NSM 2022
NORDIC STATISTICAL MEETING

Nonresponse in

Household Expenditure Surveys

Thomas Laitila, Statistics Sweden, thomas.laitila@scb.se

Abstract

Low response rates are generally obtained in household expenditure surveys collecting data via diary keeping. It is reasonable to question the quality of resulting estimates of population characteristics. However, low response rates do not necessarily mean invalid estimates and the opposite may be equally true. The first may be the case particularly for surveys involving diaries, where the respondent records events during a time interval ahead. This implies the study is prospective meaning the choice of responding is made prior to recording events. Thus, what is to be observed cannot be assumed intervening the choice to respond or not. Dropouts during the data collection period may on the other hand be directly affected by realized values on study variables.

The prospective character of diary surveys is detailed upon showing the possibility to derive valid estimates using auxiliary variables. A main issue addressed is what auxiliary variables to use and how to use them in estimation. It is suggested economic and behavioral theory to be used for the modelling of response probabilities. An illustration with an example from the Swedish household survey is presented, and includes a model derived from economic theory suggested for modelling of response probabilities. The implications of results for the design of diary-based surveys are discussed.

Keywords: diary survey, prospective studies, nonresponse, response probability, auxiliary variables

1. Introduction

In the last 40 years survey statisticians have addressed nonresponse by proposing different weighting adjustments in the estimation stage. From the timeliness and cost perspectives weighting adjustments is a brilliant idea. It avoids the costly and time-consuming double sampling approach.

There are two main approaches, direct and indirect weighting. Both are based on assuming responses being results of random trials also known as quasi-randomization (Oh and Scheuren, 1983). Considering the random trials as independent opens for estimation under a two-stage sampling design. The second stage inclusion probabilities are unknown, of course, but Little (1986) suggests estimating them by estimating a logit model with auxiliary variables assumed to explain response behavior. This is an example of direct weighting.

The formulation of a response probability model introduces a risk of misspecification bias due to wrong functional forms and omission of important explanatory variables. Särndal and Lundström (2005) propose calibration estimators where design weights are adjusted to provide known totals of auxiliary information. Calibration is an example of indirect weighting estimation. However, for showing consistency functional forms in weight calculations and choice of auxiliary variables are still equally important.

Treating response behavior as random means the decision to respond or not is considered a random trial. The trial has two possible outcomes, and the outcome of a response has an unknown probability. Thus, the QR approach is in statistical sense model based, treating a response as an outcome of a Bernoulli trial.

An important issue for the ability to adjust for nonresponse in estimation is how the study variable affects the response probability. The simplest is treating responses as purely random without being affected by any variables; responses are obtained with equal probability for all sampled units. This case is called missing completely at random (MCAR). Cases under MCAR are easily handled by e.g. adjusting estimates with the overall response rate.

Another concept is missing at random (MAR) which means response probabilities varies with some auxiliary variables but are not affected by the study variable. Correction of MAR is simple in theory where adjustment of estimates can be obtained by weighting a unit’s response with the reciprocal of the unit’s response probability. In practice the probabilities are unknown and must be estimated.

The most problematic case is when nonresponse is Not MAR (NMAR). Here the study variables affect the response probabilities. Because observations of the study variable are missing for units not responding, estimation of response probabilities are difficult under NMAR. (See Chang and Kott (2008) for an alternative.)

Most sample surveys at NSI’s collects data on past events and behavior. Such studies are retrospective. Here participation in a survey may very well be decided on the values of study variables. Traditional Household Budget Surveys (HBS) are prospective because respondents are to report future values on study variables. Thus, refusals in the recruiting stage of an HBS cannot depend on the numbers they are to report. Thereby refusals cannot be considered NMAR.

Treating nonresponses as MAR makes modelling of response probabilities less complex. As earlier mentioned, the modelling is difficult in practice because the right auxiliary variables must be included. The choice is presently based on observations of response behavior over subsets of respondents, subsets defined by household characteristics such as age, gender, region of dwelling, and income. However, we have known for a long time an observation of covariation among variables do not imply causality. Thus, reliance on correlation patterns when choosing auxiliary variables renders adjusted estimates with unknown quality properties.

The suggestion in this paper is to utilize theories on human behavior for the modelling of response probabilities. An example adapting an econometric discrete choice model is presented and implications of the model is discussed. Results from an application to HBS data is presented and compared with published estimates.

The next section contains an example on the potential errors introduced, when adjusting nonresponse using correlation based selection of auxiliary variables. Section 3 includes a development of an economic utility model for the choice of responding or not in a survey. The Section is followed by two sections on an adaption of the model to data from the Statistics Sweden HBS in 2007. The final section includes a discussion of results and the design of an HBS. Tables and figures are collected in the Appendix.

2. Inappropriate auxiliary variables

In methods for adjustment of nonresponse it is custom to use auxiliary variables that correlates with the study variable and "explains" response patterns. The meaning of "explains" is usually resorted to using variables with observed differences in response rates over its range, e.g., age and gender. Such an approach may have serious negative impact on estimates. For instance, nonresponse in an angling habit survey is likely to be only indirectly caused by demographic factors and directly by the respondent's interest in angling. In a survey of real estates, nonresponse is likely not directly affected by the size of the estate. It is rather directly affected by the length of the questionnaire and the time required to fill it out.

A simple example is used to illustrate the problem. Let the study variable be an indicator variable for units commuting to work by car. A hypothetical population is shown in Table 1a. Here the variable explaining response probabilities is Region. Units in the City area respond with a 40% probability, while those living in Rural area respond with a 60% probability.

With stratification over Region and SRS (equal sample sizes) the expected response rates are 40 and 60%, respectively. Thus, an expansion estimator weighting responses with the design weight and the reciprocal of observed response rates in the strata yields an approximately unbiased estimator.

Assume a post-stratification w.r.t. gender is made instead. Consider the cross-tabulation of Region X Mode vs Gender in Table 1b. From the table it is realized that cell frequencies in a cross tabulation of Gender vs Mode (Table 1c) are mixtures of units from both regions. In the upper left cell in Table 1c the number of females not commuting to work is 90, where 80 are from City and 10 from Rural. The response probability in City is 40% and 60% in Rural. Then the response probability among females not commuting to work is (0.4*80+0.6*10)/90=38/90. By similar calculations the rest of the cells in Table 1c is obtained.

The marginal response rates in Table 1d shows the responses from females are to be weighted with 1/0.426 while the sample of males is given the weight 1/0.529. Thus, an expansion estimator weighting observations with design weights and the reciprocal of response rates has the approximate expectation

$$22/0.462 + 64/0.529 = 168.6$$

The estimator is biased with 5.4%, approximately. Nonresponse is MAR when conditioning on region, but MAR is not obtained when conditioning on gender. One can note the variable Gender complies with recommendations. Response rates and rates of car usage differ over genders. Further calculations show increased bias when the difference in response probabilities increases between regions.

3. Response utility

If a sampled unit responds it means the utility of responding is larger than not responding. The cost of responding is the time spent and can be measured in terms of loss in value of consumption and leisure time.

Let $\boldsymbol{U}$, $\boldsymbol{C}$ and $\boldsymbol{L}$ denote utility, consumption (monetary terms) and leisure time, respectively. Also, let $\boldsymbol{R}$ denote an indicator for response ($\boldsymbol{R}$ is 1 if response, 0 if not). With this notation the sample unit’s decision problem is to maximize

$$\boldsymbol{U} = \boldsymbol{u}(\boldsymbol{C}, \boldsymbol{L}, \boldsymbol{R})$$

with respect to R. This is made by considering the maximum conditional utilities achieved if responding ($R = 1$) and if not responding ($R = 0$), respectively.

The way to do this is to consider the time τ to spend on responding and how it affects consumption and leisure. For this the following equations are introduced

$$L = T - W - \tau = 0$$

$$C = wW + E + c$$

where T is total time, W is working time, w is wage per time unit, E is income other than salary, and c is a monetary compensation for participating in the survey.

Putting a Cobb-Douglas form on the utility function gives after taking logarithms

$$U = A + \beta \log(wW + E + c) + \delta \log(T - W - \tau) + \gamma R$$

where $0 < \beta < 1$, $0 < \delta < 1$ and γ are parameters expressing utility contribution of goods, leisure, and responding the questionnaire, respectively. The restrictions on the first two is to have a function with realistic properties. For the same reason they are restricted to satisfy $\beta + \delta < 1$. A is a constant not influencing the choice to make. Differentiating w.r.t. W and setting the derivative to zero yields the solution

$$W_\tau = \alpha(T - t) - (1 - \alpha)\frac{E + c}{w}$$

which is the optimal working time when saving τ time units for responding to the survey with a compensation of c monetary units. Here $\alpha = \beta/(\beta + \delta)$.

Putting the solution in the utility function and evaluating the difference in utilities between $\boldsymbol{R}=\mathbf{1}$ $(\boldsymbol{\tau}>\mathbf{0})$ and $\boldsymbol{R}=\mathbf{0}$ $(\boldsymbol{\tau}=\mathbf{0})$ yields the criterion for responding

$$\boldsymbol{\gamma}+(\boldsymbol{\beta}+\boldsymbol{\delta})\log\left(\mathbf{1}-\frac{\boldsymbol{\tau}/\boldsymbol{T}}{\boldsymbol{\rho}}+\frac{\boldsymbol{c}}{\boldsymbol{wT}+\boldsymbol{E}}\right)>\mathbf{0} \qquad (1)$$

where $\boldsymbol{\rho}=\mathbf{1}+\left(\boldsymbol{E}/_{\boldsymbol{wT}}\right)$ and the ratio is other income than salary to maximum possible salary for the time $\boldsymbol{T}$.

In the usual case of no monetary compensation, $\boldsymbol{c}=\mathbf{0}$, the second term on the l.h.s. in (1) is negative. This means the utility of responding, measured by $\boldsymbol{\gamma}$ must be positive and large enough to compensate for the negative impact of spending time on the questionnaire. With $\boldsymbol{\gamma}$ not sufficiently large, the respondent needs a compensation for a response.

If $\boldsymbol{\gamma}$ is negative, the respondent does not find a gain in the mere reply to a survey. This means the respondent must be monetarily compensated to provide a response. To illustrate the level of compensation needed in a household budget survey with a diary, assume the data collection is over a fortnight period. The total time available under the period is assumed $\boldsymbol{T}=\mathbf{168}$ (hours), reducing total time for sleep and other necessary personal activities. The person is also assumed to have no other income than salary giving $\boldsymbol{\rho}=\mathbf{1}$, and earns 10 monetary units (munits) per hour. With $\boldsymbol{\tau}=\mathbf{7}$, half an hour per day, the criterion equals

$$\boldsymbol{\gamma}+(\boldsymbol{\beta}+\boldsymbol{\delta})\log\left(\mathbf{1}-\frac{\mathbf{7}}{\mathbf{168}}+\frac{\boldsymbol{c}}{\mathbf{1680}}\right)>\mathbf{0}$$

The compensation needed to have a positive second term on the left is $\boldsymbol{c}$ >70. When $\boldsymbol{\gamma}$ is negative, the compensation must be much larger.

The restriction $\boldsymbol{c}$ >70 can be explained. The marginal values of working time and leisure time are decreasing. Without responding utility maximization sets equal marginal values of the last hour in work and in leisure, respectively. These marginal values are measured by the wage per hour, 10 munits. The contribution to utility by previous working and leisure time hours have values larger than the marginal ones. Thus, to compensate for 7 hours of diary reporting the amount must exceed 70 munits.

If a person is not working, consumption and leisure is not affected by changes in working time. This means the decision to respond is made if $\boldsymbol{\gamma} > \mathbf{0}$ when $\boldsymbol{c} = \mathbf{0}$, and if $\boldsymbol{\gamma} > -(\boldsymbol{\beta} + \boldsymbol{\delta}) \log\left(\mathbf{1} + \frac{\boldsymbol{c}}{\boldsymbol{E}}\right)$ when $\boldsymbol{c} > \mathbf{0}$.

4. Data and discrete choice modelling

The empirical part is based on data from the 2007 HBS conducted by Statistics Sweden. The analysis is restricted to one-person households, with or without children. Households with two or more adults requires a more complex specification of the decision process because several persons may be involved in the decision. This modelling problem is avoided here.

Data contains 1 241 observations in the sample of which 530 households are in the response set. Data on background characteristics of the household head is available and includes variables such as age, gender, number of children, etc. For the modelling of choice behavior economic variables data includes total salary, other incomes, tax paid, etc.

Parameters in the l.h.s. in inequality (1) are specific for each sampled unit. These parameters are modelled as

$$\boldsymbol{\gamma_k = \gamma_0 + \gamma_1 x_{k1} + \cdots + \gamma_p x_{kp} + \varepsilon_k}$$

$$\boldsymbol{\beta_k + \delta_k = \theta_0 + \theta_1 x_{k1} + \cdots + \theta_p x_{kp}}$$

The $\boldsymbol{x}$'s represents $\boldsymbol{p}$ variables on units' characteristics, the $\boldsymbol{\gamma}$'s and $\boldsymbol{\theta}$'s are parameters. In the first equation $\boldsymbol{\varepsilon_k}$ is an unobservable zero mean random variable, independent among sample units.

The logarithmic term in (1) is represented by

$$\boldsymbol{z_k} = \mathbf{log}\left(\mathbf{1} - \frac{\boldsymbol{\tau_k}/\boldsymbol{T}}{\boldsymbol{\rho_k}} + \frac{\boldsymbol{c}}{\boldsymbol{w_k T} + \boldsymbol{E_k}}\right)$$

With $\boldsymbol{\varepsilon_k}$ as a random variable, the probability of a response equals

$$\mathbf{P}_{\boldsymbol{k}} = \mathbf{P}\left(\boldsymbol{\varepsilon_k} > -(\boldsymbol{\gamma_0} + \boldsymbol{\gamma_1 x_{k1}} + \cdots + \boldsymbol{\gamma_p x_{kp}} + \boldsymbol{\theta_0 z_k} + \boldsymbol{\theta_1 x_{k1} z_k} + \cdots + \boldsymbol{\theta_p x_{kp} z_k})\right) \tag{2}$$

Assuming $\boldsymbol{\varepsilon_k}$ normally distributed with zero mean and variance 1, Model (2) is estimated with Probit Maximum Likelihood. If $\boldsymbol{\varepsilon_k}$ is assumed logistically distributed, the model is estimated with the Logit Maximum Likelihood estimator. The two estimators give qualitatively the same results. A prior choice is here Probit because of the uncertainty in Logit ML t-statistics of estimated parameters.

It can be suspected the valuation of responding to a survey ($\boldsymbol{\gamma}$) is made different by units not working compared to those who are. For this reason, another set of interaction terms are added to yield the extended parametrization

$$\boldsymbol{\gamma_k} = \boldsymbol{\gamma_0} + \boldsymbol{\gamma_1 x_{k1}} + \cdots + \boldsymbol{\gamma_p x_{kp}}$$

$$+\boldsymbol{\gamma_{10} D_k} + \boldsymbol{\gamma_{11} D_k x_{k1}} + \cdots + \boldsymbol{\gamma_{1p} D_k x_{kp}} + \boldsymbol{\varepsilon_k}$$

where $\boldsymbol{D_k}$ is an indicator variable for units not working.

5. Results

Probit ML estimates of model (2) with the extended parametrization of $\boldsymbol{\gamma_k}$ is shown in Table 1. The Unrestricted model includes all variables and interactions. Comparing parameter estimates with standard errors shows estimates are uncertain and have high t-values. However, the model provides a significant fit with the LR statistic 106.3 with 20 degrees of freedom.

After sequentially removing variables with p-values > 0.10, the estimates in the Restricted model is obtained. In the model of $\boldsymbol{\gamma_k}$ age and income variables are significant. There is also a significant difference in the value of responding between those working and those who are not.

Concerning estimates of parameters for $\boldsymbol{z_k}$ log(age) is significant with a negative sign meaning the importance of time reduces with age. Interestingly, the logarithm of household size is also significant with a negative sign. Again, it means importance of time is reduced with household size.

Using the estimate of the Restricted model, response probabilities is estimated for each unit in the sample. In Figure 1 the estimates are plotted against age. The general picture is response probability increases with age. Four segments are identified. One group in the age interval 20 – 30 with decreasing response probabilities and another group older than 30 years with increasing probabilities. There are also two groups at different levels although with similar sinusoidal patterns.

The estimated response probabilities are used to weight observations in the response set. Estimates are obtained for two domains, single living with and without children, respectively. Estimates of consumption in total and for a few product groups are presented in Table 3. Published estimates from the 2007 HBS are included for comparison.

The striking observation is the estimates obtained here are close to those published. All estimates are within the interval estimates published. It is also noted that the estimates obtained here are all slightly larger than those published. Corrections for nonresponse of the published estimates is partly based on the variables age and disposable income, the same variables used in the Probit model. A region variable and an indicator for birthplace in Sweden are also used.

6. Discussion – Design of an HBS

Section 2 shows the importance of using appropriate auxiliary variables when adjusting for nonresponse. Use of correlations in selection of variables are not sufficient. What matters is if the auxiliary variables are the source to differences in response probabilities among sampled units. How to identify those variables is a multi-million-dollar question.

A decision to respond or not is behavioral and the scientific literature is rich on theories explaining choices made by people. This literature provides one strand of finding appropriate auxiliary variables for nonresponse adjustments. Such an approach also gives a theoretical foundation for the models and adjustments made.

This paper adapts economic theory on choice behavior and an econometric dichotomous choice model. The exercise shows what variables to include in the model and how to include them. The theory also defines ranges of appropriate values on parameters in the model. This gives a tool for evaluating the model fitted and the derived weights in nonresponse corrections.

The results in the empirical part on HBS estimates indicate no major difference between estimates obtained here and those published based on calibration adjustments. An erroneous conclusion is the method chosen does not matter. Neither correlation analyses nor theory can guarantee the right auxiliary variables are used in an adjustment. Theories, however, are empirically tested and explains how and why a variable affects the decision to respond or not.

Design of an HBS

A traditional HBS puts a demanding task in the hands of the respondents. Low response rates are no surprise. What data and how it is collected is to be carefully considered in the design of an HBS.

The sampling design is also of interest regarding the nonresponse problem, e.g. household types with low response rates are to be oversampled. Cluster sampling which allows for more versatile data collection methods can be a mean of increasing the response rates. An example is when households with low response rates can be narrowed down to smaller geographical areas.

Irrespective of the final survey design a larger rate of nonresponse is expected. As illustrated in this paper the prospective character of an HBS makes it feasible to adjust for households declining participation. The time required if joining the survey is most likely to be a determinant factor for participation. Using choice theory putting the time in the context of the individual household is an appropriate option for modelling.

Dropouts after accepting participation poses a more challenging problem. This sort of nonresponses is like nonresponses in retrospective studies and can be NMAR. A plan for handling dropouts must therefore be developed. The main information needed is the cause of the dropout. Data on some particular variables useful for nonresponse adjustments can also be collected.

References

Chang, T. and P.S. Kott (2008). Using calibration weighting to adjust for nonresponse under a plausible model, *Biometrika*, 95:3, 555-571.

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means, International Statistical Review, 54:2, 139-157

Oh, H.L. and F.J. Scheuren (1983). Weighting adjustment for unit nonresponse. In: Madow, W.G, Olkin, I. and D.B. Rubin (Eds.), *Incomplete Data in Sample Surveys:Vol 2*, Academic Press, New York, pp. 143 – 184.

Särndal, C.E. and S. Lundström (2005). *Estimation in Surveys with Nonresponse*. Wiley, New York.

Appendix

Table 1a: Hypothetical population, strata vs mode of work commuting.

Region	Not by Car	By Car	Response probability
City	100	50	0.4
Rural	40	110	0.6

Table 1b: Hypothetical population, strata vs gender.

Strata	Commuting	Female	Male	Response probability
City	Not by Car	80	20	0.4
	By Car	10	40	0.4
Rural	Not by Car	10	30	0.6
	By Car	30	80	0.6

Table 1c: Hypothetical population, gender vs mode of work commuting. (Cell response probabilities in parenthesis.)

	Not by Car	By Car
Female	90 (38/90)	40 (22/40)
Male	50 (26/50)	120 (64/120)

Table 1d: Hypothetical population, design weighted expected number of responses in cells of gender vs work commuting mode, and marginal response probabilities w.r.t. gender.

	Not by Car	By Car	Response probability
Female	38	22	0.462
Male	26	64	0.529

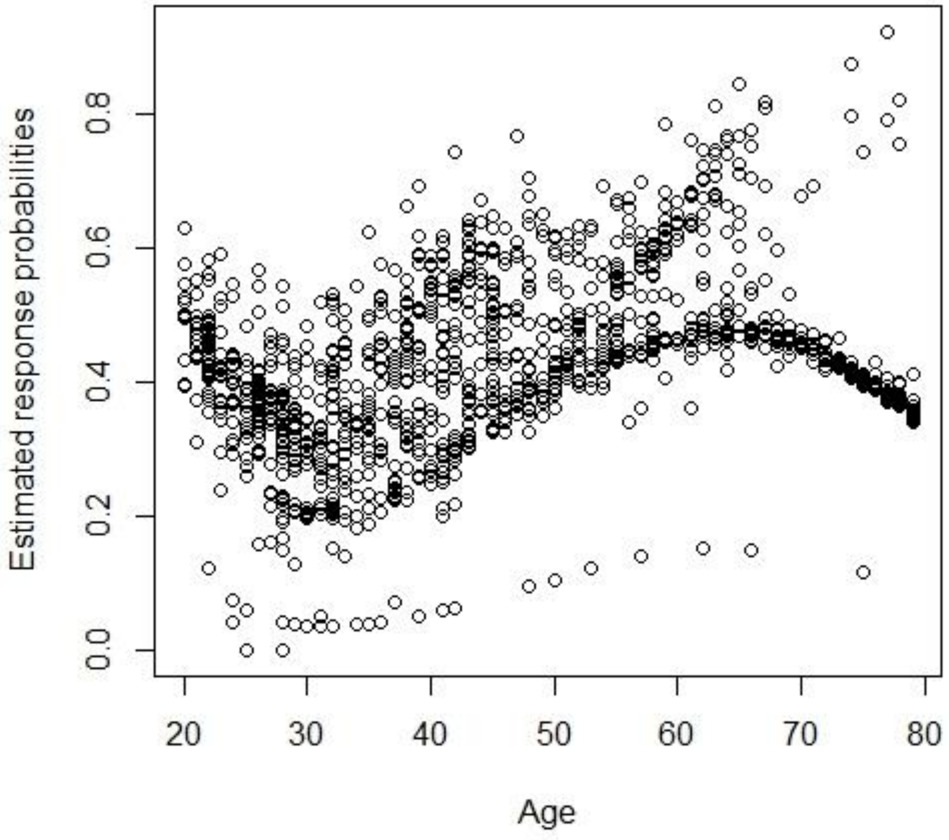
Table 2: Probit ML estimates of Model (2).

Variables	Unrestricted Model		Restricted Model	
	Estimate	St.err	Estimate	St.err
Constant	34.9	116	13.2	3.82
Main effects				
Log(N:o persons)	-.258	.573		
Age	.112	.645	.142	.027
Gender	-.157	.418		
Log(Disposable Income)	.129	.200	.284	.102
Log(Age)	-17.5	76.0	-6.17	1.35
Log(Disposable Income)^2	1.72	14.0		
$\boldsymbol{z_k}$	841	3053	108	56.2
$\boldsymbol{D_k}$	91.8	124	95.8	40.5
Interaction effects				
$\boldsymbol{z_k}$ Log(N:o persons)	-16.0	14.7	-10.8	2.77
$\boldsymbol{z_k}$ Age	-1.47	17.7		
$\boldsymbol{z_k}$ Gender	-4.91	10.9		
$\boldsymbol{z_k}$ Log(Disposable Income)	-5.39	6.17		
$\boldsymbol{z_k}$ Log(Age)	-421.9	2004	-29.3	14.8
$\boldsymbol{z_k}$ Log(Disposable Income)^2	61.9	370		
$\boldsymbol{D_k}$ Log(N:o persons)	.347	.607		
$\boldsymbol{D_k}$ Age	-.783	.688	-.713	.205
$\boldsymbol{D_k}$ Gender	.417	.440		
$\boldsymbol{D_k}$ Log(Disposable Income)	-.059	.202	-.202	.106
$\boldsymbol{D_k}$ Log(Age)	-66.1	80.8	-65.6	25.9
$\boldsymbol{D_k}$ Log(Disposable Income)^2	13.5	14.8	13.0	4.64
Deviance	1587.6	df=1220	1594.7	df=1229

Table 3: Estimated mean yearly household expenditures. Estimates obtained by Discrete Choice modelling (DC) and reported estimates in HUT 2007.

	Single with children		Single without children	
Expenditures	DC model	HUT 2007	DC model	HUT 2007
Total expenditures	234 535	229 290	168 595	167 540
		±17 100		±9 910
Food	29 232	28 310	16 508	17 280
		±2 360		±1 080
Clothes and shoes	11 899	11 590	9 172	8 230
		±2 640		±1 670
Healthcare	4 140	4 060	4 282	4 020
		±1 190		±1 270

Figure 1: Estimated response probabilities versus age of household head. (Single living households, with or without children. Model estimates reported in Table 1.)


Estimated response probabilities
Age