimulation models, which are essentially creating synthetic simulated micro population data to produce 'simulated estimates' at small area level. To obtain the reliable microdata at small area level is the key task for microsimulation modeling, where the reweighting and synthetic reconstructions are two commonly used methods. (*Small area estimation: the term" small area" refers to a small geographical area or a small sample size. If a survey has been carried out for the population as a whole, the sample size within any particular small area may be too small to generate accurate estimates from the data. To deal with this problem, the "borrow" strength is used, which means borrow the additional information from the other domains.)

Table 2. Summary of dynamic models.

Country	Model Name	Aim of use	Initial database	Programming	Sample size	Organization
Australia	APPSIM	Demography, pensions	Survey data		188000	NATSEM
	DYNAMOD	Demography, labor market, education, wealth	ABS dataset	C	150000	NATSEM
Canada	DYNACAN	Demography, pensions	Registered data	C, SAS, TPL	212000	HRDC
	LifePaths	Demographic development, income	Survey and registered data	C++		Statistics Canada
United States	CORSIM	Demography, social health	Survey data	C	180000	Social Security Administration
	PENSIM	Pensions	Survey data	C++	100000	Policy Simulation Group
Sweden	LaMPSim	Labor market	Registered data	C#		Arbetsförmedlingen
	SESIM	Demographic development, labor market, pension	Registered data	Visual Basic	786000	Department of Finance
Great Britain	PENSIM2	Private and State pensions	Survey data	SAS, Excel		DWP
	SAGE MODEL	Demography, pensions	Survey data	C++	53985	University of Southampton
Norway	MOSART	Demographic development, pensions	Registered data	C#	40000	Statistics Center
	TRIM	Demography, pensions, labor supply	Registered data	SAS		NAV
Country	Model Name	Aim of use	Initial database	Programming	Sample size	Organization
Denmark	DENSIM	Demography, pensions	Registered data		147000	SFI
	DREAM	Demography, income, taxation, public economy	Registered data			DREAM Group
	MILASMEC	Direct taxation	Registered data		3500000	Danish Economic Council
France	DESTINIE	Demography, labor market, pensions	Survey data		16300, 66000 households	INSEE
Irland	LIAM	Pensions		C++, Excel		RERC
Japan	INAHSIM	Demographic development, pensions	Survey data	FORTRAN	128000, 49000 households	Research Institute for Policies on Pension and Aging

Source: Honkanen Pertti. Katsaus malleihin. Helsinki: Kela, unpublished document.

The reweight algorithm is listed below:

If the given sampling design weights are for element k, $d_k=1/i_k$, ($k\in s$), where i_k is the inclusion probability, and the we have to generate a new set of weights $w_k$for k is in the sample s, so for the calibration equation

$\sum_{k\in s} w_k x_k = T_x$,

Where d_k is the design weight, and T_x is the known population Total of the auxiliary variable x. The new weights w_k will be as close as possible to d_k.

The distance measure, for example, Azizur (2009) uses it in the GREGWT algorithm is known as truncated Chi-squared distance function and it can be defined as

$G_k=(w_k-d_k)^2/2d_k$
for $L_k \leq w_k/d_k \leq U_k$

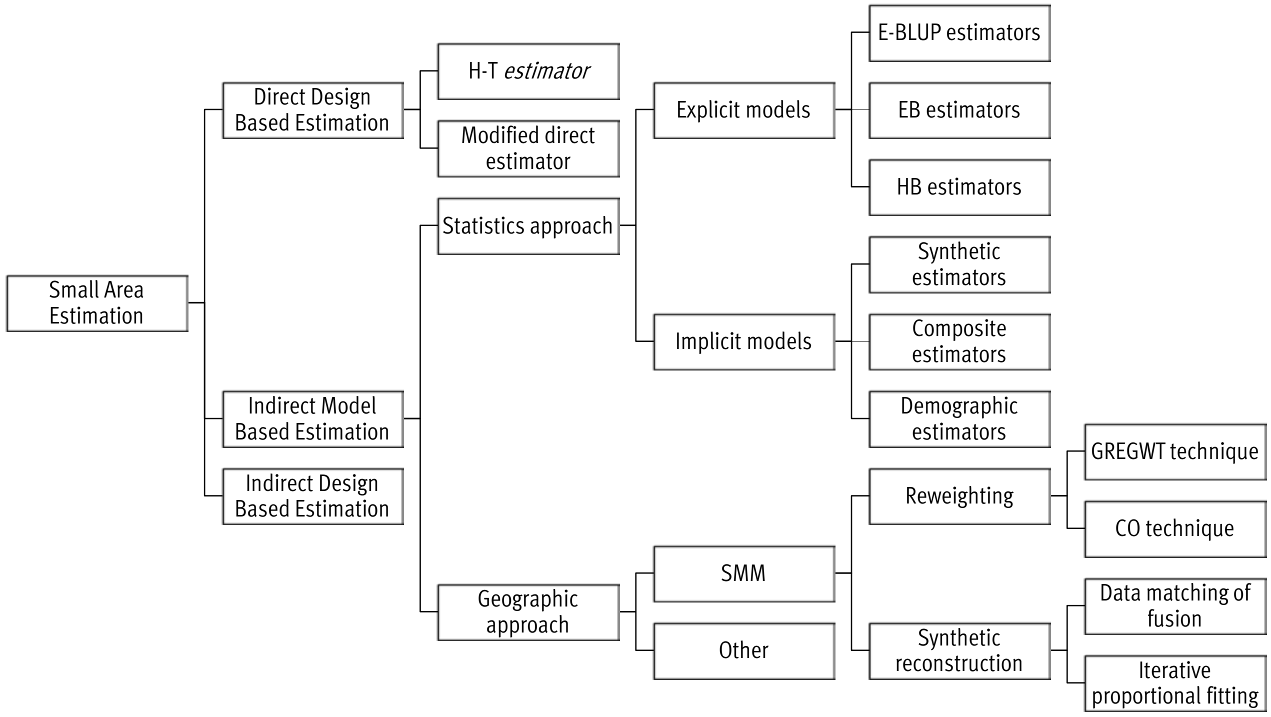
where L_k and U_k are pre specified lower and upper bounds respectively for each unit $k\in s$ (Azizur et al. 2010).

Figure 1 (p. 12) illustrates the microsimulation model (geographic approach) and its position in the whole small area estimation world (Azizur 2009).

3 Dynamic microsimulation procedure

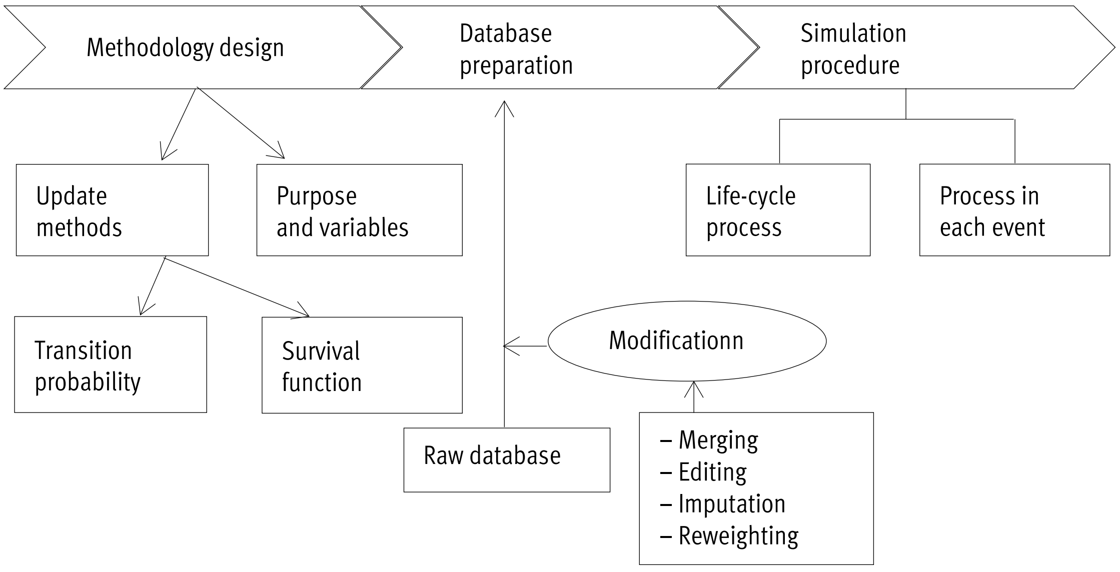
There are three components in Dynamic Microsimulation: methodology design, database preparation and simulation procedure. Figure 2 (p. 12) presents the basic structure.

Figure 1. Small area estimation methods in spatial microsimulation.



Source: Azizor 2009.

Figure 2. Basic structure.



3.1 Methodology design

The design for the model should suit for obtaining the desired results. Generally, the demographic items such as age, gender, fertility, mortality are default variables in the simulation. The other necessary variables are inputted in the original database according to which direction the model wants to estimate. The directions are often classified into three main fields: Taxation, Pension and labor market. In some cases, different directions could be combined in one model, for instance, the SESIM (a simulation model of the Swedish population) contains both labor market and pension system.

A mainstream dynamic microsimulation in the sense is that the variables (events) are updated in a sequence based on transition probabilities, and the space in time between the updating processes is a year. There are plenty models in this group: DYNACAN, SESIM, LIFEPATHS, TRIM3. Also, there exist the outliers that the events happen do not depend on the transition probabilities, but survival functions, such as DYNAMOD-2. The transition probabilities are simulated by the Monte Carlo method, which is a key point in the dynamic microsimulation and it will be discussed in details later in this paper. A small example of using survival function will also be illustrated in this paper as well, see section 3.4 *Fertility*.

3.2 Database preparation

3.2.1 Raw database

Normally, the initial survey data (register datasets) could not be applied to the microsimulation process directly, they needs some modification. But, firstly, we will start choosing the suitable raw database from different resources. The important figure for the raw database is that it should represent the whole population as close as possible.

3.2.2 Modification

Modification refers here to merging of different data sources, editing and imputing the wrong or missing information.

Merging: If the initial database contains data from different raw databases, the raw databases should be merged into one piece dataset. This case usually comes from the family-rooted model, such as

DYNACAN, The resulting file consists of individuals grouped within family units in term of both demographics and the initial distribution of employment earnings. (Rick 1998.)

Editing: Deletion often used in difficult cases, such as children whose parents could not be identified, and married people for who no partner could be identified. Addition applies when a non-identified partner was indicated as being temporarily absent, however, partners were created with reference to the characteristics of the partners of similar. Notice here, that the weight for each observation will be changed according to addition and deletion automatically. (Anthony et al. 1999.)

Imputation: The imputed characteristics include the state of the residence, disability status, earnings and a number of variables according to the case conditions. For instance, usually, the earnings for individuals are grouped into range level, where their accurate numbers are unknown, therefore the exact earnings are imputed from the official Income distribution survey based on the individuals' earnings in the previous period given the condition that people remain in employment. The other typical case is children' age specification from group classification, this could be imputed from their education levels.

3.3 Simulation procedure

Simulation procedure is the core in microsimulation model. In this step, the relevant demographic and economic events will be simulated. And in simulation procedures, the statistical models are used.

3.3.1 Applications of generalized linear model and mixed model in microsimulation

Linear regression: In statistics, the linear regression is a method to modeling the relationship between a scalar variable y and one or more explanatory variables vector *X*. In linear regression, data are modeled using linear functions, and unknown model parameters are estimated from the data. The function could be denoted as:

$Y=X\beta+\epsilon$,

where X is a matrix of explanatory variables, Y is the study vector and β is the parameter vectors which should be estimated, ϵ is the error term. It has been widely used in estimating the transition probabilities in dynamic microsimulation. In this paper, Section 3.4, *Young people leaving home* is one application of linear regression. X is the personal characteristics matrix, β is its coefficient vector and Y is the binary vector, with value 0 and 1. β could be estimated by fitting the current year known Y and X using ordinary least square method, that is:

$$\widehat{\beta} =(X^{'}X)^{-1}X'y.$$

Generalized linear model (GLM): The generalized linear model is a flexible generalization of ordinary linear regression. The linear model can be transformed to a generalized linear model by link function $g()$. The model could be represented as:

$$E(Y)=\mu=g^{-1}(X\beta),$$

Where $E(Y)$ is the expected value of Y, $X\beta$ is the linear predictor, a linear combination of unknown parameters vector β, $g()$ is the link function.

There are many common used link functions, listed below:

Identity
$g(\mu)=\mu$
Log
$g(\mu)=log(\mu)$
Logit
$g(\mu)=log(\mu/(1\text{-}\mu))$
Probit
$g(\mu)=\phi^{-1}(\mu)$
Log-Log
$g(\mu)=log(\text{-}log(1\text{-}\mu))$

The link functions are chosen wisely by the respondent data denoted as y. The common situations are:

Binary response: the outcomes are zeros and ones, it holds

$P(Y=1)=E(Y)=g^{-1}(X\beta)$,

where the possible values of the inverse link function $g^{-1}(\,)$ belongs to the interval(0,1). The commonly used linked functions are the logit link;

Multinomial response: the probit link;

Count response: log link;

Continuous response: Identity link.

Here the Table 3 shows us the general rules when considering applying the link functions.

Table 3. Applications of link functions.

y	Distribution	Name	Link function
Countable	Poisson	Log	$g(\mu)=log(\mu)$
Continuous	Normal	Identity	$g(\mu)=\mu$
Binary	Binomial	logit	$g(\mu)=log(\mu/(1-\mu))$
Nominal and ordinal	Multinomial	probit	$g(\mu)=\phi^{-1}(\mu)$

In Dynamic microsimulation, the logit and probit are the two most useful models when we are estimating the transition probabilities. In this paper, Section 3.4 *Employment status*, the logit model is used for the outcome *Y* is the binary vector with value 0 and 1(Status of Employed and unemployed). The probit could be applied the outcome *Y* is a vector with binary values, and also, the outcome Y is restricted to ordinal values, denoted by *1, ….., k, k+1*. The examples can be seen in SESIM model when estimating the probability of income of capital and probability of capital loss, the unemployment and male participation rate in “Accounting for Unobserved Heterogeneity in discrete-time, discrete-choice Dynamic Microsiluation models”.

Mixed model: The applications of GLM(Generalized Linear Model), Linear Regression mentioned above are under the classical statistics assumption that the observations are independent and identically distributed (iid). However, when we consider the initial database for microsimulation, as can be seen from the Table 1 and Table 2, some countries use registered data, and some use survey data. The sample from the registered database satisfies the classical assumption, that the observations are iid, where the sample from the survey database could not guarantee that the observations are iid, since the initial database may have clustered structure for the reason that the survey data come from different regions and the data in the same region are not iid anymore. Due to this reason, mixed model will be the good solution.

In contrast to the GLM, the mixed model treats clustered data adequately and assumes two sources of variation, within cluster and between clusters. It is defined below:

$Y=X\beta+Zu+e$,

where

Y is a vector of observations, with mean $E(Y)=X\beta$

β is a vector of fixed effects

u is a vector of independent and identically-distributed (IID) random effects with mean $E(u)=0$ and variance-covariance matrix $var(u)=G$.

e is a vector of IID random error terms with mean $E(e)=0$ and variance $var(u)=R$

X and Z are matrices of regressors relating the observations Y to β and u.

In dynamic microsimulation, mixed model could also be used in the panel data when countering the estimation of person's salary, as can be seen from chapter 3.4 *Earnings*, the person's salary for each year is independent to other people, but it is correlated to the salaries he/she earned before, mixed model is used. After obtaining the estimated salary, we could continue to the next step-simulating the income tax in the current year. The mixed model is used this way:

$$y_{it} = \alpha + X_{it}\beta + \mu_{it}$$
$$Y = X\beta + Zu + e \quad ,$$

$$\mu_{it} = \mu_i + v_{it} \text{ ,}$$

$$\mu_i \sim N(0, \sigma_\mu^2) \text{ ,}$$

$$v_{it} \sim N(0, \sigma_v^2) \text{ ,}$$

where μ_i are individual-specific, time-invariant effects, and v_{it} is a time-dependent random effect. The parameters of a linear mixed model are usually estimated by ML (maximum likelihood) or REML (Restricted ML) (Demidenko 2004). In SAS, the function "proc mixed" gives us the mixed model.

Generalized linear mixed model (GLMM): Generalized linear mixed model is a special case in mixed model, because it is an extension to the generalized linear model in which the linear predictor contains random effects in addition to the usual fixed effects referring the variation within clusters. These random effects are usually assumed to have a normal distribution or Gaussian distribution. So the GLMM could be obtained by adding random effects on the linear scale. It is defined as this:

$$E(Y) = g^{-1}(X\beta + Zu),$$

where $g(.)$ is a link function, Z is a design matrix for the random effects. The random effects are assumed to be normally distributed with mean 0 and variance matrix G.

Here is the logistic regression with random effects:

$$\log(p/(1-p)) = X_i\beta + Z_iu\ .$$

Model parameters are usually estimated by ML (maximum likelihood) (Demidenko 2004).

3.3.2 Model structure

In dynamic microsimulation procedure, the underlying data base is aged by one year, and it is run repeatedly to generate the multi-year demographic evolution needed for the whole simulation.

Its “kernel” ages an input database by one year in any given pass. During each such pass, it simulates all of the births, deaths, marriages, labour force entry and exit and earnings, etc., that occur during that simulation year, and ages each of the individuals in the database by one year. It then outputs another database that is itself a new, representative population, but one that reflects the situation one year later than did the previous input database. The cycle is repeated over and over again for the length of the simulation run; in each cycle, the output data base from one pass through the kernel is used as the input for the next pass. (Anthony et al. 1999)

The collective output from the DYNACAN-B (Anthony et al. 1999) stage is thus a sequence of cross-sectional databases, one per simulated year. The key characteristic for the outputs’ collective suitability for subsequent longitudinal analyses is that each individual has a unique personal identifier; that identifier is included in each of the sequence of year-specific data bases generated as the simulated individuals age between one database to the next.

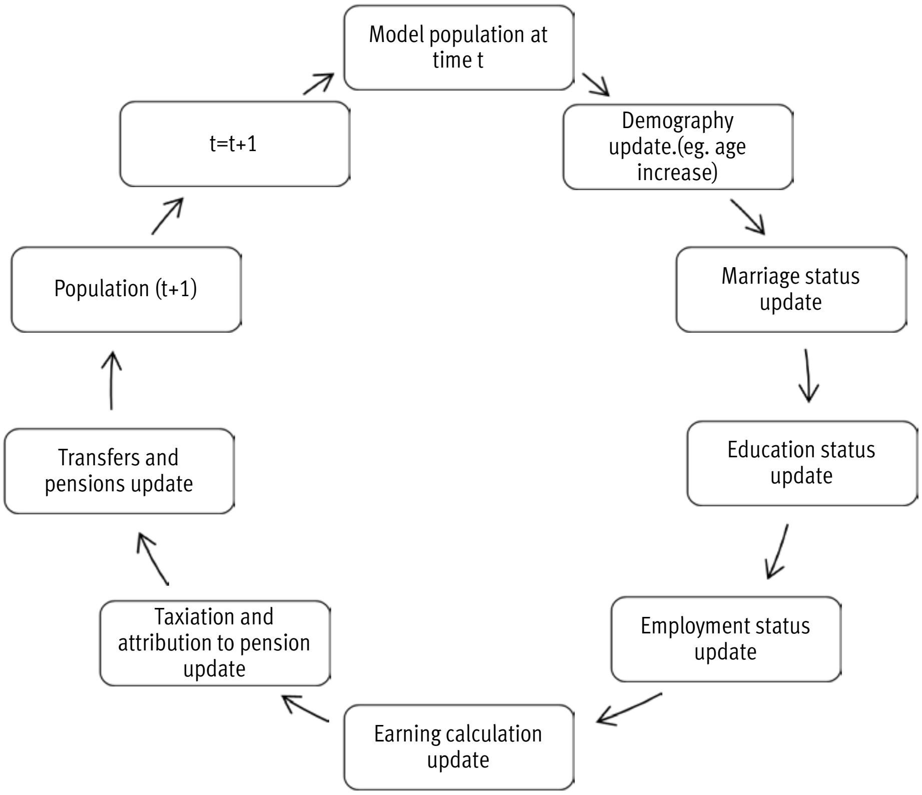
We could consider the aging process as a sequence of modules that this step consists of a number of modules (events) executed in sequence, each of them modifying the in-memory population for that module's event for the current year. Each module processes all of the population for which that module/event is relevant, updating that aspect of the individuals' lives. However, not all individuals are eligible for all modules; e.g. individuals who have previously died will not give birth, and individuals who are presently married are not, in the same year, eligible to enter the marriage market.

Once the full set of modules has been executed, they have collectively aged the in memory population by one year. That is, they have implemented all of the events that effectively transform the base from one year's representative population to the next year's representative population.

3.3.3 The design for the sequence of the event

To describe the sequence of events, I have generated-Life Cycle Graph (Figure 3).

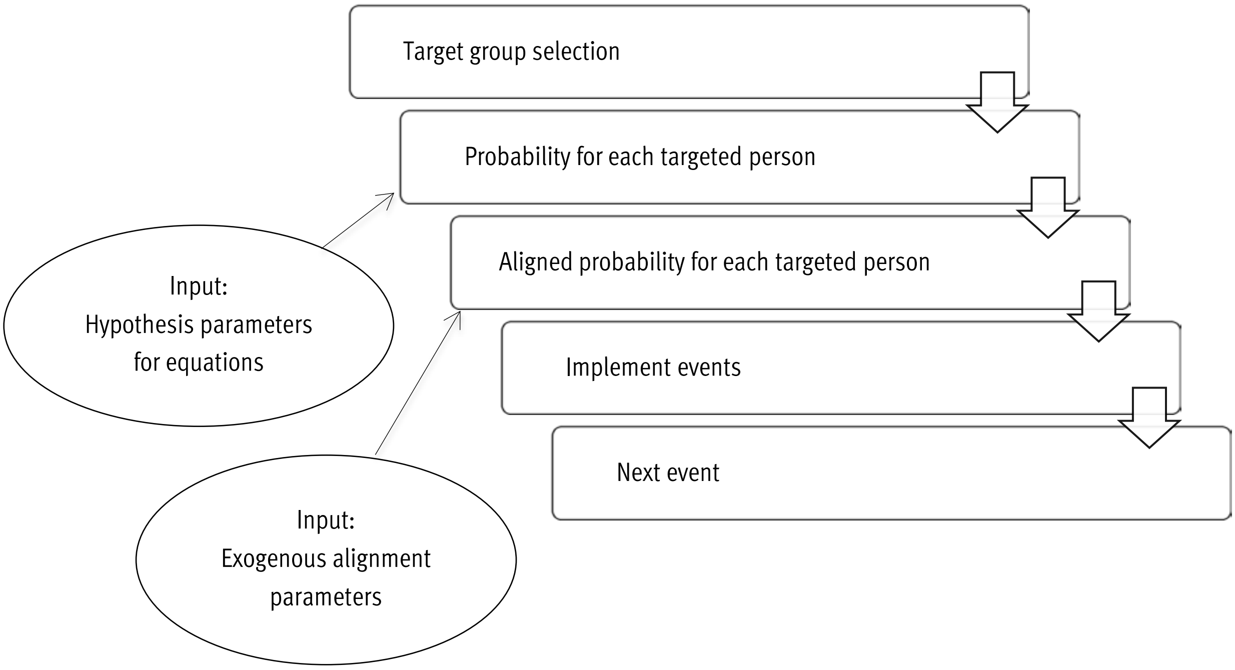
Figure 3. Life cycle process.


Model population at time t
Demography update.(eg. age increase)
Marriage status update
Education status update
Employment status update
Earning calculation update
Taxiation and attribution to pension update
Transfers and pensions update
Population (t+1)
t=t+1

In each event, the transition probability plays this way (Figure 4):

- Select target group through the certain rules for certain events. For instance, females aged from 18 to 40 will give birth, only disabilities would get the disability pension, only the 65 or above people would get retirements.
- Uses equations to calculate probabilities for each individual or family. (How to find a good equation, regression model, logit model?) and sum them up to the aggregation level, say expected total events.
- Calculate alignment factors. Using exogenous data (aggregate level) to derive an optimal adjustment factor to obtain the more practical number.
- Implement events using aligned probabilities.

Figure 4. Procedure in each event.



3.4 Modules

Modules in the life cycle include the people's basic demography and their life economics, they are considered as the life events in the dynamic microsimulation model. The basic demography contains the mortality, fertility, age growing, marriage, divorce and so on. The economics means the individual's employment status, salary income and so on. Here we will introduce some typical modules.

Mortality: Each simulation year, this module decides whether an individual living at the start of the year will be selected to die during the year. The probabilities of death depend on calendar year, age, gender, marital status, income, disability status, and region.

Fertility: It would be a suitable case applying the survival function. The method used to model childbirth involves simulating three distinct types of birth – premarital, first marital, and second and subsequent marital births. The premarital births refer to the births among women who have never been married. First marital birth refers to the first live birth after entry into the first marital union, while second and subsequent marital births refer to all births after the first.

Survival functions are used to predict the time until occurrence of pregnancy among women between 15 and 50. The estimation of the time until childbirth for the three types of childbirth depends on selected characteristics of the women, particularly educational participation, educational qualifications, employment status and age. The first and subsequent marital births are influenced by additional characteristics including marital status, employment status of husband, duration of the marital relationship and time since last childbirth. Thus, three piecewise exponential hazard regression models could be estimated from official survey database.

Births are modeled on the estimation of continuous-time functions of the following type (Cox model):

$$h_{ki}(t) = h_{k0} e^{\beta_k(t) X_{ki}(t)}$$

Where $h_{ki}(t)$ is the probability that individual i experiences event k at time *t* conditional on the explanatory variables in $X_{ki}(t)$ and exposure of individual i to event k at time *t*. The baseline (corner) hazard h_{k0} is constant and coefficient vector of explanatory variables $\beta_k(t)$ is time invariant. (Anthony et al. 1999)

As can be seen from the model, it is a typical Cox model. There are some similarities between the regression model and Cox model, the only difference is that the dependent variable Y is the hazard function in Cox model at a given time. The hazard function-denoted by *h(t)* is presented as:

$$h(t) = \lim_{h \to 0} \frac{P(T \in [t, t+h] \mid T \geq t)}{h},$$

That is the rate of change of the conditional failure probability, and it could be estimated by maximum likelihood resulting:

$$\hat{h}(t) = \frac{total\ number\ of\ failures}{total\ person-time},$$

The total person-time is summing up all persons' risk time in a certain cohort, and total number of failures it the number of failures observed in the risk cohort.

Therefore, the basic Cox model form is given by:

$$h(t) = h_0(t)\exp(\beta_t X_t),$$

The term $h_0(t)$ is the baseline hazard function and it represents the probability of failure when all the explanatory variables are zero. The coefficient vector β gives the proportional change in the hazard, corresponding to the change in the explanatory variables, and it could be estimated by the maximum likelihood method by using computer software such as SAS proc phreg. There is important property of vector β that a positive sign means that the hazard is higher (probability of failure is higher), vice versa.

This Cox model used in dynamic microsimulation helps estimating the probability of person *k* giving birth at time *t,* latter the new born persons will be inserted to the original database as the extra synthesis people.

Aging: Each simulation year, each individual who has not died or just been born, becomes a year older.

Young people leaving home: Leaving the parental family to form one's own family is a big step in one's life cycle. Generally, single person aged 18 years or more living with their parents is assessed using probability distribution, which is the transition probability in this model. The algorithm is presented below:

The normal regression model is applied:

$$\pi_i = \alpha + X\beta + \varepsilon_i$$

Where π_i is the transition probability for individual i, α is the intercept, X is the explanatory variables vector, including age, sex, disability, parental occupation, parental education, education obtainment, presence of siblings and parental income, β is the coefficient vector of the explanatory variables.

Thus, the estimated transition probability is calculated:

$$\hat{\pi}_i = \hat{\alpha} + X\hat{\beta}$$

Where $\hat{\alpha}$ and $\hat{\beta}$ are estimated using outsource database.

Then, a random number is chosen from the uniform, distribution, $u_i \sim U(0,1)$. Finally, if $\hat{\pi}_i$ is larger than u_i, we say that this person will leave home, otherwise, will not.

Education: Each year, individuals currently in school may complete another year of school or leave school, potentially to join the labor force. This module decides, for each individual already in school at the beginning of the year whether to increment that person's years of schooling by one additional year during the simulation year. It assumes a specific, fixed relationship between the years of education completed and educational attainment (high school diploma is received after 12 school years).

Employment status: It would be a good case to illustrate the Monte Carlo Simulation (Lennart Flood et al. 2005) Monte Carlo technic gives the model stochastic property. For the binary variable employment status, we 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 denote employment, 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 u_i. 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 these attributes also determine the probability of unemployment. This is typically accomplished by a logit regression model. 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 like gender, age, working history or any other characteristic relevant for explaining unemployment, i.e. rate of regional unemployment and β is a vector of parameters.

In order to calculate the estimator $\hat{\pi}_i$, firstly, we need to know the expected parameter vector $\hat{\beta}$, where it could be estimated from the outsource databases, such as registered database and official survey database. Then, the dependent variable $\hat{\pi}_i$ is calculated this way:

$$\hat{\pi}_i = [1 + \exp(-X_i\hat{\beta})]^{-1}$$

After this, u_i is chosen randomly from the uniform distribution: $u_i \sim U(0,1)$. Finally simulated binary variable employment status is assigned to 1 or 0 by comparing $\hat{\pi}_i$ and u_i.

Earnings: Earning simulation also applies for Monte Carlo method. The model is given as
$Y_{ij} = X_{ij}\beta + \gamma_i + \varepsilon_{ij}$, 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 correlation, for each individual, over time is $\rho = \tau^2 / (\tau^2 + \sigma^2)$.

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.

The simulations of the earnings equation is based on the individual attributes in X_{ij}, the estimated parameters $\hat{\beta}$ and the random numbers $\tilde{\gamma}_i$ and $\tilde{\varepsilon}_{ij}$. 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 $\tilde{Y}_{ij} = X_{ij}\hat{\beta} + \tilde{\gamma}_i + \tilde{\varepsilon}_{ij}$.

Since $\tilde{\gamma}_i$ is specific for each individual and constant over time, only one draw at the start of the simulation is need, but draws for $\tilde{\varepsilon}_{ij}$ have to be repeated for each year (and new individual). (Lennart Flood et al. 2005.)

Taxation and attribution to pension are simulated by the variables employment status and earning simulation according to different government legislations.

Transfers and pensions: They are relevant with the retirement status and attribution to the pension, mainly calculated according to the legislations.

Aggregation level: After the micro-level simulation, some variables such as tax, pension, and income could be summed up to the aggregation level, this may help the government make wiser legislation rules.

4 Special case study: Norway

4.1 Introduction to effects of family policy reforms in Norway

Main idea: Ageing becomes a problem in most of the developed countries like Norway. In near future, we will facing the condition that there is a decreasing share of the population available for work, which means the people who are capable to work have to work more. Mothers of preschool children could be considered as a contributor to the labor supply. In this paper, the four reformed legislations are measured to increase the participation and the working hours. (Kornstad and Thoresen 2006)

Why mothers of preschool children: In paper by Tom Kornstad and Thor O.Thoresen's paper, the figure showing that the differences between male and female parents of preschoolers are calculated. Nearly 90 percent of the males work full-time, while only about 40 percent of the females are full-time workers. If the preschoolers' mother could work as much as the fathers do, then there are 50000 man-years working time will be increased in the labor supply, this is a substantial input to the Norwegian labor market.

Background for female's choice: In Norway, the preschoolers' mothers have different choices that are: 1. sending child to the child-care center; 2. taking care of child at home, 3. buying other child-care services (e.g. babysitter); 4. combining option 1 and option 2. The people who take care of children at home will get the tax free transfer (allowance) from the government in cash.

4.2 Modeling framework

The discrete choice model is the main technique applied to the Norway case. So the general idea of this model will be introduced.

Discrete choice: Discrete choice problems involve choices between two or more discrete alternatives. The discrete choice models are statistical procedures that model choices made by people among a finite set of alternatives. The models take many forms, including: Binary Logit, Binary Probit, Multinomail logit, Conditional logit, Multinomial Probit, Nested Logit and Mixed Logit, they were introduced by Daniel McFadden in 1978. Nowadays, the models could be applied through some statistical software, such as Stata and SAS (MDC).

Defining Choice Probabilities: A discrete choice model specifies the probability that a person chooses a particular alternative, with the probability expressed as a function of observed variables that relate to the alternatives and the person. In its general form, the probability that person *i* chooses alternative *k* is expressed as:

$$p_{ik} \equiv \Pr\ (\textit{person i chooses alternative k}) = G(x_{ik,} x_{im,} \forall k \neq m, s_i, \beta)\ ,$$

where

x_{ik} is a vector of attributes of alternative *k* faced by person *i*,

x_{im} is a vector of attributes of the other alternatives (besides *k*) faced by person *i*,

s_i is a vector of person i's characteristics and β is a set of parameters of the vector x_{ik}, which will be estimated later.

Consumer Utility: Discrete choice models can be derived from utility theory. This derivation is useful for three reasons:

- The choice probability p_{ik} gets the meaning from the utility function.
- It motivates and distinguishes alternative model specifications.
- It gives the meaning when changing the consumer's surplus from choose the alternative choices.

U_{Ik} is person *i*'s utility of choosing alternative *k*. The value of utility should be maximized: that is person *i* choosing the alternative that provides the highest utility. The choice of the person is designated by dummy variables, y_{ik} , for each alternative:

$y_{ik} = 1$, if $U_{ik} > U_{im}$, for $\forall k \neq m$;

$y_{ik} = 0$, otherwise.

The choice probability is:

$$P_{ik} = pr(y_{ik} = 1) = pr(U_{ik} > U_{im}, \forall k \neq m)$$
$$= pr(\beta x_{ik} + \varepsilon_{ik} > \beta x_{im} + \varepsilon_{im}, \forall k \neq m)$$
$$= pr(\varepsilon_{im} - \varepsilon_{ik} < \beta x_{ik} - \beta x_{im}, \forall k \neq m)$$

where ε_{ik} and ε_{im} are the random errors, x_{ik} and x_{im} are the vectors of attributes person *i* choose *k* and *m*. (Random Utility Models.)

From Probit to Multinormial logit:

The simplest case which only involves two choices:
We define the dependent variable such that

y_i=*1*, if yes
y_i=*0*, if no

For each individual i. Now if we define the x_i as one explanatory variable vector and e_i as a random, independent error. We could fit the linear model showed below:

$$y_i = \beta_0 + x_i\beta_1 + e_i ,$$

$$\hat{y}_i = \hat{\beta}_0 + x_i\hat{\beta}_1 ,$$

thus the expectation probability of individual *i* choosing yes is:

$$p_{i1} = \Pr[y_i = 1] = \Pr[Yes],$$

And not choosing yes is:

$$p_{i2} = \Pr[y_i = 0] = \Pr[No].$$

We also have:

$p_{i1}+p_{i2}=1$, and

$$E(y_i) = 1 \times p_{i1} + 0 \times p_{i2} = p_{i1} = \beta_0 + x_i\beta_1.$$

The logistic function for this model is:

$$\hat{p}_{i1} = F_L(\hat{\beta}_0 + x_i\hat{\beta}_1) = \exp(\hat{\beta}_0 + x_i\hat{\beta}_1) / (1 + \exp(\hat{\beta}_0 + x_i\hat{\beta}_1)),$$

$$\hat{p}_{i2} = 1/(1 + \exp(\hat{\beta}_0 + x_i\hat{\beta}_1)).$$

The log odds of these two probabilities is:

$$\ln(p_{i1} / p_{i2}) = \beta_0 + x_i\beta_1.$$

If we define:

$u_i = \beta_0 + x_i\beta_1$, then

$p_{i1} = \exp(u_i)/(1+\exp(u_i))$ and

$p_{i2} = 1/(1+\exp(u_i))$.

We can make it more general by using the following notation:

$$p_{ij} = a_{ij} / \sum_m^J a_{im}.$$

In this case, $J = 2$, and $a_{i1}=exp(u_i)$, $a_{i2}= exp(0) =1$.

In most of the cases, the choices are more than two options, that is called multinomial logit model, or conditional logit model, which is given below:

$$P(y_i = j) = p_{ij} = e^{(X'_{ij}\beta_j)} / \sum_{k=0}^{J} e^{(X'_{ik}\beta_k)} \ ,$$

Where x'_{ij} a vector for the independent variables, the choice set is C_i which includes the values from $k=0$ to $k=j$.

Discrete model applied to the Norwegian case: The model used in this paper is based on a discrete choice approach, which is an economic model coming from the utility function. The choice is depending on the jobs and various care alternatives characterized by a number of pecuniary and non-pecuniary attributers. The chosen of the jobs depends not only by the wage rates, working hours, but also by their feeling about the jobs-satisfaction, the jobs location and so on. In the same situation, the child-care options vary not only with the opening hours and fees, but also the facilities and quality of staff. Here, the author assumes that the labor supply and choice of child care are outcomes of discrete choices from finite sets of jobs and child are arrangement, where each job has fixed working hours, a wage rate and a number of non-pecuniary characteristics, and each care alternative has fixed opening hours, a care fee and specific quality attributes.

Generally, the standard framework of the labor supply and child care choice literature is to assume that:

$$U = v(L, C, Q) \ ,$$

This is the family utility function (*U*), and *L, C, Q,* represent leisure, consumption and child care quality respectively.

At the same time, the quality of child care *Q* could be given as:

$$Q = Q(T, F, A, I; X, \eta_2) \ ,$$

Where *T* is maternal child care time; *F* is non-parental child care time; *A* is a vector that contains the attributions to quality of non-parental care, it usually consist of group size, staff/child ratio and provider training; *I* represents the different combinations of child care type and job type; *X* is a vector of observed exogenous variables; and η_2 means unobserved household characteristics. (Zaidi et al. 2009.)

The Norwegian model has borrowed the basic idea of the family utility function without accepting the leisure. In the model, the jobs have been divided into groups according to working hours (*j*), and the child care arrangements are divided into three different modes of care (*m*). The Table 4 shows the total plan:

Table 4. Labor supply stimulating policies by Korstad and Thoresen (2006).

	Weekly working hours / weekly child care hours (j)				
Mode of care (m)	**0**	**1–16**	**17–24**	**25–32**	**32+**
Day care center	$m=1,\ j=1$	$m=1, j=2$	$m=1,\ j=3$	$m=1,\ j=4$	$m=1,\ j=5$
Other paid care	$m=2,\ j=1$	$m=2,\ j=2$	$m=2,\ j=3$	$m=2,\ j=4$	$m=2,\ j=5$
Parental care	$m=3,\ j=1$	-	-	-	-

Now we could apply the conditional logistic model to the Norway case, and it is easy to understand that from the paper by Kornstad and Thoresen (2006), the utility function by chosen time j and caring arrangement is:

$$U_{jm} = \max U(C_{kr}, H_k, k, r) = \log(\sum_{k\in B_j} \sum_{r\in S_{jm}} \exp(v(C_{kr}, H_k))) + \varepsilon_{jm},$$

Where C_{kr} is the disposable income corresponding to job k and child care arrangement r and H_k is the hours of work in job k for the mother. The stochastic error term is i.i.d distributed and B is the finite job choice set, *S* is the child care arrangement choice set. (Kornstad and Thoresen 2007)
By T.Kornstad and T.O.Thoresen's assumption:

$$(1/n_{jm})\sum_{k\in B_j} \sum_{r\in S_{jm}} \exp(v(C_{kr}, H_k)) \approx \exp(v(\tilde{C}_{jm}, \tilde{H}_j)),$$

Where n_{jm} is the number of opportunities in the set $B_j \times S_{jm}$, $\tilde{H}_j$ is the median working time in hours of work group j, and $\tilde{C}_{jm}$ is the consumption corresponding to working time $\tilde{H}_j$.

The choice probability is given by:

$$p_{hjm} = \frac{\exp(v(\tilde{C}_{hjm}, \tilde{H}_j, X_h) + \log n_{jm})}{\exp(v(\tilde{C}_{h13}, \tilde{H}_1, X_h) + \log n_{13}) + \sum_{i=1}^{5} \sum_{l \in \Omega_h} \exp(v(\tilde{C}_{hil}, \tilde{H}_i, X_h) + \log n_{il})},$$

where $\Omega_h = \{1\}$ if household h is constrained in the market for care at centers, and $\Omega_h = \{1, 2\}$ otherwise. So p_{hjm} is the probability that household h chooses a job with hours of work in group j and a childcare arrangement in model m.

The indirect utility function is represented as following:

$$v(\tilde{C}_{hjm}, \tilde{H}_j) \equiv \gamma_0 \frac{\tilde{C}_{hjm}^{\alpha_1} - 1}{\alpha_1} + \frac{(1 - \frac{\tilde{H}_j}{M})^{\alpha_2} - 1}{\alpha_2} X_h \beta,$$

where the M=8760 is the total number of annual hours, and $\gamma_0, \alpha_1, \alpha_2, \beta$ are parameters, in them, α_1, α_2 are less than 1, X_h is the number of children below 19 years old.

4.3 Data

The data comes from the Home Care Allowance Survey 1998. The consumption and hours of work are measured annually, and the consumption is defined as disposable family income. As there might be some price differences between centers run by local governments and private owners, the price measures are weighed averages by market shares.

4.4 Measures of distributional effects

Definition: Compensating variation (CV)

It is a measured utility change, referring to the amount of additional money an agent would need to reach its initial utility after a change in prices/quality. It is used to find the effect of a price change

on an agent's net welfare. The CV is used to measure the distributional effects. The indirect utility function according to the utility function mentioned above could be written as:
$V_{jm} = v_{jm}(P_{jm}, I) + \mu_{jm}$,

where μ_{jm} is an error term that has the same distribution as $\varepsilon^*(C_{kr}, H_k, k, r)$, $P_{jm} = [w, Q_{jm}]$ is a price vector consisting of the wage rate and the fee in child care group *jm*. Then equivalent variation is defined implicitly according to

$$\max_{jm}\left(v_{jm}(P^1_{jm}, I^0) + \mu_{jm}\right) = \max_{jm}\left(v_{jm}(P^0_{jm}, I^0 + EV) + \mu_{jm}\right),$$

where superscript 0 denotes initial attributes and superscript 1 denotes attributes after policy interventions. According to this equation, EV is that value of a tax-free transfer that makes the household as well off under the reference tax and transfer system with prices P^0_{jm} and taxable and exogenous household income I^0 as it is under the new regime with prices P^1_{jm} and the initial income I^0. As usually done in simulation studies, it is assumed that the policy experiment does not influence on the random error terms. (Compensating Variation 2012.)

4.5 Effects of four different reforms

1) Increasing the number of spaces at child care centers, the abolition of queues
In Norway, the governments want to achieve the goal" full-coverage" in the market for non-parental child care. However, even though the number of child care centers has increased a lot in the last decades, there is still an excess demand for center-based care, in the sense that there are waiting lists in many municipalities. In their data from 1998, there are about 16 percent of the parents reported that they had applied for care at centers without success.

Since many parents seem to prefer day care centers for other types of non-parental care, abolition of queues in this market might have significant effects on female labor supply.

2) Rate cuts
In Norway, the discussion about reducing the prices for center-based care has been quite hot. A system of maximum prices similar to the Swedish system has been proposed: The fee for full-time care in day care centers should not exceed 1500 NOK per month for the first child, 1200 NOK for the

second child, whereas the maximum fee is 900 NOK for the third child. Thus, the maximum fee for three children in full-time day care centers is 3600 NOK. However, the fees are income dependent. If the pre-tax family income is less than 330,000 NOK, the fees are proportional to income. In this case the fee for the first child is 5 percent of pre-tax family income, for the second child it is 4 percent and for the third child it is 3 percent of family income. For child number 4 and more there is no payment at all.

3) Rescinding the home care allowance

This reform legislation is reversing the rules made before, that the parents of preschool children receive a tax free transfer in cash. However, this is one advantage that it bears the promise of being labor supply stimulating and costs decreasing at the same time.

4) A modified working families' tax credit system

This simulation is related to the motivation behind various tax system arrangements in other countries. This system aims at improving work incentives and thereby raises the earned income of the poor, by letting tax allowances depend on the connection to work.

The system assumes that the family is entitled to NOK 30,000 per year if the female works at least 17 hours per week (the basic credit). If she works more than 24 hours per week (*m*=3 or *m*=4 in Table 4), the family also qualifies for the 30-hours credit of NOK 20,000 per year.

4.6 Conclusion

The Table 5 shows the effects of these four reforms based on average rate. In a more detailed calculation that the population has been divided into 10 groups, from the poorest 10 percent to richest 10 percent, the data shows that all the reforms are more effective to the poor ones rather than the rich ones. Also, when comparing the effect of these four reforms, the last one "*A modified working families' tax credit system*" shows the strongest positive effect.

Table 5. Effects of four reform.

	Reference system (NOK)	**Income Change (NOK)**	**Change in child care-center /hours**	**Change in labor supply /hours**
QUEUES ABOLITION	223446	810	2.7	1.0
RATE CUTS	223446	7531	2.6	0.9
ALLOWANCE RESCINDING	223446	−3876	3.0	1.6
Modification tax credit system	223446	9972	−0.3	0.9

4.7 Model development

The Norwegian model assumed that fathers play no role in the child caring event, but in reality, males do share child caring responsibility, which may leads the over estimation of the effects in reducing the fee. On the other hand, there are some other similar models, which analyzes the female labor market and child caring, such as Australia model, it overcomes the difficulty of ignoring the fathers' role by setting up the model based on the couples rather than only on female individuals.

5 Special case study: Finland

In this section, the practical work is done by assessing the Finnish static model-JUTTA, also certain statistical models will be discussed to improve the JUTTA model.

5.1 Key terminologies for the variables in Finland

All the definitions are from Statistics Finland.

Absence from working life due to child care: Absence from working life due to child care means generally a longer absence from work caused only by family leave, child home care leave or other child care. Absence is also the time when in addition to child care the persons has worked only occasionally or little (under 5h per week). Meanwhile, if the respondent has been on maternity or paternal leave or taken care of children direct after school/studies before starting paid employment, this time is also counted as absence from working life.

Age-Specific death rate: The age-specific death rate refers to the number of deaths per 1000 of the mean population in the age group in question.

Age-specific fertility rate: The age-specific fertility rate indicates the number of live births per 1000 women of the mean population in the age group in question.

Age-specific marriage rate: The age-specific marriage rate indicates the number of married women per 1000 non-married women of the mean population in the age group in question.

Family: A family consists of a married or cohabiting couple or persons in a registered partnership and their children living together, or either of the parents and his or her children living together; or a married or cohabiting couple and persons in a registered partnership without children. Starting from 1 March 2002, same-sex couples have been able to register their partnerships. Persons living in the household-dwelling unit who are not members of the nuclear family are not included in the family population, even if they are related, unless they form their own family. Brothers and sisters or cousins living together are not a family and do not belong to the family population. The same applies to people who live alone or with a person of the same sex. Families living in residential homes are included in the family population. In contrast, persons who live in institutions are not included. A family can consist of no more than two successive generations. If the household-dwelling unit comprises more than two generations, the family is formed starting from the youngest generation. This means, for example, that a mother-in-law or father-in-law living with their child's family will not be included in the family population unless they live together with their spouse, in which case the old couple forms their own family.

Household: A household is formed of all those persons who live together and have meals together or otherwise use their income together. The concept of household is only used in interview surveys. Excluded from the household population are those living permanently abroad and the institutional population (such as long-term residents of old-age homes, care institutions, prisons or hospitals). The corresponding register-based information is household-dwelling unit. A household-dwelling unit is formed of persons living permanently in the same dwelling or address. More than one household may belong to the same household-dwelling unit.

Unemployed:

Definition 1: In the income distribution statistics persons who have been unemployed for at least six months during the year are classified as unemployed. Months of unemployment are asked from persons in the interview. Interview months are checked and where needed, corrected on the basis of register data (Social Insurance Institution's register data on unemployment allowances and times of receipt, the tax register's unemployment allowances). It is used in “Income distribution statistics”

Definition 2: The unemployed labour force comprises persons aged 15-74 who were unemployed on the last working day of the year. Data on unemployment are obtained from the Ministry of Labour's register on job applicants. It is used in “Employment”, “Transition from school to further education and work”.

Definition 3: A person is unemployed if he/she is without work during the survey week, has actively sought employment in the past four weeks as an employee or self-employed and would be available for work within two weeks. A person who is without work and waiting for an agreed job to start within three months is also classified as unemployed if he/she could start work within two weeks. Persons laid off for the time being who fulfil the above-mentioned criteria are also counted as unemployed. It is used in “Labour force survey”.

Wages and salaries: Wages and salaries include compensations in money of all employees of the enterprise for work done during the month. Wages and salaries comprise all income taxes and social security contributions collected from employees as well as diverse additional work (overtime work, night work) bonuses and holiday bonuses. Wages and salaries exclude incentive stock options, expenditure arising from the performing of the work and employer's social security contributions. It is used in “Wages and salary indices”(Statistics Finland)

5.2 Assessment of uncertainty of the JUTTA model and innovation methods

5.2.1 Description of JUTTA Model

The JUTTA model is a static microsimulation model developed by Social Insurance Institution of Finland, it is also called tax-benefit model. In year 2009 JUTTA has 10989 households and around 30000 individuals sample size. It has ten sub-models and one main model. The sub-models are designed for each branch of legislations and the main model is designed for running all the sub-models and producing the final results of the key data based on household level. The sub-models include: SAIRVAK, TTURVA, KOTIHUKI, OPINTUKI, KANSEL, VERO, LLISA, ELASUMTUKI, ASUMTUKI, TOIMTUKI. They represent sickness insurance benefits, unemployment benefits, child care benefits and day-care fees, study grant, the national pension system, personal taxes, benefits for families with children, pensioner’s housing allowances, general housing allowances, means-tested income support, respectively. For each of these sub-models, parameter system and function system were built. (Honkanen. P, 2009)

5.2.2 Analysis design

After setting up the JUTTA model, an assessment system is necessarily built by evaluating the accuracy of the model in micro-level: personal level and household level, which depends on the unit type in particular sub-model. The practical frame work has been divided into two parts: one is measuring the accuracy of the JUTTA model (including sub-models) and finding out the most inaccurate model, the other one is improving this model by using statistical strategy.

5.2.3 Assessment of JUTTA MODEL

Database preparation: For each model, the raw database is from the JUTTA model korotus. However, the initial korotus is weighted database. In order to measuring the difference easily, the weight is deducted for each variable in each model.

Also, because for most of the sub-models, only except the VERO model, not all benefits are applied to each person, the observations are deleted in case that both real value and estimated value are zeros. Thus, the number of observations for each variable declines. However in VERO model, we assume that all the samples pay the taxes, thus there is no need to exclude the both zeros.

All the sub-models are calculated based on personal levels, only except the LLISA model, although its raw database is also in personal level. The reason is that the real value and the estimated value are not calculated on the same person in the same household in LLISA model, thus they could not be compared directly. So the real values and estimated values are summed up to household level, and then the values are compared.

Measuring the accuracies in two different forms: In all the models, the accuracy is calculated in two different forms, one is the absolute difference percentage and the other one is the relative difference percentage. The algorithm of absolute difference percentage is:

1) Calculating the absolute difference between the observed value in the database and generated value by JUTTA for each individual:
   $d_i=|y_i-\hat{y}_i|$;
   where d_i is the absolute difference, $y_i$represents the observed value and $\hat{y}_i$ represents the estimated value by JUTTA.

2) Classifying the absolute difference into five intervals by giving them numbers from one to five, the interval are [0, 1), [1, 10), [10, 100), [100, 1000) and [1000, ∞).
3) Calculating the number of the observations in each interval, and then divided them by the total number of observations to obtain the percentage.

The algorithm of relative difference percentage is:

1) Calculating the absolute difference absolute difference between the real value and estimated value for each individual.
2) Calculating the relative difference based on the real value, that means multiplying real values by 0.1%, 1% and 10%.
3) Obtaining the four relative difference intervals, which are [0%, 0.1%), [0.1%, 1%), [1%, 10%), [10%, ∞).
4) Classifying the observations into these four different levels by falling the absd values into the four relative difference intervals.
5) Calculating the number of the observations in each interval, and then divided them by the total number of observations to obtain the percentage.

The final results of these two forms of accuracies are carried out Table 6 (p. 39–40).

5.2.4 Conclusion and comments of the assessment

From the results showed in Table 6, it is clear to see that most of the models perform quite well, with their variables' accuracy high enough in the interval (60%, 100%] for both absolute different and relative different in first level called [0, 1) and [0, 0.1%) respectively. However, there is one extremely inaccurate model called TOIMTUKI, which with both zero percentage in the first level intervals and more than 60% in the last intervals([1000, ∞) and [10%, ∞)). In the next section, we will search for the reasons for the inaccuracy of TOIMUKI model and try to improve it with statistical model.

The Toimtuki is the last benefit the people could apply after house benefit, health benefit, student benefit and so on. In the other word, the toimtuki could be regarded as the "residual" benefit in the JUTTA model, where the people apply when no other benefits could be applied. So, obtaining the accurate estimated value from TOIMUKI is based on accurate estimation on all other models, which brings more challenges to TOIMUKI model. Another possible reason could come from the vary

Table 6. Comparison of models for year 2009.

Model	Variable	Number of observation	Absolute error percentage					Relative error percentage			
			[0, 1)	[1, 10)	[10, 100)	[100, 1000)	[1000, ∞)	[0%, 0.1%)	[0.1%, 1%)	[1%, 10%)	[10%, ∞)
ASUMTUKI	ASUMTUKI	367	0.00545	0.01090	0.12534	0.49046	0.36785	0.00817	0.03542	0.24796	0.70845
ELASUMTUKI	SUM	235	0.00851	0.06809	0.19574	0.59149	0.13617	0.01277	0.08936	0.29362	0.60426
KANSEL	KANSEL	2288	0.98820	0	0.00044	0.00524	0.00612	0.98820	0	0.00306	0.00874
	LAPSKOR	84	0.94048	0.05952	0	0	0	0.41667	0.52381	0.05952	0
	RILI	179	1	0	0	0	0	1	0	0	0
	YLIMRILI	92	0.75000	0.04348	0.08696	0.11957	0	0.61957	0.16304	0.07609	0.14130
	EHOITUK	502	0.91235	0.02191	0.05976	0.00598	0	0.90637	0.03386	0.05976	0
	LHOITUKI	257	0.82101	0.04669	0.10895	0.02335	0	0.77821	0.10895	0.08560	0.02724
	VAMMTUKI	66	0.68182	0.03030	0.13636	0.15152	0	0.63636	0.12121	0.21212	0.03030
	KELIAK	155	1	0	0	0	0	1	0	0	0
KOKOSIMUL	EI_MALL_REKONSTR/ KAYT_TULO_MALLI	10989	0.69051	0.21421	0.00328	0.03258	0.05942	0.90518	0.00910	0.04259	0.04313
KOTIHTUKI	KOTIHTUKI	571	0.87916	0	0.01051	0.10683	0.00350	0.87916	0	0.05954	0.06130
	OSHOIT	90	0.62222	0.11111	0.26667	0	0	0.62222	0.03333	0.3000	0.0444
LLISA09	LLISAT	3334	0.89982	0	0.00840	0.06899	0.02280	0.89922	0.00420	0.03029	0.06629
	ELTUKI	158	0.07595	0.24684	0.67722	0	0	0.17089	0.31646	0.50000	0.01266
	AITAVUST	277	1	0	0	0	0	1	0	0	0
OPINTUKI	TUKIKESK	722	0.36842	0.01801	0.20776	0.36288	0.04294	0.36842	0.00277	0.11634	0.51247
	TUKIKOR	872	0.74197	0.00115	0.01835	0.19151	0.04702	0. 74197	0	0.02408	0.23394
	ASUMLISA	964	0.78838	0.00622	0.04876	0.14627	0.01037	0.78216	0.00934	0.05290	0.15560

Table 6 continues.

Model	Variable	Number of Observation	Absolute Error Percentage					Relative Error Percentage			
			[0, 1)	[1, 10)	[10, 100)	[100, 1000)	[1000, ∞)	[0%, 0.1%)	[0.1%, 1%)	[1%, 10%)	[10%, ∞)
OPINTUKI	TUKIKESK	722	0.36842	0.01801	0.20776	0.36288	0.04294	0.36842	0.00277	0.11634	0.51247
	TUKIKOR	872	0.74197	0.00115	0.01835	0.19151	0.04702	0. 74197	0	0.02408	0.23394
	ASUMLISA	964	0.78838	0.00622	0.04876	0.14627	0.01037	0.78216	0.00934	0.05290	0.15560
SAIRAVAK	SAIRPR	803	0.66127	0.33250	0.00623	0	0	0.88543	0.10959	0.00125	0.00374
	VANHPR	753	0.75166	0.14343	0.03851	0.05179	0.01461	0.86985	0.03984	0.06507	0.02523
	SAIRPR_TA	1056	0.84943	0.15057	0	0	0	0.97348	0.02652	0	0
TOIMTUKI	TUKI	1012	0	0.00296	0.02569	0.33103	0.64032	0	0.00494	0.03458	0.96047
TOIMTUKI(N1)	TUKI	621	0.00161	0.01288	0.05314	0.37037	0.56200	0.00322	0.01771	0.08535	0.89372
TOIMTUKI(N2)	TUKI	894	0	0.00224	0.02685	0.42953	0.54139	0	0.00224	0.01454	0.98322
TTURVA	PERUSP	252	0.61905	0.05952	0.22619	0.07540	0.01984	0.57143	0.23810	0.11905	0.07143
	ANSIOSID	1604	0.90087	0.00499	0.00436	0.02182	0.06796	0.90274	0.00312	0.00748	0.08666
	TMTUKI	718	0.38858	0.20474	0.20334	0.14067	0.06267	0.42618	0.30641	0.11838	0.14903
	VUORKORV	106	0.42453	0.25472	0.31132	0	0.00943	0.55660	0.20755	0.22642	0.00943
	KOULTUKI_ANS	153	0.57516	0.16340	0.15033	0.07190	0.03922	0.65359	0.21569	0.07190	0.05882
VERO	VALT_ANS	27009	0.96346	0.01018	0.01507	0.01000	0.00130	0.93384	0.03906	0.01755	0.00955
	POVERO	27009	0.94643	0.01751	0.02595	0.00826	0.00185	0.88130	0.04254	0.03932	0.03684
	SVMAKSU	27009	0.71432	0.16469	0.09726	0.02325	0.00048	0.50294	0.25099	0.14617	0.09989
	KUNNVERO	27009	0.86001	0.04184	0.06794	0.02843	0.00178	0.85623	0.05957	0.06342	0.02077
	KIRKVERO	27009	0.92162	0.04791	0.02310	0.00703	0.00033	0.69177	0.21278	0.06272	0.03273
	PALKVAK	27009	0.99985	0	0.00007	0.00007	0	0.99985	0	0	0.00015
	MAKSVEROT	27009	0.66663	0.12251	0.14014	0.06379	0.00692	0.71472	0.13810	0.11237	0.03480

original information, all the data was collected based on every year, but not every month, the situations could be changed during one year, for instance, the person was not employed last year in May, but he was again employed since July, if the data is collected in June, the his employ status would be no, so he may apply the TOIMTUKI benefits, however, in reality, his status is yes, and he may not apply TOIMTUKI. The other potential reason is that not all the people would like to apply TOIMTUKI in real life even they meet the criterions to apply this benefit, because the action is depending on each person.

5.3 Innovation methods in TOIMTUKI

5.3.1 Method one

The algorithm:

1) Giving the estimated value zero in case the real data is zero.
2) Fit the regression with the rest of the individuals, and estimate the values.
3) Comparing this model with the JUTTA model by assessment the results.

The hypothesis variables that would affect the real value are: htyotper tyot tyotseu martul04 tyotkmuu lpaktyva ttyotpr palkm vvvmk1 vvvpvt1 svatva maksvuok jasenia desmod lapsia. Where the meanings of variables are:

htyotper: Basic unemployment allowance paid by KELA in Euros.
tyot: Number of month of person’s unemployment or forced leaving.
tyotseu: Number of month of person’s unemployment or forced leaving in year 2010.
martul04: An earning related unemployment allowance which is awarded in November.
tyotkmuu: Other unemployment compensation.
ttyotpr: Unemployment allowance.
lpaktyva: Employee compulsory unemployment insurance.
vvvmk1: Paid earnings-related unemployment allowance.
vvvpvt1: Paid earning-related unemployment allowance days in total.
svatva: Personal annual income.
maksvuok: Household rent for last month.
jasenia: Number of people in the household.
lapsia: Number of children in the household.
desmod: Decile, from 0 to 9. According to OECD, the average household consumption, each decile group has 10% people.

However after fitting the regression, the significant effecting variables are: tyot ttyotpr svatva maksvuok jasenia, with their p-values under 0.0001. See Table 7.

Table7. All parameters' estimators.

Variable	Parameter estimate	Standard error	t value	Pr>t
Intercept	186.9843	229.2549	0.82	0.415
htyotper	0.32126	0.39592	0.81	0.4174
tyot	161.746	23.69606	6.83	<.0001
tyotseu	−1.93579	15.80183	−0.12	0.9025
martu104	1828.867	662.6417	2.76	0.006
tyotkmuu	−0.02276	0.05638	−0.4	0.6865
lpaktyva	11.77559	5.94797	1.98	0.0482
ttyotpr	−0.16711	0.04094	−4.08	<.0001
palkm	−12.5026	11.17277	−1.12	0.2636
vvvmk1	0.19488	0.10074	1.93	0.0535
vvvpvt1	−10.5159	4.85797	−2.16	0.0308
svatva	−0.12977	0.01577	−8.23	<.0001
maksvuok	1.90105	0.31597	6.02	<.0001
jasenia	1461.561	190.1344	7.69	<.0001
lapsia	−943.908	203.7964	−4.63	<.0001
desmod	508.47509	74.76715	6.80	<.0001

From the table 7, we could see from the t Value, where Pr > t, if this probability is smaller than 0.0001, we say that variable is statistically highly significant. So, after fitting the model, the variables: tyot, ttyotpr, svatva, maksvuok, jasenia should be kept, hence, Table 8 is carried out.

Table 8. Significant variables' parameters.

Variable	Parameter estimate	Standard error	t value	Pr>|t|
Intercept	195.67209	230.39284	0.85	0.3960
tyot	158.40442	21.28756	7.44	<.0001
ttyotpr	−0.17366	0.03029	−5.73	<.0001
svatva	−0.10823	0.01098	−9.86	<.0001
maksvuok	2.02230	0.31497	6.42	<.0001
jasenia	1308.86787	183.66519	7.13	<.0001
lapsia	−786.69217	198.34392	−3.97	<.0001
desmod	520.13951	74.80554	6.95	<.0001

Now estimates the values by the equation:

$\hat{y}=X\hat{\beta}$,

where X is the characteristics vector and $\hat{\beta}$ is the vector of estimated coefficients listed in Table 7. Now we plug the numeric in this equation:

$\hat{y}$ =195.67209+158.40442×tyot-0.17366×ttyotpr-0.10823×svatva+2.02230×maksvuok +1308.86787×jasenia-786.69217×lapsia+520.13951×desmod; (R-Square=0.3744)

So, from this method we see that the moth of unemployment, unemployment allowance, household monthly income, household rent, number of persons in the household, number of children and household decile number, they play significant roles in estimating the benefit for TOIMTUKI. Table 9 shows us absolute and relative percentages under JUTTA and Linear Regression model (M1), so we see there are clearly improvement when using M1 both for absolute and relative percentages.

Table 9. Comparison of JUTTA model and Linear regression model.

Model	Variable	[0,1)	[1, 10)	[10, 100)	[100, 1000)	[1000, ∞)	[0, 0.1%)	[0.1%, 1%)	[1%, 10%)	[10%, ∞)
JUTTA	TUKI	0	0.00296	0.02569	0.33103	0.64032	0	0.00494	0.03458	0.96047
M1	TUKI	0.00161	0.01288	0.05314	0.37037	0.56200	0.00322	0.01771	0.08535	0.89372

5.3.2 Method two

The algorithm:

1) Estimate the binary variable status, which describes the weather the person gets this benefit or not, meaning if he/she gets, then status=1, if he/she doesn’t get, then status=0. This step is using Monte Carlo method. Firstly, by logistic regression, the estimated parameters are calculated, then by using $\pi_i = \exp(X\beta)/(1+\exp(X\beta))$, where π_i is the probability t of being status=1. Finally, generating random value from the uniform distribution, and compare this value with π_i the probability, if the random value is larger than the probability, giving status value 0, if not, giving value 1.
2) Estimating the TUKI value by regression model in case the status=1, otherwise, give value 0. However, the estimated value could be negative, but in reality, it should be nonnegative value, so change the negative value to 0.

So, by calculating $\pi_i = \exp(X\beta)/(1+\exp(X\beta))$, where β is the vector and its estimation is shown in Table 10. Next, a random number u_i is drawn from the standard uniform distribution, that is $u_i \sim U(0,1)$. Finally, by comparing u_i and π_i, we give the estimated status 1 and 0. In cases that the individual estimated status is 1, regression model is set to calculate the TUKI, while in other cases, TUKI will be given value 0 directly. The regression is shown in Table 11.

Table 10. Estimators of the logistic model.

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	−2.1593	0.1185	332.0326	<.0001
tyot	0.2023	0.0107	354.7804	<.0001
svatva	−0.00005	6.957E-6	59.0412	<.0001
maksvuok	0.00216	0.000184	137.2076	<.0001
vvvmk1	−0.00011	0.000019	35.2189	<.0001
desmod	−0.1751	0.0390	20.1498	<.0001
tyotseu	0.0429	0.0120	12.8695	0.0003
lpaktyva	0.00952	0.00287	10.9831	0.0009

Table 11. Estimators of the linear regression.

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	−152.15070	177.19949	−0.86	0.3909
tyot	202.66061	17.99137	11.26	<.0001
ttyotpr	−0.13827	0.02625	−5.27	<.0001
svatva	−0.07511	0.00867	−8.67	<.0001
maksvuok	1.78485	0.26905	6.63	<.0001
jasenia	394.85153	70.85511	5.57	<.0001
desmod	341.67733	64.15313	5.33	<.0001

Then, plug the estimator to the equation:

$$\hat{y}=-152.15070+202.66061\times tyot-0.07511\times svatva+1.78485\times maksvuok-0.13827\times ttyotpr$$
$$+394.85153\times jasenia+341.67733\times desmod;(R\text{-}Square=0.3930)$$

where $\hat{y}$ is the estimated TUKI. However, in this equation, $\hat{y}$ could be negative value, but in real case, it could not be, so when it is negative, we change it to value 0. Now compare the real TUKI and estimated one.(Tables 12 and 13.)

From the table, we see that the second method is better than the original method in the absolute difference view, however, it is almost the same as the original method in the relative difference point of view.

Table 12. Comparison of TOIMTUKI and Method2.

			Absolute Error Percentage					Relative Error Percentage			
Model	**Variable**	**Number of observation**	**[0, 1)**	**[1, 10)**	**[10, 100)**	**[100, 1000)**	**[1000, ∞)**	**[0%, 0.1%)**	**[0.1%, 1%)**	**[1%, 10%)**	**[10%, ∞)**
TOIMTUKI	TUKI	1012	0	0.00296	0.02569	0.33103	0.64032	0	0.00494	0.03458	0.96047
MEHTOD2	TUKI	894	0	0.00224	0.02685	0.42953	0.54139	0	0.00224	0.01454	0.98322

Table 13. Comparisons of TOIMTUKI, Method1 and Method2.

			Absolute Error Parentage					Relative Error Percentage			
Model	**Variable**	**Number of observation**	**[0, 1)**	**[1, 10)**	**[10, 100)**	**[100, 1000)**	**[1000, ∞)**	**[0%, 0.1%)**	**[0.1%, 1%)**	**[1%, 10%)**	**[10%, ∞)**
TOIMTUKI	TUKI	1012	0	0.00296	0.02569	0.33103	0.64032	0	0.00494	0.03458	0.96047
TOIMTUKI(N1)		621	0.00161	0.01288	0.05314	0.37037	0.56200	0.00322	0.01771	0.08535	0.89372
TOIMTUKI(N2)		894	0	0.00224	0.02685	0.42953	0.54139	0	0.00224	0.01454	0.98322

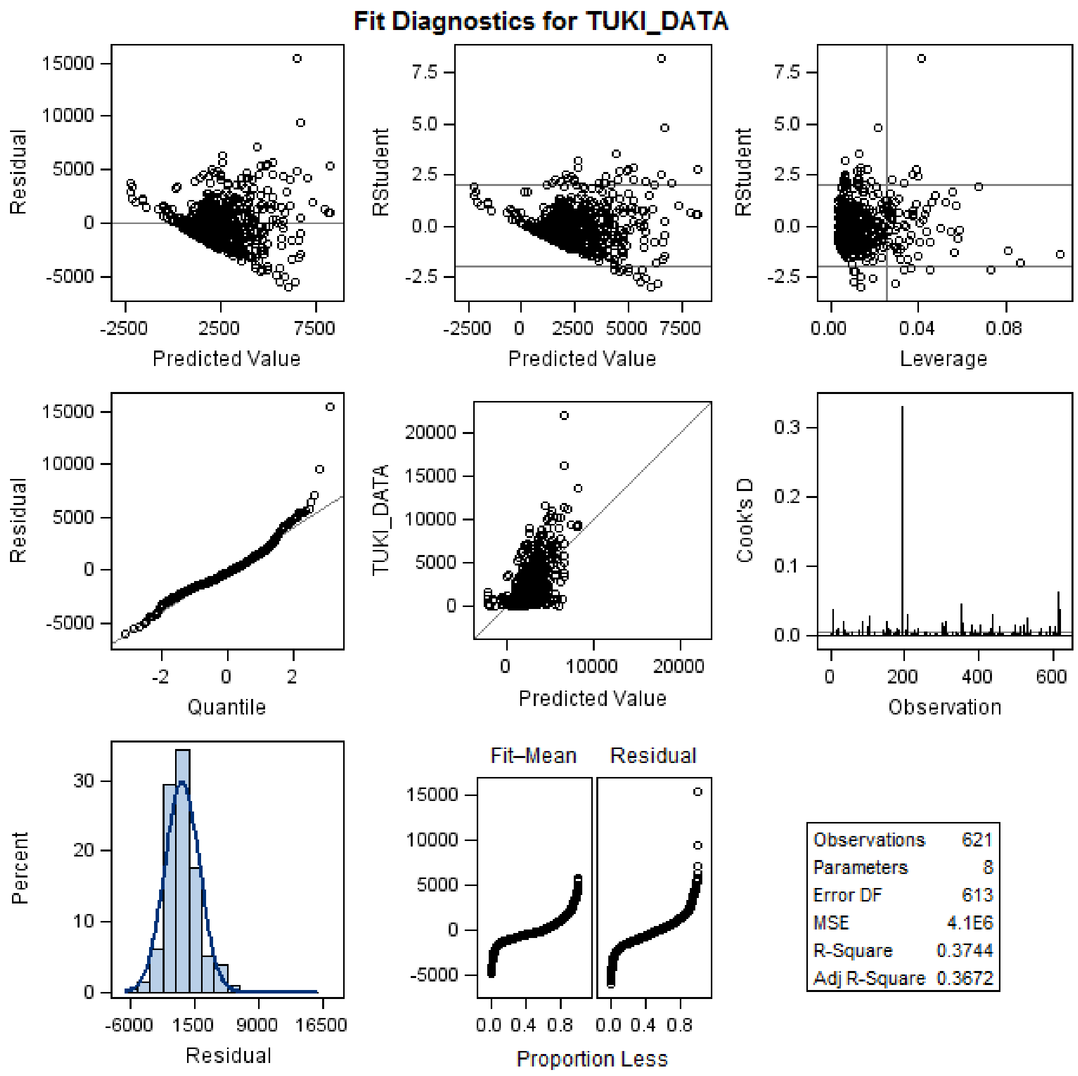
5.3.3 Diagnostics of the two methods

Although the results of effect of these two methods could be seen above, we also should see their effect in a more statistics way. So, the diagnostics of methods are given below.

Regression method

From the graph shown in Figure 5, we see that there are 621 observations which the real value is not zero. The R-Square has value 0.3744, meaning compared with null hypothesis all the parameters are zero, only the intercept is left, the 6 expected value of variables could explain 37.74% more information from the data. R-Square is defined as (Joseph E. Cavanaugh, Criteria for Regression):

Figure 5. Linear regression fitting diagnostics.



Observations	621
Parameters	8
Error DF	613
MSE	4.1E6
R-Square	0.3744
Adj R-Square	0.3672

$R^2=1\text{-}SS_{err}/SS_{tot}=SS_{reg}/SS_{tot};$

where $SS_{tot}=\sum_i(yi-\bar{y})^2$, y_i is observation individual, and $\bar{y}$ is the mean of the observed data, $SS_{reg}=\sum_i(\hat{y_i}-\bar{y})^2$, $\hat{y_i}$ is the fitted value, $SS_{err}=\sum_i(y_i\text{-}\hat{y}_i)^2$

For the adjusted R-Square, it is defined as:

$$R^2_{adj}=1\text{-}\frac{(n\text{-}1)SS_{err}}{(n\text{-}p)SS_{tot}}=1\text{-}\frac{SS_{err}/(N\text{-}P)}{SS_{tot}/(N\text{-}1)}=1\text{-}\frac{MSE}{SS_{tot}/(N\text{-}1)}$$

where p is the number of parameters. So, the maximum of value of adjusted R-Square is equivalent to choosing the fitted model corresponding to the minimum value of MSE. The MSE is donated as:

$MSE= SS_{err}/(n\text{-}p).$

For the graph residual-Predicted Value, the data is averagely spread above and below zero. And for the RStudent-Predicted Value graph, only a few observations are above 2 and below than −2, which means only this few observations could influence the estimated parameters. The Cook's D graph also explains this influence, and from it, the observation number 193 is the outlier, compared with other observations it influence the estimates mostly. The more detailed statistics could be found by using the option influence in SAS code:

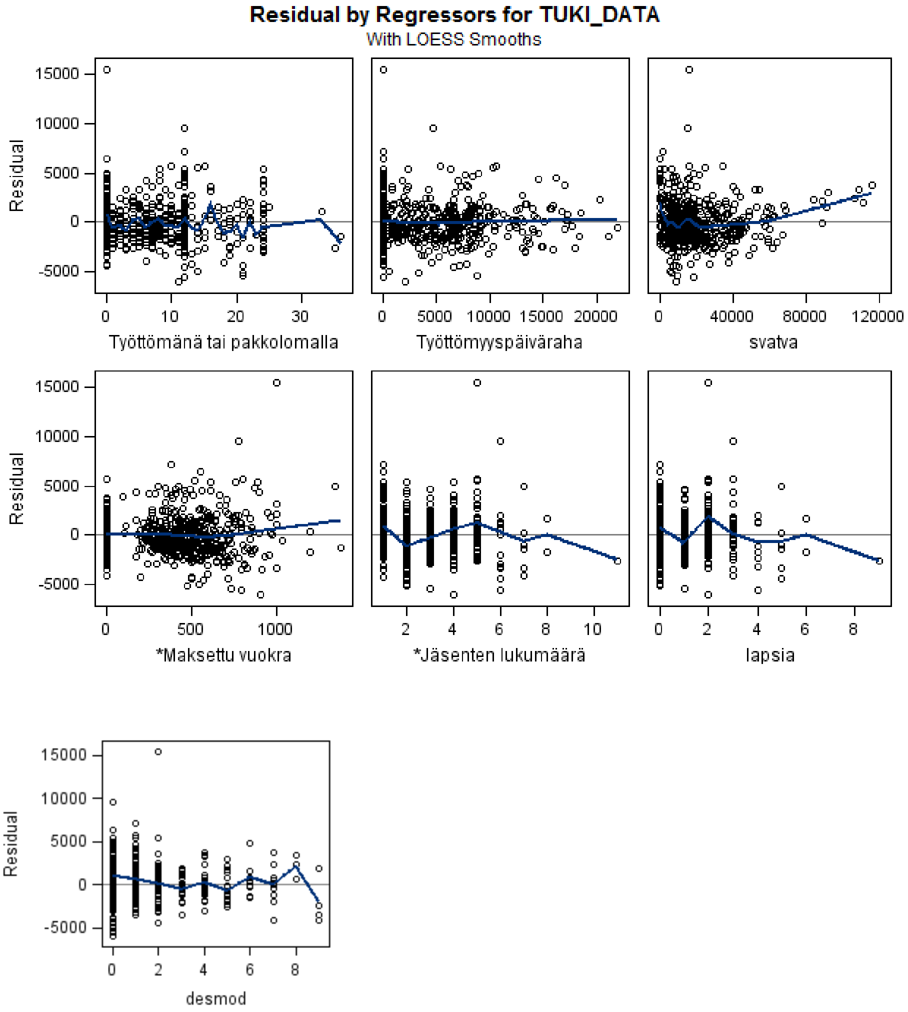
```
model tuki_data=tyot ttyotpr svatva maksvuok jasenia / influence;
```

There, every observation's influence figure is calculated. And we should pay attention to the observations whose absolute values are larger than 2.

From the Residual-Quintile plot, we see that the residuals fall approximately on a diagonal straight line which means the residuals are normally distributed (T.Krishnan, Regression Diagnostics). This point could be confirmed when we check the Residual percent, it is almost normal distributed as we want, only except it is a little bit right skewed.

Next, Figure 6 shows the smoothed residuals against repressors.

Figure 6. Diagnostics for residuals.



After looking at the outlier observation 193, this household has 15658 euro income for one year with 5 persons, and with quite general monthly rent 1005 euros. From this point of view, we could not treat it as the outlier, and I will keep it in the original database.

2 Stage Least Squares Method

1. Logistic regression

From table 14, AIC, SC and -2logL, we see that this model has been improved from the hypothesis that there only exists the intercept, see statistics have been fallen from more than 4500 to less than 3000.

Table 14. Model fit statistics.

Criterion	Intercept Only	Intercept and Covariates
AIC	4776.888	2842.531
SC	4784.193	2900.968
-2LogL	4774.888	2826.531

R-Square 0.1625 Max-rescaled R-Square 0.4610

From the Influence diagnostics in Figure 7: From the Person Residual fitting, we could identify that a few cases when event=1 (obtaining benefits) do not fit the model, since some red crosses do not lie down to the horizontal line=0.

The CI Displacement CBar tells us that the certain outliers undue effect the expected parameter, in this case, the condition is quite good since there are not many outliers. About the DfBetas, for each parameter estimate, the procedure calculates a DfBetas diagnostic for each observation. The DfBetas diagnostic for an observation is the standardized difference in the parameter estimate due to deleting the observation, and it can be used to assess the effect of an individual observation on each estimated parameter of the fitted model. From the Figure 8, we could see that the variable svatva and vvvmk1 have relatively more outliers than the other three variables, so the estimator for these two variables in this model are more effected by the observations.

Figure 7. Influence diagnostics 1.

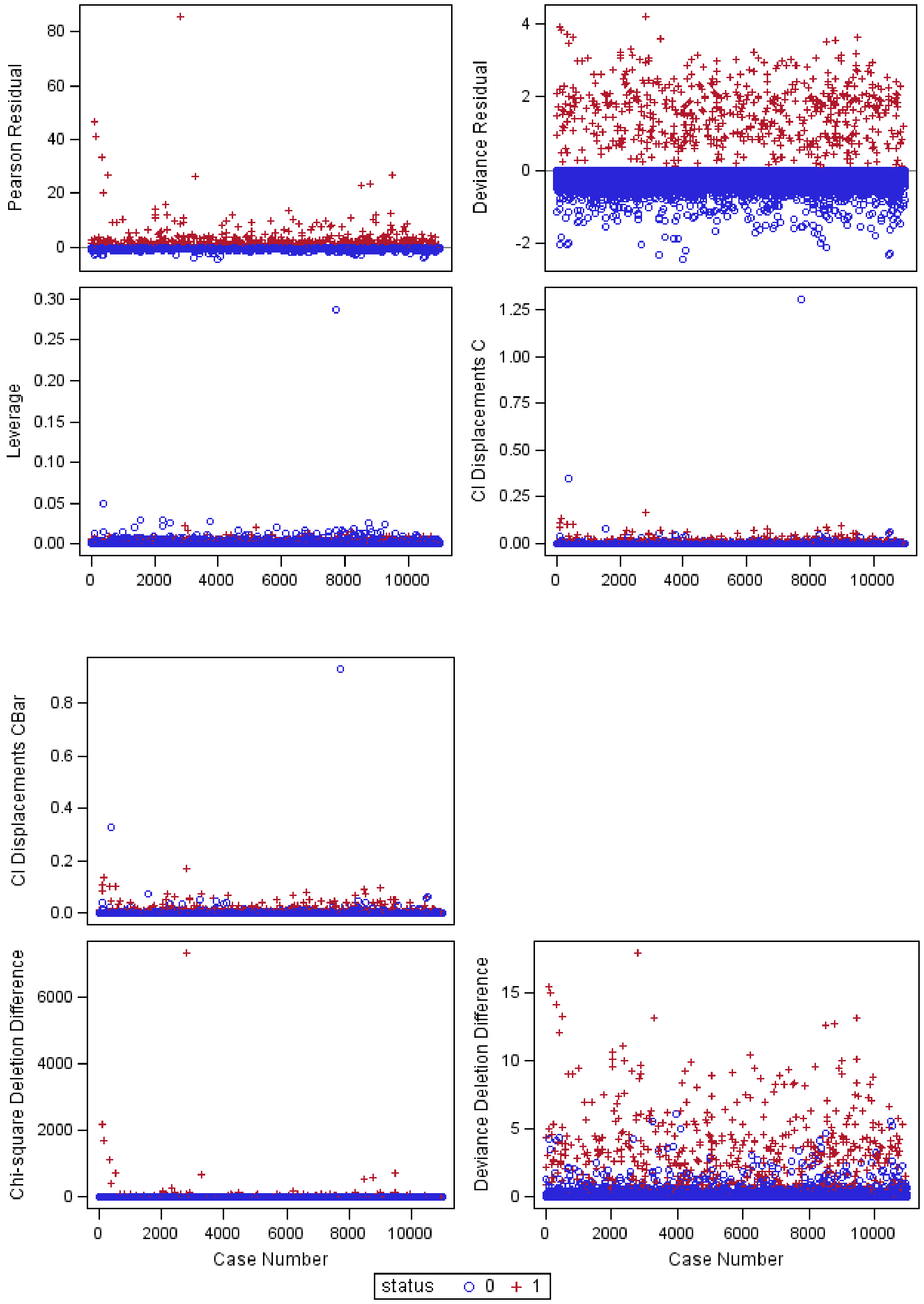
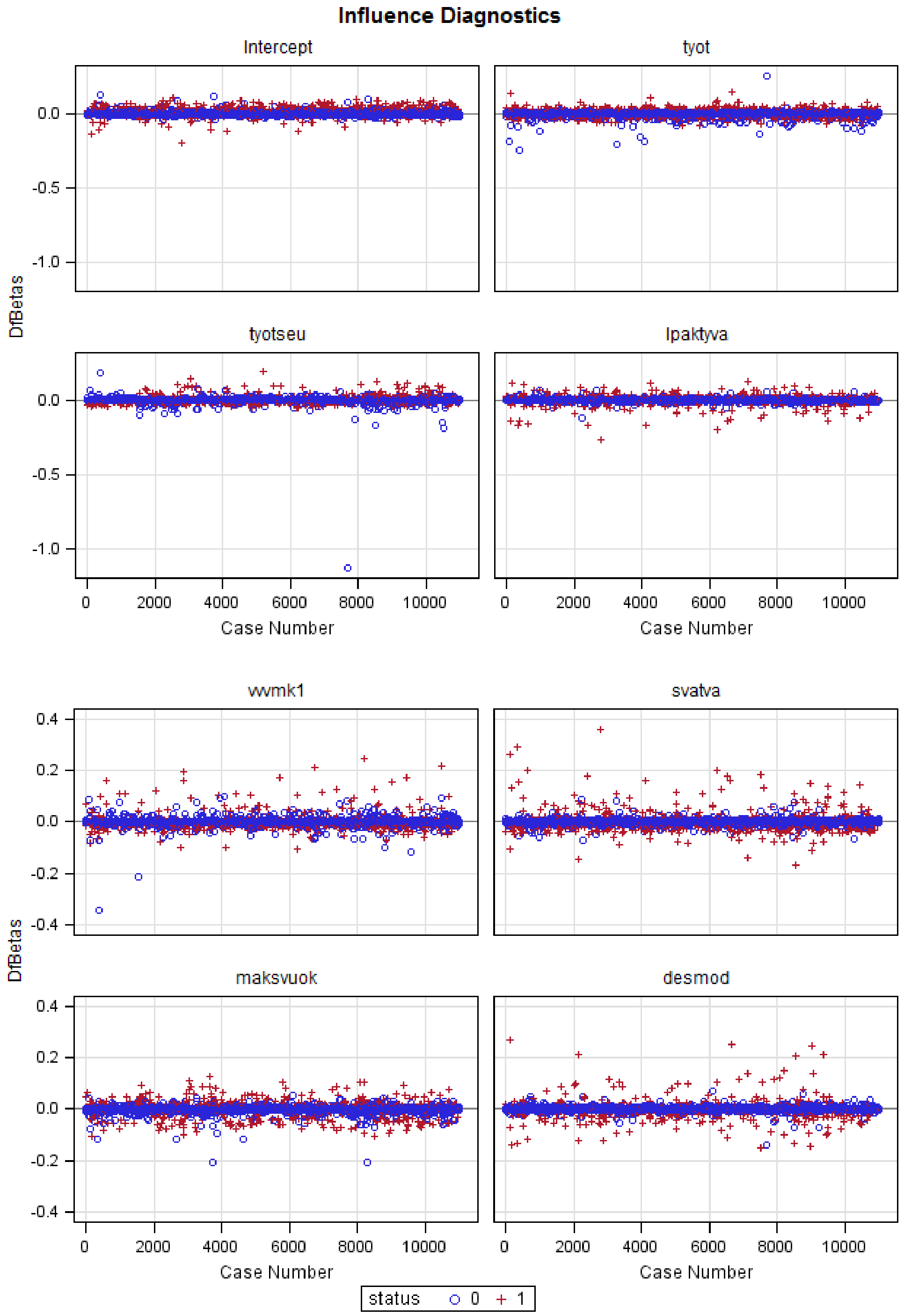


Figure 8. Influence diagnostics 2.



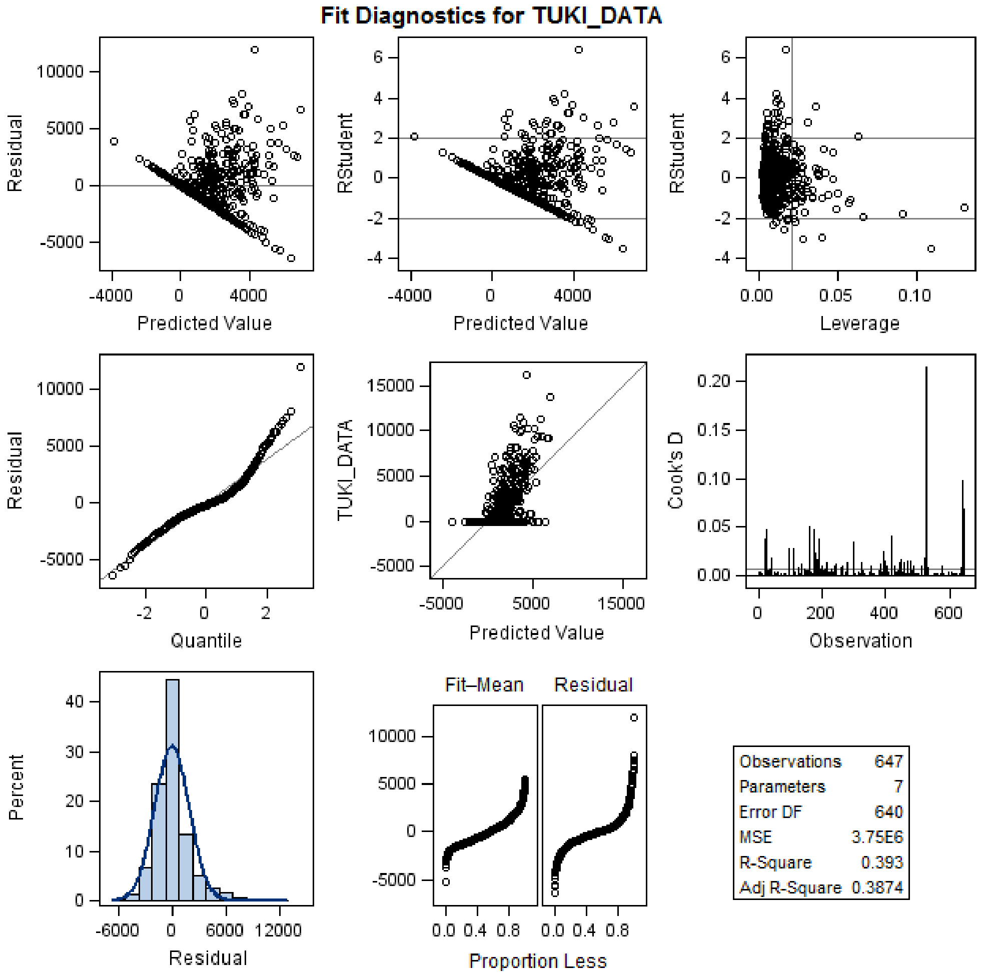
As can be seen from the Figure 7, there are one red cross and one blue cycle outliers, which are observations 2811 and 7702 respectively. When checking them from the database, these two observations do not appear any extreme characteristics, so they are remained in the database.

2. Regression model

When we look at the Figure 9, its R-Square value is larger than the previous simple linear regression model, so it is more presented the information from the data. The other figures show us the same characteristics as the previous model, but Residual has the certain pattern as there is a line, which should be paid attention on.

When we only consider the linear regression of these two models, we will choose the second one, since its MSE is smaller, and both R-Square and Adj R-Square are larger than the first method.

Figure 9. Fitting diagnostics.



Observations	647
Parameters	7
Error DF	640
MSE	3.75E6
R-Square	0.393
Adj R-Square	0.3874

6 Conclusion

The study above firstly went through the static and dynamic microsimulation and focused on the statistical methods applied in them by giving detailed examples of mixed model in the salary estimation, logistic regression model in employment status estimation.

Secondly, a special case study-Norwegian labor force and child care is followed showing a detailed dynamic microsimulation application. In this study case, the effects of four reform options are calculated.

Finally, a Finnish static model-JUTTA has been assessed, the results demonstrate that JUTTA microsimulation model performs quite well in all sub-models, only except for the "residual" model-Toimtuki (income-related supplementary benefit). For Toimtuki, two statistical methods: Linear Regression model and 2SLS model were applied, and they both improved the accuracy the model to some extent, especially in the absolute difference percentage point of view. Also, there might be more potential significant variables need to be found in future that could help TOIMTUKI to be more accurate.

References

Athey S, Imbens GW. Discrete choice models with multiple unobserved choice characteristics. Berkeley, CA: University of California, Economics to Econometrics Conference, 2007.

Balls D, Rossiter D, Bethan T, Graham C, Dorling D. Geography matters: Simulation the local impacts of national social policies. University of Leeds: Joseph Rowntree Foundation p.491, 2005.

Cavanaugh EJ. Model Selection-Lecture VIII: Criteria for Regression Model Selection. Department of Biostatistics, Department of Statistics and Actuarial Science, the University Of Lowa, 2009.

Compensating Variation. Available at: <http://en.wikipedia.org/wiki/Compensating_variation>. Accessed 2012.

Eugene D. Mixed Models Theory and Applications. : Wiley, 2004.

Figari F, Iacovou M, Skew A, Sutherland H. Approximation to the truth: comparing survey and microsimulation approaches to measuring income for social indicators. Institute For Social and Economic Research, UK, 2010.

Flood L, Jansson F, Pettersson T, Pettersson T, Sundberg O, Westerberg A. SESIM III- a Swedish dynamic micro simulation model. Handbook of SESIM051222. : Swedish Ministry of Finance, 2005.

Gribble S, Hicks C, Rowe G. The LifePaths Microsimulation Model. Canberra, International Microsimulation Conference on Population, 2003.

Honkanen P. JUTTA-käsikirja. Tulonsiirtojen ja verotuksen mikrosimulointijärjestelmä. Helsinki: Finnish Social Insurance Institution, 2010.

Keegan M, Kelly S. APPSIM-Dynamic microsimulation modelling of social security and taxation. NATSEM, University of Canberra, 2009.

King A, Bækgaard H, Robinson M. DYNAMOD-2. An overview. NATSEM, University of Canberra, Technical Paper no. 19, 1999.

Kornstad T, Thoresen OT. A discrete choice model for labor supply and childcare. Journal of Population Economics 2007 (20): 781.

Kornstad T, Thoresen OT. Effects of family policy reforms in Norway. Results from a joint labour supply and childcare choice microsimulation analysis. Oslo: Statistics Norway, 2006.

Krishanan T. Regression diagnostics. Bangalore: Cranes Software International Limited.

Lehtonen R, Pahkinen E. Practical methods for design and analysis of complex surveys. Chichester: Wiley, Statistics in practice, 2004.

Lehtonen, R, Veijanen A. Design-based methods of estimation for domains and small areas. In: Pfeffermann D, Rao CR, eds. Hanbook of statistics. Sample surveys, inference and analysis. Volume 29B. Amsterdam: Elsevier, 2009: 219–249.

McFadden D. Computing willingness-to-pay in random utility models. Berkeley, CA: University of California, Econometrics Laboratory Software Archive, 1995.

Morrison R, Dussault B. Overview of DYNACAN. A full-fledged Canadian actuarial stochastic model designed for the fiscal and policy analysis of social security schemes. Statistics Canada, 1998.

O'Donoghue C. Dynamic microsimulation. A methodological survey. Brazilian Electronic Journal of Economics, 2001.

Rahman A. Small area estimation through spatial microsimulation models. Some methodological issues. Ottawa, Second International Microsimulation Association Conference, 2009.

Rahman A, Harding A, Tanton R, Liu S. Methodological issues in spatial microsimulation modelling for small area estimation. International Journal of Microsimulation 2010; 3 (2) 3–22.

Random utility models. Mathematical marketing. Available at: <http://www.docstoc.com/docs/51948559/Chapter-13-Random-Utility-Models>. Accessed 2012.

The urban institute of US, 2012b. Available at: <http://trim3.urban.org/documentation/Static%20versus%20Dynamic%20Microsimulation.html>. Accessed 2012.

Todd P. Classical discrete choice theory. Philadelphia, PA: University of Pennsylvania, 2007.

TRIM3. The Urban Institute of US, 2012a. Available at: <http://trim3.urban.org/T3IntroMicrosimulation.php>. Accessed 2012.

Zaidi A, Harding A, Williamson P. New frontiers in microsimulation modelling. Farnham: Ashgate, 2007.

Appendix: variables in different models

Model

Asumtuki:

Asumtuki: General Housing Allowance. It is intended for low-income households, and it is available for both rental and owner-occupied homes.

ELASUMTUKI:

SUM: Housing benefits for pensioners.

KANSEL:

KANSEL: National pensions.

LAPSKOR: Child increase.

RILI: Front-veteran’s supplements.

YLIMRILI: Additional-veteran’s supplements.

EHOITUK: Pensioners medical aid.

LHOITUKI: Child care payments.

KELIAK: Celiac disease dietary allowances.

KOTIHTUKI:

KOTIHTUKI: Child home care allowance.

OSHOIT: Partial home care allowance.

LLISA:

LLISAT: Child allowances.

ELTUKI: Maintenance support.

AITAVUST: Maternity grant.

OPINTUKI:

TUKIKESK: Secondary school students study grant.

TUKIKOR: University students study grant.

ASUMLISA: Housing supplement for students.

SAIRAVAK:
SAIRPR: Sick leave allowance.
VANHPR: Parental money payments.
SAIRPR_TA: Sickness benefit for employers.

TOIMTUKI:
TUKI: Support benefits.

TTURVA:
PERUSP: Basic unemployment allowance.
ANSIOSID: Earning related allowance.
TMTUKI: Labor market subsidy.
VUORKORV: Job alternation compensation.
KOULTUKI_ANS: Financial assistance for unemployed persons undergoing training.

VERO:
VALT_ANS: State income tax.
POVERO: Income tax.
VARALV: The property tax.
SVMAKSU: Health insurance premium.
KUNNVERO: Municipal tax.
KIRKVERO: Church tax.
PALKVAK: Employee' insurance premium.
MAKSVEROT: MAKSUUN PANNUT VEROT?

KOKOSIMUL:
KAYT_TULO: Disposable income.
KAYT_RAHATULO: Disposable money income.
TULONS_VERONAL: Taxable income transfers.
KANSPERHEL: National pensions and survivors' pensions.
ASUMTUKI: Housing allowances.
LAPSIP: Child allowances, maternity allowances and maintenance support.
TOIMTUKI: Income support.