Economics
The Open-Access, Open-Assessment E-Journal

Vol. 3, 2009-30 | June 23, 2009 | http://www.economics-ejournal.org/economics/journalarticles/2009-30

Social Policy Targeting and Binary Information Transfer between Surveys

Daniel Gottlieb and Leonid Kushnir
Ben-Gurion University; CONSIST Ltd.

Abstract

In this paper we develop a methodology for identifying a population group surveyed latently in the (target) survey relevant for further processing, for example poverty calculations, but surveyed explicitly in another (source) survey, not suitable for such processing. Identification is achieved by transferring the binary information from the source survey to the target survey by means of a logistic regression determining group affiliation in the source survey by use of variables available also in the target survey. In the proposed methodology we improve on common matching procedures by optimizing the cut-value of the probability which assigns group affiliation in the target survey. This contrasts with the commonly used "Hosmer-Lemeshov" cut-values for binary categorization, which equates between the sensitivity and specificity curves. Instead we improve group identification by minimizing the sum of total errors as a percent of total true outcomes.

The Jewish ultra-orthodox population in Israel serves as a case study. This idiosyncratic community, committed to the observance of the Bible is only latently observed in the surveys typically used for poverty calculation. It is explicitly captured in the social survey, which is not suitable for poverty measurement.

This procedure is useful for ex-post enhancement of survey data in general.

JEL: C15, D63, I38, Z12

Keywords: Group identification; binary variables; optimal cutoff value; poverty; targeting

Correspondence

Daniel Gottlieb, National Insurance Institute and Economics Department of Ben-Gurion University, danielgt@nioi.gov.il; Leonid Kushnir, CONSIST Ltd.

We bear sole responsibility for any views or mistakes presented here. We thank an anonymous referee for useful comments and the Central Bureau of Statistics, Israel for the support in preparing the data base.

1 Introduction

The Haredi (Jewish ultra-orthodox) population in Israel is an idiosyncratic community, committed to the observance of the Bible and its commandments, as interpreted by its sectarian religious leaders. Haredi poverty incidence is exceptionally high at 67.5%, with a share of 20% of all Israeli poor while its share in the total population is only half that size. Its major causes are a very high Haredi fertility (a population growth of 6% p.a.), reducing both household income per capita and the mother's earning capacity; its education system is largely independent from the national school system and neglects (particularly among boys) materially important subjects for the buildup of future earning capacity such as Mathematics, English and digital skills; a low labor-force participation of Haredi men, due to prolonged learning in religious seminars (Yeshiva), often deep into the prime working age. A further cause for the sharp increase in short-term poverty has been the recent large cuts in child benefit payments.[1]

The share of Haredi children up to age 4 is nearly 3 times higher than in the rest of the Jewish society. This, together with the empirical regularity of a negative relationship between poverty and age implies an upward-drift for Haredi and overall Israeli poverty over time. Haredi Poverty, as measured by the distribution-sensitive Sen-poverty index, nearly doubled over the last 3 years after a previous significant improvement. This deterioration stands in contrast to developments in the rest of the Israeli-Jewish society, whose poverty intensity increased only slightly over the last couple of years.[2]

In 2004 Haredi male labor force participation of 37% hardly exceeded one half that of the other Jewish male population, mainly due to the Haredi high enrollment in religious seminars (Yeshiva) during their prime working age. Despite their much higher fertility the women function as the family's main providers, with a participation rate of 48%, compared to 58% of non-Haredi women. Preliminary and still statistically insignificant empirical evidence points to a recent increase in Haredi labor-market involvement, both among men and women, probably related to increased economic hardship, maybe due to the drastic cut in child allowances from 2002 to 2004.

Empirical evidence shows job training to affect labor force entry positively, particularly among Haredi men, though at low wages.[3] These schemes proved successful tools when conceived with a high sensitivity towards the particular cultural needs of the Haredi society.

Proper identification of the poor is essential for social policy targeting. When the poor belong to a specific cultural group with exceptionally high poverty incidence, the basic determinants of their poverty are typically centered on family size, educational deficiencies and labor market behavior. However, these characteristics might have more deep-seated cultural roots, reflected in collective preferences concerning fertility and gender-related differences in education and labor-force participation. Such underlying cultural determinants which are of a qualitative rather than quantitative nature should be included explicitly in the poverty analysis.

[1] See Gottlieb (2007) (in Hebrew), "Poverty and Labor Market Behavior in the Ultra-Orthodox Population in Israel".

[2] Haredi poverty, labor market behavior and the effect of training are analyzed in Gottlieb (2007) op.cit.

[3] See Gottlieb (2007), op.cit., Table 6, 32.

Such information is usually lacking in standard income or consumption surveys typically used for poverty calculations. Such information is to be found in special social surveys, based on the same population, but not suitable for standard poverty calculations. The question arises whether an efficient procedure can be devised by which such information could be transferred ex-post from the original survey, the source-survey, to the relevant surveys for poverty calculation, i.e. the target-surveys. Such a procedure may be a useful tool for ex-post enhancement of the information content of survey data in general.

The major social and economic surveys of an economy focus typically on different aspects of the same population. While some of the questions recur in more than one survey, other information is unique to a specific survey. Since the gathering of information is not costless and some of the survey-specific information might be useful to researchers of another survey or to policy makers, we suggest an efficient method for optimal binary information transfer (BIT) from one survey (the "source" survey) to another (the "target" survey).

The optimal method for transferring the information depends crucially on three aspects of the process: (1) an overlap of the set of variables that contain explanatory power of the variable to be transferred (thus ensuring a reasonable goodness of fit of the Receiver Operating Characteristic-curve, henceforth ROC-curve), (2) the rule for determining the cutoff value, i.e. the value by which the logistic probability forecast is translated back into a binary variable and (3) a quality test of the procedure. Our quality test, while performed in the source survey, still provides a clue to the quality of the synthetic information in the target data set.

Such enhancement of socio-economic data by an ex-post information transfer is particularly useful when additional data collection by a survey is either too expensive or impossible. Our method for choosing the cutoff value of the forecasted probability is shown to improve on that suggested by Hosmer and Lemeshow, 2000.

BIT has several possible applications. It can be useful in the targeting of policies to specific population groups, which is one of the purposes of poverty mapping.[4] The present results might also be used in medical research and other research employing logistic regression and cutoff values.[5]

We then illustrate the application of the method to the measurement of poverty in a specific group, known for its high poverty incidence – the Israeli Jewish Ultra-orthodox ("Haredi") population. Due to the lack of information on religious affiliation in the surveys typically used for poverty calculations (the income- and expenditure surveys), and the lack of sufficiently detailed income and consumption data in the survey that does provide information on religious affiliation there arises a need for the transfer of

[4] In recent years poverty mapping has become an important tool in improving targeting. This technique utilizes information from surveys, amenable to poverty calculations, but too small for efficient targeting of the poor, by transferring information to large scale data bases such as census data, which provide less detailed information, but on a larger share of the population. Such a transfer is carried out by use of econometric tools. The purpose is to enable the calculation of policy variables, for example binary information on poverty, for small geographic areas. See for example Hentschel, J., Lanjouw, J.O., Lanjouw, P. and Poggi, J. (2000), Bigman and Srinivasan (2002) or Small Area Estimation at www.worldbank.org.

[5] See for example Hadjicostas Petros and George C. Hadjinicola (2001), G. Schares et al. (2003), Schutter E.M.J. et al. (1998) and Stegeman, J.A. et al. (2006)

information on Haredi membership from the Social Survey to the Income and Expenditure surveys.

The paper is organized as following: In chapter 2 the model of BIT is presented. Chapter 3 describes the process of BIT in more detail. In chapter 4 we report on a case study of BIT applied to the Israeli Haredi population for the purpose of poverty calculations.[6] Concluding remarks complete the paper.

2 The Model

Assume a sampling of two Household surveys, one which we call the Source-survey (*S*), consisting of $n_S = 1\ldots S$ households, and another survey sampled on the same population[7], which we call the Target survey (*T*), $n_T = 1\ldots T$. Let there be a dichotomic binary group variable, say of group *H*, with a value of 1 for success and 0 for failure. We denote the household's probability of event $H = 1$ occurring, as π_i and its estimate as $\hat{\pi}_i$.

The estimate is conditional, based on vector x of explanatory variables, $\hat{P}(H\text{=}1|\ \mathrm{x}')$ = $\hat{\pi}$(x) where vector $x' = (x_1, x_2, \ldots x_k)$. $\hat{H}_i^{S,T}$ is a binary estimate of *H* for individual i in the respective sample of the source (*S*) or the target survey (*T*). Obviously, the suggested procedure requires vector x' to appear in both *S* and *T*. The logistic probability function for event *H*=1 is given by

$$\pi_i(x) = \frac{e^{g_i(x)}}{1+e^{g_i(x)}} . \tag{1}$$

The logit equation includes continuous (k=1…K) and categorical variables (D_{jl}, j=1…J), such as simple dummy variables or dummy variables with more detailed coding levels (l=1…L-1):

$$g_i(x) = \beta_0 + \beta_1 x_1 + \ldots + \sum_{l=1}^{L_j-1} \beta_{jl} D_{jl} + \beta_K x_K \tag{2}$$

3 The BIT Process

Step 1: Search for an Efficient Logistic Regression in the Source Survey

The quality of BIT depends crucially on the explanatory power (not necessarily in a causal sense) of equation (2) of group membership probability in the Source survey ($\hat{\pi}_i^S$). The better the explanatory power, as reported in the regression's log-likelihood

[6] Gottlieb (2007), op.cit..

[7] Since the households are chosen by specific mechanical processes the chances that the same household will appear in more than one survey is negligible. Of course if it does, and the researcher knows that information, then the information transfer becomes trivial.

ratio, Wald test, the z-values and additional statistical parameters, the better is the chance for a successful BIT of household i's group membership.

Step 2: A Forecast of Group Membership, Using a 'Continuous' Cutoff Value ($\hat{\pi}_c^S$)

We choose any cutoff point $0 \leq \hat{\pi}_c^S \leq 1$ in the source survey, above which the forecast of household *i*'s group membership ($\hat{H}_i^S$) is either 1 or 0. We repeat this procedure, covering the whole range of $0 \leq \hat{\pi}_c^S \leq 1$. Consequently, $\hat{H}_i^S = \hat{H}(\hat{\pi}_c^S)$ for *i*=1…*S*. For each cutoff value we then organize the binary outcomes of $\hat{H}_i|_{\hat{\pi}_c}$ into 4 mutually exclusive categories:
True Positive Outcomes: $TP(\hat{\pi}_c^S)$ for all $\hat{H}_i = H_i = 1$,
True Negative Outcome: $TN(\hat{\pi}_c^S)$ for all $\hat{H}_i = H_i = 0$,
False Positive Outcome: $FP(\hat{\pi}_c^S)$ for all $\hat{H}_i = 1$ and $H_i = 0$,
False Negative Outcome: $FN(\hat{\pi}_c^S)$ for all $\hat{H}_i = 0$ and $H_i = 1$,
These steps are repeated for a near-continuous number of cutoff values.

Step 3: Assessment of the Forecast Quality in the Source Survey

The error rate or forecast quality can only be estimated in the source survey since the target survey includes only the set of explanatory variables and not the dependent variable. We characterize the forecast quality using the ROC curve as a measure.

Step 4: Searching for the Optimal Probability Cutoff Value ($\hat{\pi}_c^{S,*}$)

We choose the optimal cutoff value by using the outcomes of the previous step, i.e. the cutoff value that minimizes the sum of total squared errors FP and FN. Notice that Hosmer and Lemeshow (henceforth HL) suggest that the optimal cutoff value is at the level $\hat{\pi}_c^{S,*}$ for which sensitivity equals specificity. In the following we show that in the present case our choice yields a significant improvement on the HL choice.

Step 5: The BIT - Calculation of the Forecast $\hat{H}_i^T$ in the Target-Survey

After having ascertained that we have elicited the best possible forecast we move to the target survey. As mentioned before there is no way of testing the quality of BIT, except by new data collection. We calculate $\hat{H}_i^T$ by use of the regression equation and the optimal cutoff value as estimated in the source-survey.

4 A BIT Case Study: Poverty among the Jewish Ultra-Orthodox in Israel

The Israeli Ultra-Orthodox Jewish society, also called Haredi society,[8] has long been known to have an exceptionally high poverty incidence. However, since there is no indication of Haredi affiliation in the surveys used for estimating poverty, available poverty studies and in particular the official ones do not report separate poverty estimates for this population group. Some studies have attempted to estimate poverty in this population group on a national level and we shall discuss them below.

Due to its idiosyncratic cultural features and a lack of their explicit inclusion in survey questions the Haredi society is an interesting example for applying the BIT process. Their heterogeneous labor market behavior leads to extreme poverty situations of many Haredi households. Consequently, there arises a need for statistical enhancement concerning Haredi group membership in the major surveys used for rational policy formulation and implementation. In order to model a logistic group membership probability of the Haredi (the first step of the BIT process) we briefly characterize them here.

4.1 The Israeli Haredi Society – Roots and Characteristics

The Israeli Haredi society is fragmented into several subgroups, each emphasizing different aspects of Judaism and obeying its own spiritual leaders. For simplicity we concentrate on three main factions: The Hassidic, the Lita'i and the Sephardic[9] groups. They all share strict observance of the Torah and the Jewish commandments and a high degree of compliance to their spiritual leaders' decisions concerning a wide range of public and Family issues. The leadership maintains a strong sense of hierarchy and issues and looks over detailed rules for individual and family behavior through its organs. When in the early 20th century secular Jewish nationalism emerged as a rapidly growing alternative to the religious way of life, the Haredi rejected its anti-religious character strongly. Since a historical compromise in 1948 between David Ben-Gurion (then Prime Minister) and Hazon Ish[10] (then Leader of the Haredi society) male religious scholars, whose main occupation is Torah study, Haredi women and men in drafting age are by and large exempted from serving in the Israeli army. Over the years the number of exempted men from the army grew rapidly (8 to 9% p.a., reaching more than 30,000 in the early 21st millennium from about 400 in 1948. In response to court appeals and a general public discontent over that exemption from army service and to a mounting social problem of young drop-outs from the Haredi religious seminars, an

[8] "Haredi" is the Hebrew name of the Ultra-Orthodox society. It has the meaning of a person who "trembles in awe of God". It includes distinct groups, common in their unequivocal commitment to the study and observance of the Torah and its commandments, as interpreted by their religious leaders. See also Friedman (1991).

[9] Sephardic originally indicated the Judeo-Spanish origin. In the Israeli context it is sometimes used in a wider context to indicate also other Jews originating from North Africa or the Middle East.

[10] Rabbi Abraham Yishayahu Karelitz (1878-1953) The compromise included also an exemption of Haredi girls. In the years 1951 and 1952 the argument over drafting Haredi women developed into a government crisis, causing the Haredi party Agudat Israel to leave the government.

official commission, appointed in 1999, proposed a change in government policy with the purpose of reducing the number of exemptions from the army and of improving Haredi men's labor force participation.[11]

4.2 Estimates of Haredi Population Size

Several attempts to estimate the size of the Haredi population were based on the question about the "last school visited" in the household surveys of the ICBS. This education-based approach was pioneered by Berman and Klinov, 1997 and also Dahan, 1998. It was elaborated in Berman, 2000, and has since then been used for analyzing Haredi poverty and labor market behavior.[12] Other population estimates were based on election results due to typically monolithic Haredi voting-patterns.[13]

4.2.1 The Education-based Approach

According to this approach a household is assumed to be Haredi if at least one of its male members indicates a Yeshiva (a religious seminary)[14] as the last school attended. Berman (2000) forecasted Haredi population to reach 280,000 in 1995 and 510,000 people by 2010, based on expected fertility and death rates. Such forecasts are bound to produce unsatisfactory results for a number of reasons: Yeshiva studies do not constitute a necessary condition for Haredi belief. Indeed, Yeshiva attendance among Hassidic Jews, a large group within Haredi society, is believed to be lower than in the Lita'i and Sephardic Haredi groups.

4.2.2 The Elections-based Approach

This approach was chosen by Gurovich and Cohen (henceforth GC), based on the 2003 elections[15] and a geographic identification of localities with a high percentage of voters for the two political parties of Haredi orientation out of the 13 party lists represented in the parliament: United Torah Judaism (UTJ, or in Hebrew "Yehadut HaTorah") and "Shas"[16]. While the voters for UTJ are supposedly mainly Haredi, many "Shas" supporters are less religious but rather traditional or ethnically oriented voters. In order to identify this subgroup, GC included Shas supporters among the Haredi only if they lived in the vicinity of areas with a high percent of UTJ support, assuming that the

[11] See the report of the Tal Commission (2000).

[12] See for example Flug and (Kaliner) Kasir (2003), Gottlieb and (Kaliner) Kasir (2004) and Gottlieb and Manor (2005).

[13] See Degani and Degani (2000), and more recently Gurovich and Cohen (2004).

[14] These seminaries are not to be confounded with religious High schools (Yeshiva Tihonit), which combine religious studies with a high school curriculum. The latter are typically frequented by orthodox rather than ultra-orthodox Judaism. Orthodox Jews, distinctly from the Ultra-orthodox are fully integrated in the Israeli society, its labor market as well as in the army.

[15] An earlier study by Degani and Degani (2000), was based on the 1996 elections.

[16] The "Shas" party of Torah-observant Sephardis was founded in 1984. It has many non-orthodox supporters.

Haredi like to live whitin each other's proximity. GC concluded that only 1/3 of the Shas voters are Haredi. The population estimate is calculated as following:

$$H_{\mathrm{Pop}} = \sum_{j}(\sum_{i} i\,\mathrm{voters} / p_j)/(1\text{-}x_j)$$

where i = number of voters for each party, j = UTJ party/Shas party, where p_j = election-participation rate of the j^{th} party. x_j = percent of population under voting age of the j^{th} party supporters. In areas with a high rate of UTJ voters, the researchers report a high participation rate compared to other areas. In areas with 90% and more UTJ votes the general participation rate was 94%. In areas with 80% and more UTJ votes, the general participation rate was 85%. The study assumes a significantly higher Haredi election participation rate than that of the general public.[17] Based on fertility rates derived from the Social survey for Haredi women of Ashkenasi[18] background, GC used a fertility of 7.5 births per woman yielding an estimate of the share of people below the voting age (based on a model of stable populations) of 56% of the population. The total population of actual and potential UTJ voters is estimated to be 361,000. The Sephardic Haredi estimate amounted to 204,000 and the total Haredi population was estimated at 565,000 by the end of 2002.

4.2.3 The Estimate Based on the Social Surveys of the ICBS

The first Social Survey with a sample size of some 10,000 persons aged 20 or more and their household was published in 2002. The estimate of Haredi affiliation is based on question Nr. 26 of the questionnaire[19]. In order to estimate the population size including children, the weights need to be adjusted to account for the fact that in a household there may be more than one person aged 20 or more. We calculate population size by use of the following formula[20]:

$$H_{\mathrm{Pop}} = \qquad + \mathit{under}20_{\mathrm{i}} \times \qquad)$$

where H_{Pop} = Haredi population, i = people declaring themselves as Haredi and nn_i = population weight for each respondent. According to the Social Survey they were 194.9 thousand by end of 2002. Under20_{i} = number of people aged under 20 in the i^{th} (Haredi) household and over20_{i} = number of additional people (to those questioned)

[17] The general participation rate in the 2003 elections was 67.8%. When adjusted for the very low Arab election participation rate and for the Israelis who were absent during the elections, the general participation rate is somewhat higher but still lower than the Haredi participation rate.

[18] In the present context this indicates a European (including Eastern European and Russian) and Anglo-Saxon background.

[19] Question 26: "Do you consider yourself (1) Haredi, (2) religious, (3) traditional-religious, (4) traditional and "not so" religious, (5) non-religious or atheist. In order to estimate the population size, the detailed data set is needed, including information on the other household members, their age and the weights attached to the interviewed person. These and more data were kindly provided by the CBS.

[20] We thank Tsahi Makovki from the ICBS for providing the formula.

aged 20 or more, in that household. According to this calculation, the Haredi population reached about 550,000 by end of 2002.[21]

As shown in Table 2 the new estimate exceeds the commonly accepted estimate of empirical economists by 62 percent.

We find the population estimate of the Social Survey to be consistent with the calculations of GC, who calculated the size of the Haredi population based on revealed (party-) preference from the 2003 election results.

4.2.4 The BIT Estimate

The variable reflecting true group membership and chosen as a benchmark for the competing estimates is the sampled person's own declaration of Haredi affiliation (group membership). In the present case study this seems to be the most natural approach since religious affiliation is first of all a subjective cognition. In other examples of group membership one might be looking for a variable reflecting an objective recognition of group membership (e.g. a valid passport for citizenship, a university degree for being an academic etc.) as the benchmark that might be preferable.

Step 1: Search for an Efficient Logistic Regression in the Source Survey

Based on prior knowledge of the distinctly high Haredi fertility and fundamental changes in fertility over the last generations we decided to split the data into 3 subgroups by the mother's age in order to improve the overall empirical results. In order to identify Haredi families we analyzed differences in fertility patterns and demographic characteristics (the mother's age child ratio, country of origin of the head of household), educational characteristics (Yeshiva as the last School attended by one of the male household members or No High School or University diploma), geographic concentration (areas with high Haredi concentration), differences in social behavior (philanthropic behavior, no use of internet), living conditions (car ownership, number of children per room).

The regression results for the families, grouped by the mother's age and the variable definitions are given in Appendix Table 1. The coefficient vectors $\hat{\hat{\beta}}_i (i = 1, 2, 3)$ are needed for the BIT process in order to calculate the estimated probabilities by use of the logit functions $g_i(x)$ with the relevant coefficients for each group respectively as mentioned in equations (1) and (2). The logistic regression model (Appendix Table 1) is statistically significant as can be seen from the log likelihood statistic and the Wald-test in Table 1.

[21] The total population should add up to 6.59 million people, but yields only 6.19 million. The discrepancy may be due to the de-facto exclusion of the Eastern-Jerusalem Arab population, Bedouins in non-recognized settlements and people staying in non-sampled institutions. Furthermore the weight adjustment reflects only an approximation of the true weight.

Table 1: Model Significance of the Logistic Regression of Haredi Group Membership

(Regression results from Appendix Table 1)

	20-30	31-40	41+
LR test			
Log likelihood	-138.074	-122.1076	-315.4222
Probability(LR stat)	0.0000	0.0000	0.0000
Wald test			
Value	213.48	210.61	727.74
Probability	0.0000	0.0000	0.0000

Step 2: A Forecast for Group Membership, Conditional upon a 'Continuous' Cutpoint ($\hat{\pi}_c^S$)

In the next step we calculate a binary variable of group membership from the estimated model based probability, conditional upon the cutoff value $\hat{\pi}_c^S$. Any probabilities exceeding the cutoff value receive a value of 1. All others receive a value of 0. This step is repeated for a near-continuous number of cutoff values in small steps (say 0.01). The results can then be categorized in a classification table such as Table 2 for any specific cutoff value.

Table 2: Classification Table of the Logistic Regression Model at Cutoff Value $\hat{\pi}_c^S$ =0.5.

		Observed (True value)								
		20-30			*31-40*			*41+*		
		0	**1**	**Total**	**0**	**1**	**Total**	**0**	**1**	**Total**
Classified	**0**	911	32	943	918	36	954	3773	74	3847
	1	8	69	77	11	32	43	11	55	66
	Total	919	101	1020	929	68	997	3784	129	3913

Such tables measure forecast quality, as reflected in the calculation of sensitivity (the share of correctly forecasted non-members out of all true non-members), and specificity, (the correctly forecasted members as a share of all true members at any given cutoff value). They are conditional upon specific cutoff values.

Step 3: Assessment of the Forecast Quality ("goodness-of-fit") in the Source Survey

A well known indicator for the assessment of binary model forecasts is the ROC (Receiver Operating Characteristic) curve which juxtaposes sensitivity (truly identified positive response) with the percent of outcomes wrong positive responses (truly negative).

Table 3: Sensitivity and Specificity for the Mother's Age Group 20-30 at Cutoff Values from 0 – 1 by Increments of 0.05.

Cutpoint	Sensitivity	Specificity	1-Specificity
0.00	100.00%	0.00%	100.00%
0.10	87.02%	83.58%	16.42%
0.15	80.05%	90.24%	9.76%
0.20	76.20%	93.02%	6.98%
0.25	72.84%	94.36%	5.64%
0.30	72.12%	94.86%	5.14%
0.35	66.35%	97.15%	2.85%
0.40	64.18%	97.43%	2.57%
0.45	61.78%	97.71%	2.29%
0.50	58.89%	97.96%	2.04%
0.55	57.93%	98.06%	1.94%
0.60	53.13%	98.73%	1.27%
0.65	51.68%	98.80%	1.20%
0.70	49.52%	99.08%	0.92%
0.75	45.91%	99.26%	0.74%
0.80	40.38%	99.37%	0.63%
0.85	34.86%	99.47%	0.53%
0.90	28.37%	99.68%	0.32%
0.95	20.19%	99.72%	0.28%
1.00	0.00%	100.00%	0.00%

Figure 1: The ROC Curve

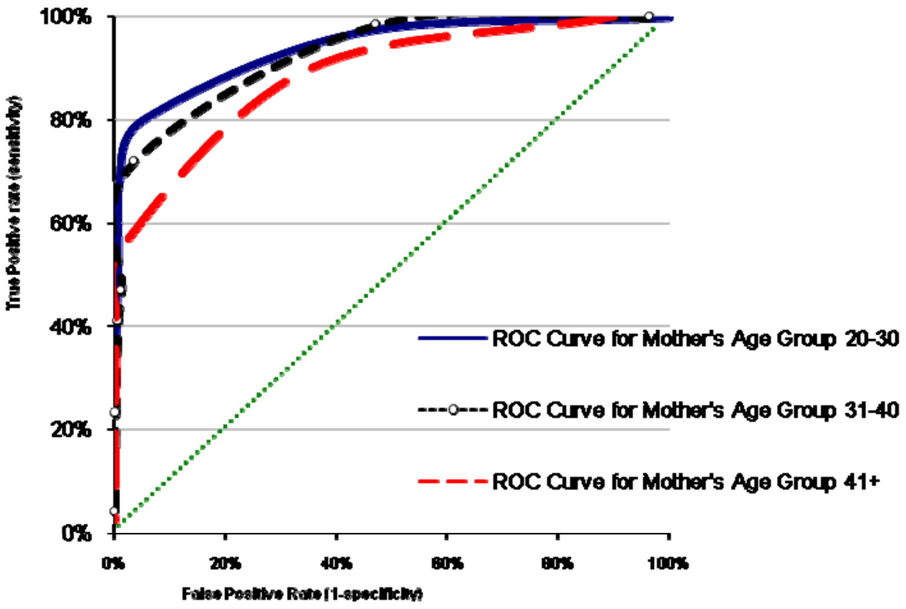


Figure 1 indicates that the youngest age group's regression performance is best among the 3 groups when evaluated by the integral below the curve. Estimates for all 3 groups are better than the 45o line of random assignment. Figure 1 emphasizes the importance of splitting up the model estimation according to the mother's age-group, thereby allowing for age-dependent parameter coefficients in the regression.

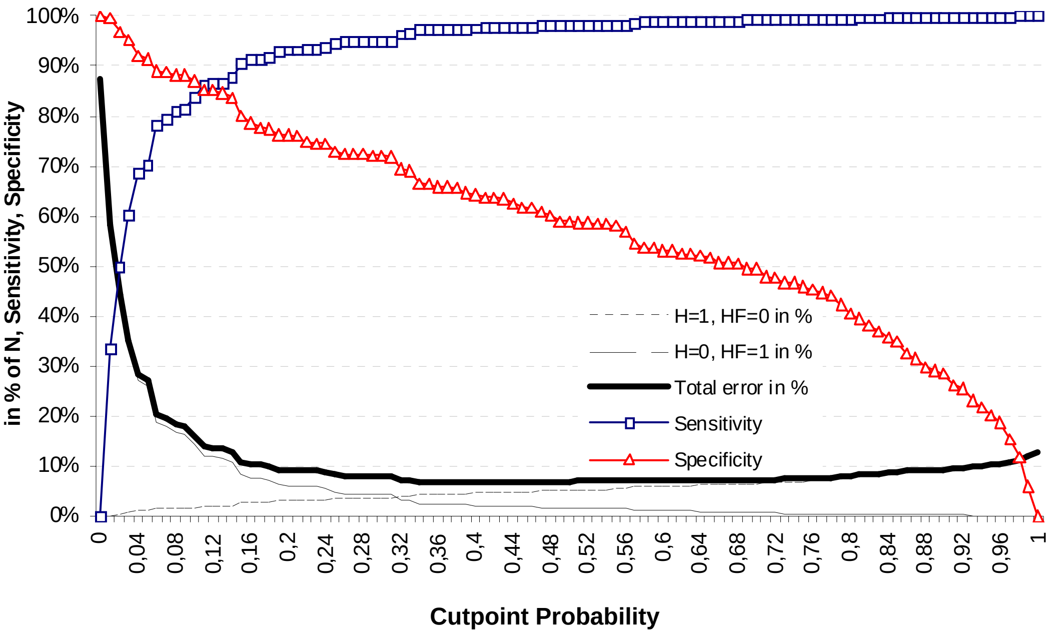
Step 4: Searching for the Optimal Probability Cutoff Value ($\hat{\pi}_c^{S,*}$)

After the ROC curves have been calculated, the optimal probability cutoff value needs to be located among all the possible cutoff values. The question how to translate logistic probabilities back into a binary variable is not conclusively dealt with in the literature. In medical research the question arises frequently and is sometimes related to the improvement or the damage in health caused by a specific treatment under review. For example, if the effect of a certain vaccine can only be known ex post and it is found to cause an important health improvement for some, while for others there is a negligible negative effect a general vaccination policy might be a reasonable course of action.

If, like in the present case, there is no a priori case for a particularly large net cost of either error (FP, FN in section 3, step 2) we opt for a cutoff value that minimizes the total sum of squared errors FP and FN.

Hosmer and Lemeshow (2000) suggest that the optimal cutoff value lies at the intersection of sensitivity and specificity. Their choice seems to hinge on the argument that we attach equal importance to each group in a relative sense. In figure 2 this cutoff value is at $\hat{\pi}_c^{S,*}$=0.11.[22] This probability level is surprisingly low, allowing for a great number of mistaken binary forecasts. The cutoff value according to the Minimum-squared-error-rule (MSE) is at $\hat{\pi}_c^{S,*}$=0.35, yielding more reliable forecasts of Haredi group membership.

Figure 2: The Optimal Cutoff Value



[22] See also Hosmer and Lemeshow (2000: 162).

The optimality rule for chosing the cutoff value crucially affects the number of forecasting errors. In Table 4 we compare the incidence of errors. While the MSE-rule reduces the number of FP cases significantly, the opposite occurs in the FN cases. This pattern repeats itself in all three age-groups. However, we also observe that the deterioration in FN is more than offset by the improvement in FP. This is again the case in all three age-groups, such that we can conclude that the MSE approach meaningfully improves the forecast, reducing the sum of errors to half compared to the HL approach.

Table 4: Forecasting Errors in Percent of the True Haredi in Each Age Group

Probability Cutpoint	Sum of Errors (FN+FP)	False negative (FN)	False Positive (FP)	True Positive (sensitivity)	Forecast
		H1 HF0	H0 HF1	H1 HF1	HF 1
		Age of Female Partner, 18-30			
Hosmer Lemeshow Model	78.6%	13.0%	65.6%	87.0%	199.0%
Minimum Squared Error Model	52.2%	33.7%	18.5%	67.3%	85.8%
Actual		-	-	100.0%	100.0%
		Age of Female Partner, 31-40			
Hosmer Lemeshow Model	78.1%	17%	110%	83%	193%
Minimum Squared Error Model	36.3%	49%	9%	51%	60%
Actual		-	-	100%	100%
		Age of Female Partner, 41+			
Hosmer Lemeshow Model	167.5%	39%	109%	61%	170%
Minimum Squared Error Model	82.5%	63%	9%	37%	46%
Actual		-	-	100%	100%
		Total			
Hosmer Lemeshow Model	324.3%	24%	93%	76%	186%
Minimum Squared Error Model	170.9%	49%	13%	51%	63%
Actual	-	-	-	100%	100%

Step 5: The BIT - Calculation of the Forecast $\hat{H}_i^T$ in the Target-Survey

In the final step we estimate the Haredi population $\hat{H}_i^T$ by use of the regressions in appendix Table 1. Table 5 illustrates the importance of the quality of the model to be used for forecasting group membership. In each of the observed years the downward bias of the Haredi population is much smaller in the recommended approach compared to the traditional approach as used in Berman and Klinov, 1997, or in Dahan, 1998, and elsewhere.

Table 5: Alternative Estimates of Haredi Population Size
thousands, percent*, based on data from 2002–2004)

	1997	1998	1999	2000	2001	2002	2003	2004
Source Survey (Soc.S)								
Based on the respondents' declaration	-	-	-	-	-	469,017	512,442	658,669
						0%	0%	0%
Optimal BIT	-	-	-	-	-	401,182	344,930	439,990
						-14%	-33%	-33%
Education based model						171,511	165,034	215,966
						-63%	-68%	-67%
Target Survey (HES)								
Optimal BIT	361,344	423,143	358,553	416,427	365,321	395,628	403,329	409,566
						-16%	-21%	-38%
Education based model	331,590	376,783	307,696	343,319	321,739	326,550	358,117	360,585
						-30%	-30%	-45%
Election based model (E)								
			525,000			565,000		
						20%		

*Percentages indicate deviations from the population size based on the respondents' declarations

Source: Social Survey, Household Expenditure Survey, Election results, Central Bureau of Statistics.

(

5 Conclusions

This methodology can be usefully applied to many fields, since it allows us to optimally enhance a given data base by adding a binary variable that does not exist in the relevant data base. The present study shows how our methodology can be used for poverty estimations of a population subgroup that is not sampled in the major surveys used for poverty calculations – the income survey or the expenditure survey.

The main results are reported in Table 6. Poverty incidence in the Haredi population is nearly three times higher than in the general population. It is among the poorest population groups in Israel, making it obviously highly necessary to monitor efforts of poverty reduction. Inequality among the poor, though it is lower than in the general population, does not compensate for the higher poverty incidence and income gap. Poverty intensity, as measured by the Sen-index is almost double its level for the total population. A similar conclusion can be drawn concerning child poverty.

Table 6: Relative Poverty in Israel among the Haredi and the Total Population

	1/2 Median Equivalized Poverty	
	Haredi Population	Total Population
Headcount	60%	23%
Gini Index of the Poor	0,158	0,333
Income Gap	31,6%	33,3%
Sen Poverty Measure	0,255	0,130
Child Poverty Headcount	63%	33%
Haredi Population Size	409 566	6 274 115

Source: Expenditure Survey, 2004, C.B.S

While until recently Haredi poverty has been approximated only roughly, the present methodology improves the accuracy of poverty measurement for the targeted group, a desirable feature, the more expensive and the longer the time lag of policy implementation.[1]

The proposed method may also be usefully applied in the context of poverty mapping by Small Area Estimation (see for example Hentschel et al., 2000). In such an exercise we might be interested in attaching a binary forecast of poverty incidence to a household in a small area, not covered by the household surveys typically used for official poverty calculations (the Source Survey, *S*). Data on households in the small area, collected in an extensive but superficial large scale survey such as a survey accompanying a population census could be used as the vector x' in a Target Survey, *T*, as outlined in sections 2 and 3. Use of the coincidental vector x' in *S* and *T* and the choice of an optimal cutoff value would allow for the production of an estimate of poverty incidence in *T*.

References

Berman, Eli and Ruth Klinov (1997). Human Capital Investment and Nonparticipation: Evidence from a Sample with Infinite Horizons (Or: Mr. Jewish Father Stops Going to Work), Jerusalem, The Maurice Falk Institute for Economic Research in Israel, Discussion Paper (97.05), 1-36.

Berman, Eli (2000). Sect, Subsidy, and Sacrifice: An Economist's View of Ultra-Orthodox Jews, *The Quarterly Journal of Economics*, 904-952.

Bigman, D., and P.V. Srinivasan (2002). Geographical Targeting of Poverty Alleviation Programs: Methodology and Applications in Rural India. *Journal of Policy Modeling,* 24, 237-255.

Dahan, M. (1998). The Ultra-Orthodox Jews and Municipal Authority, Part 1 – Income Distribution in Jerusalem, in Hebrew, The Jerusalem Institute for Israel Studies, *Research Series*, 79, 1-50.

Degani, Avi and Rina Degani (2000). The Demand for Housing in the Haredi Sector, Institute for Spatial Analysis Ltd., September, 1-170.

Flug, Karnit and Nitsa (Kaliner) Kasir (2003). Poverty and Employment and the Gulf between Them, *Israel Economic Review,* Vol. 1, p. 55-80.

Frenkel, Alona, Pavel Soyfer and Yoram Mayshar (2003). Potential Income as a Measure of Poverty in Israel, Working Paper, Maurice Falk Institute, (in Hebrew), 1-30.

Friedman, Menachem (1991). The Haredi (Ultra-Orthodox) Society – Sources, Trends and Processes, in Hebrew, Summary in English, The Jerusalem Institute for Israel Studies, Jerusalem.

Glewwe, Paul and Jacques Van der Gaag (1990). Idenifying the Poor in Developing Countries: Do Different Definitions Matter? *World Development*, 18 (6), 803-815.

[1] See Glewwe and Van der Gaag (1990).

Gottlieb, Daniel and Nitsa Kasir (2004). Poverty in Israel and a Strategy for its Reduction, in Hebrew, The Bank of Israel, www.bankisrael.gov.il, 1-46.

Gottlieb, Daniel and Roy Manor (2005). On the Choice of a Poverty Measure: The Case of Israel, 1997 to 2002, in Hebrew, Abstract in English, forthcoming, The Bank of Israel, 1-54.

Gottlieb, Daniel (2007). Poverty and Labor Market Behavior in the Ultra-Orthodox Population in Israel, (in Hebrew), *Economics and Society Program*, The Van Leer Institute, Jerusalem, 1-55.

Gurovich, Norma and Eilat Cohen-Kastro (2004). Ultra-Orthodox Jews – Geographic Distribution and Demographic, Social and Economic Characteristics, 1996-2001, in Hebrew, Summary in English, *Working Paper Series, No. 5*, Central Bureau of Statistics – Demography Sector.

Hadjicostas, Petros and George C. Hadjinicola (2001). The Asymptotic Distribution of the Proportion of Correct Classifications for a Holdout Sample in Logistic Regression, Journal of Statistical Planning and Inference, Vol. 92 (1-2), January, 193-211.

Hentschel, J., J.O. Lanjouw, P. Lanjouw and J. Poggi (2000). Combining Survey Data to Trace the Spatial Dimensions of Poverty: A Case Study of Ecuador, *World Bank Economic Review*, 14, (1), 147-65.

Hosmer, David W. and Stanley Lemeshow (2000). *Applied Logistic Regression.* 2nd Edition, New York: John Wiley & Sons Inc.

Schares, G., et al. (2003). Regional Distribution of Bovine Neospora Caninum Infection in the German State of Rhineland-Palatinate Modeled by Logistic Regression, *International Journal of Parasitology,* Vol 33 (14), 1631-1640.

Schutter, E.M.J. et al. (1998). Estimation of Probability of Malignancy Using a Logistic Model Combining Physical Examination, Ultrasound, Serum CA 125, and Serum CA 72-4 in Postmenopausal Women with a Pelvic Mass: An International Multicenter Study", *Gynecologic Oncology*, Vol. 69, (1), 56-63.

Stegeman, J.A. et al. (2006). Establishing the Change in Antibiotic Resistance of Enterococcus Faecium Strains Isolated from Duch Broilers by Logistic Regression and Survival Analysis. *Preventive Veterinary Medicine*, 74 (1): 56–66.

Tal Commission (2000). Report on the arrangement concerning the recruitment of Yeshiva students to the IDF (in Hebrew), July,

Appendix Table 1: Logistic Regression for Haredi Affiliation

		20-30		31-40		41+
Variable	Coefficient	Prob.	Coefficient	Prob.	Coefficient	Prob.
C	-3.390632	0.00000	-4.936651	0.00000	-4.910893	0.00000
AC15_2	1.899704	0.10270	-	-	2.638771	0.04140
LSY	4.033626	0.00000	1.689088	0.05200	4.025401	0.00000
DIST_11	0.967045	0.01930	1.320375	0.00270	1.836252	0.00000
DIST_51	-	-	0.58212	0.20220	0.87233	0.00140
PHILANT	0.582083	0.16180	1.92117	0.00000	1.191979	0.00000
IL_HH_m	0.853245	0.01480	1.298233	0.00220	0.098535	0.80580
CH_ROOM	2.368207	0.00000	2.00989	0.00000	1.807127	0.00000
CAR	-2.20409	0.00000	-0.795949	0.03270	-0.816382	0.00100
NO_DIPL	2.143235	0.00000	1.499992	0.00160	1.885641	0.00000
INTERNET	-1.131852	0.02510	-1.812845	0.00060	-1.791576	0.00000

The Variable List:

AC15_2 - Binary variable indicating ratio between number of children in household and the age of the mother at values 0.15-0.2

LSY - Binary variable indicating, that the last school attended by any of the male members of the household was a religious seminar (Yeshiva).

DIST_11 - Binary variable indicating that the household was sampled from the Jerusalem district (1,0).

DIST_51 - Binary variable indicating that the household was sampled from the Tel-Aviv district (1,0).

PHILANT - Binary variable indicating philanthropic activity by head of household (1,0).

IL_HH_m - Binary variable indicating country of birth of household head as Israel (1,0).

CH_ROOM - The number of children divided by the number of rooms, in the household.

CAR - Binary variable indicating car ownership (1,0).

NO_DIPL - Binary variable indicating that head of household never got any school/university diploma (1,0).

INTERNET - Binary variable indicating household's use of internet (1,0).