AN OPTIMAL METHOD OF BINARY INFORMATION TRANSFER (BIT) BETWEEN SURVEYS OF AN IDENTICAL POPULATION

Daniel Gottlieb and Leonid Kushnir

Discussion Paper No. 06-06

August 2006

Monaster Center for Economic Research
Ben-Gurion University of the Negev
P.O. Box 653
Beer Sheva, Israel

Fax: 972-8-6472941
Tel: 972-8-6472286

An Optimal Method of Binary Information Transfer (BIT) Between Surveys of an Identical Population:

Poverty among the Jewish Ultra-Orthodox in Israel – A Case Study of BIT

by

Daniel Gottlieb and Leonid Kushnir[1]

Abstract

This paper deals with the optimal transfer of information on group identification between different data sets of an identical population. Such a need might arise frequently in the analysis of socio-economic surveys and in the implementation of social and economic policy. Due to the limited number of questions asked in any given survey, analysts of one survey may need some binary information available in another survey of the same population. We suggest an efficient method for transferring such binary information between the survey in which the information is available (the "source-survey") and the survey in which the information is needed (the "target-survey"). We show that an optimal method for transferring the information depends crucially on two aspects of the process: (1) an efficiently estimated statistical model of of group-membership (the goodness of fit of the ROC-curve) and (2) the determination of the optimal cutoff value (the value chosen for turning the logistic probability forecast back into a binary variable). The proposed method can be useful in the social sciences, in medical research and in any field that requires probabilistic binary data enrichment between data sets drawn from a common population. It is an efficient method for ex-post enhancement of given data sets, when such enhancement by use of other methods is either expensive or impossible.

In this paper we perform a BIT on membership in the Jewish Ultra-orthodox community in Israel, known as an extremely impoverished group. Since the information on religious group membership exists only in the "Social Survey" and information on poverty exists only in the Income Survey or the Expenditure Survey, there arises a need for optimal BIT. We present the results of such an optimization and its importance for social policy.

Key Words: Poverty**,** Targeting, ROC curves, Binary Variables, Group Identification, Poverty research, Social Policy Planning

JEL Classification Numbers: C15, D63, I38, Z12

Jerusalem, July 2006

[1] Daniel Gottlieb, Bank of Israel and Ben-Gurion University and Leonid Kushnir, Bank of Israel. Of course, we bear the responsibility for any remaining mistakes.

1. Introduction

The major social and economic surveys of an economy focus typically on different aspects of the same population sample. 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, some of the survey-specific information might be useful to researchers of another survey and to policy makers using the combined information. In this paper we suggest an efficient method for optimal binary information transfer (BIT) from a source-survey to one or more target-surveys by means of econometric analysis. The method allows for a quality test within the source survey. The optimal method for transferring the information depends crucially on two aspects of the process: (1) a sufficient overlap of variables that contain explanatory power of the variable to be transferred (thus ensuring a reasonable goodness of fit of the ROC-curve) and (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. 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. The proposed method improves that of Hosmer and Lemeshow (2000).

BIT has many possible applications. It can be useful in the targeting of policies to specific population groups, in poverty mapping or in the analysis of simulation results.

We illustrate the method by applying it 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 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.[2] Concluding remarks complete the paper.

[2] A detailed analysis of poverty in the Israeli Haredi

2. The Model

Assume a sampling of two Household surveys, one which we call the Source-survey (S), consisting of n_S = 1…S households, and another survey sampled on the same population[3], which we call the Target survey (T), n_T = 1…T. Let there be a binary group variable, say of group H, which receives the value of 1 if the household is a member of H and 0 if not. 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, P($\hat{H}$=1| **x'**) = $\hat{\pi}$(**x**) where vector $x' = (x_1, x_2, ...x_k)$. $\hat{H}_i^{S,T}$ 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

$$\pi_{ij}(x) = \frac{e^{g_i(x)}}{1+e^{g_i(x)}} . \qquad (1)$$

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

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

3. The BIT process

Step 1: Search for a logistic regression in S with a high (statistical) explanatory power

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 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 for group membership, conditional upon a 'continuous' cutpoint ($\hat{\pi}_c^S$)

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

[3] 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.

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.[4]

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] Gottlieb (2006b) presents a general model of optimal choice of a cutoff value. The optimal outcome is shown to depend on a minimization of the expected net costs caused by each of the two types of errors – FP and FN.

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

The Israeli Ultra-Orthodox Jewish society, also called Haredi society,[5] 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.

The Haredi society is an interesting example for BIT application since it is an idiosyncratic population group with distinct cultural features. That has important implications for public policy, concerning aspects of anti-poverty policy as well as others. When the population is highly heterogeneous it is therefore crucial to be able to break down the statistics into sub-groups. To some extent this could of course be done fairly accurately by use of GIS data on the specific addresses of the people involved in the sample. While such information could be usefully integrated into the BIT approach outlined here, it is usually not provided to the researcher due to the right to privacy.

4.1 The Israeli Haredi Society – Roots and Characteristics

The Israeli Haredi society consists of three more or less independent movements, each emphasizing different aspects of Judaism and obeying its own spiritual leaders: The Hassidic, the Lita'i and the Sephardic[6] groups. They all share a 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 itself maintains a strong sense of hierarchy and it issues and looks over detailed rules for individual and family behavior through its legal and semi-legal organs part of which are recognized by the state.

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. The majority of Haredi women and men in drafting age do not serve in the Israeli army, due to a special exemption from army service, issued by the state. This exemption dates from a historical compromise, struck in 1948 between David Ben-Gurion (then Prime Minister) and

[5] "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.

[6] 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.

Hazon Ish[7] (then Leader of the Haredi society) and was at the time relevant for some 400 male religious scholars, whose main occupation was the Torah study. Over the years their number grew at an average annual rate of 8 to 9 percent, reaching more than 30,000 in the early 21st millennium. In response to court appeals and a general public discontent concerning the exemption from the army of Haredi Students at Jewish Seminars (Yeshivas) on the one hand and a mounting social problem due to an increasing number of young drop-outs from the religious seminars, an official commission was appointed in 1999 with the purpose of proposing a change in government policy concerning the exemption of Haredi men from military service. In order to improve the Haredi men's participation in the labor market the commission suggested a "year of decision", during which Yeshiva students at the age of 23 would be allowed to leave the Yeshiva in order to work or enroll in job training, without losing their right of exemption from army service. Thereafter they would have to decide about returning to the Yeshiva, join the labor force or the army for a short service or else volunteer in a civilian task.[8]

4.2 Existing Estimates of the Israeli Haredi Population Size

Estimating the size of the Haredi population is not trivial, since information on Haredi group membership has until recently not been included in the standard surveys of the Israeli Central Bureau of Statistics (henceforth ICBS). Earlier attempts to estimate the size of the Haredi population were based on the survey 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). This approach was further elaborated in Berman (2000). It has been used by additional authors for analyzing Haredi poverty and labor market behavior.[9] Another approach used to estimate the size of the Haredi society has been to analyze election results by utilizing the monolithic Haredi voting-pattern. Such studies were undertaken by Degani and Degani, 2000 (henceforth DD), and more recently by Gurovich and Cohen, 2004 (henceforth GC).

[7] 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.

[8] See the report of the Tal Commission, 2000.

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

4.2.1 The Education-based approach

This approach was developed in a paper by attributes a household to the Haredi population if at least one of its male members indicates a Yeshiva (a religious seminary)[10] 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. There are various reasons why a forecast based on male Yeshiva attendance might produce unsatisfactory forecasts: Yeshiva studies do not constitute a necessary condition for being Haredi. Indeed, Yeshiva attendance among Hassidic Jews, which are the largest sect within Haredi society, is much lower than in the Lita'i and Sephardic Haredi groups.

4.2.2 The Elections-based Approach

This approach was chosen by GC, based on the 2003 elections[11] 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"[12]. While the voters for UTJ are supposedly mainly Haredi, many "Shas" supporters may be less religious traditional or ethnically oriented non-religious Sephardic voters. Therefore GC had to rank only part of the Shas supporters should be counted as Haredi. In order to identify this subgroup, GC included Shas supporters among the Haredi only if they lived in the vicinity of areas in which there was a high percent of UTJ support, reasonably assuming that the 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:

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

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

[10] 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.

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

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

general public.[13] Based on fertility rates derived from the Social survey for Haredi women of Ashkenasi[14] background, GC used a fertility rate 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[15]. 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 the population size by use of the following formula[16]:

$$\mathrm{H}_{\mathrm{Pop}} = \sum_{i} \mathrm{nn}_{i} + \mathrm{under}\,20_{i} \times \frac{\mathrm{nn}_{i}}{\mathrm{over}\,20_{i}}$$

where $\mathrm{H}_{\mathrm{Pop}}$ = 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 ith (Haredi) household and over20_i = number of additional people (to those questioned) aged 20 or more, in that household. According to this calculation, the Haredi population reached about 550,000 by end of 2002.[17]

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.

[13] 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.

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

[15] 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.

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

[17] 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.

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 S

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.

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 Based on the Logistic Regression Model at Cutpoint $\hat{\pi}_c^S = 0.5$.

Classified		**Observed (True value)**								
		20-30			***31-40***			***41+***		
		0	**1**	**Total**	**0**	**1**	**Total**	**0**	**1**	**Total**
	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 allow for the calculation of the forecast quality indicators called Sensitivity and Specificity. They are conditional upon specific cutoff values. Sensitivity is defined as the correctly forecasted non-members as a share of all true non-members. Specificity is defined as the correctly forecasted members (nonmembers) as a share of all true members at any given cutoff value.

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%

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).

Figure 1: The ROC Curve

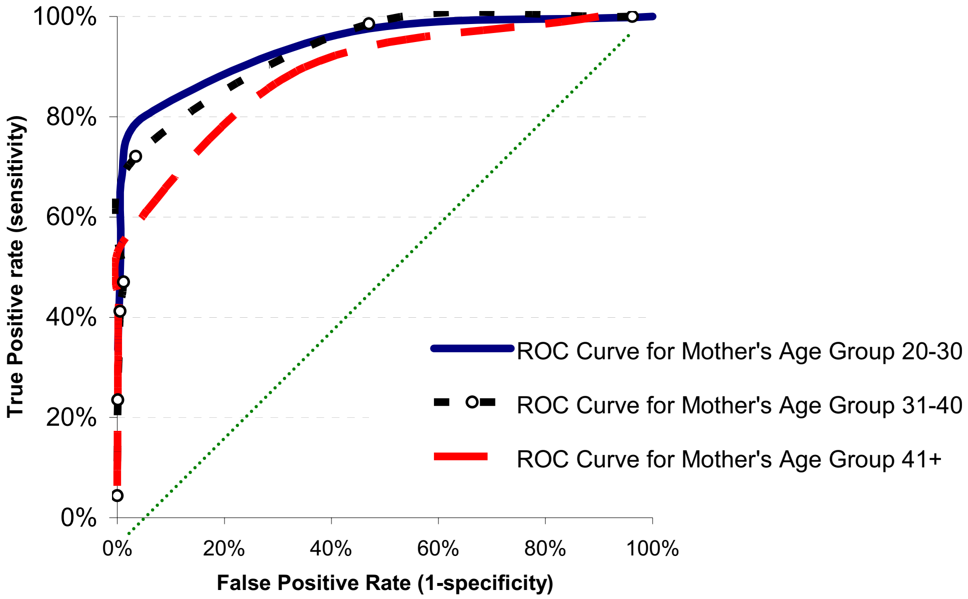


Figure 1 indicates that the youngest age group's regression performance is best among the 3 groups when evaluated by the integral below the curve. Estimates for all 3 groups are better than the 45° line of random assignment. Figure 1 emphasizes the importance of splitting up the model estimation according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

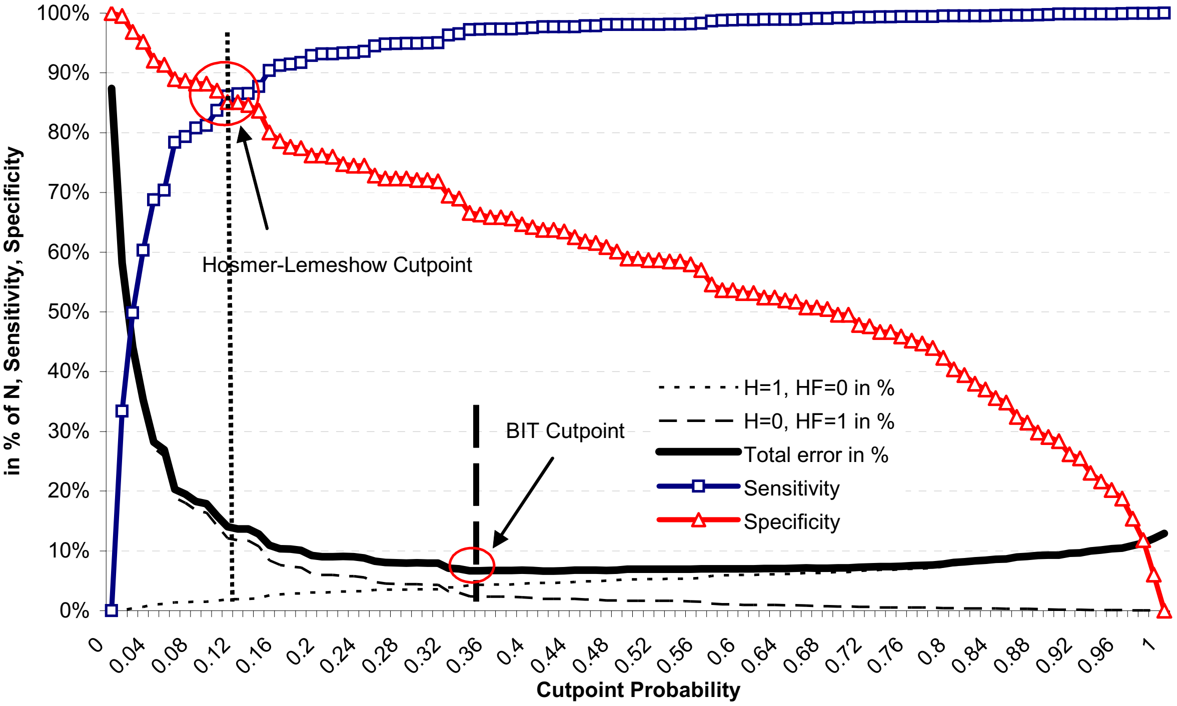
Step 4: Searching for the Optimal Probability Cutoff Value ($\hat{\pi}_c^{S,*}$)

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.[18] 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



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.

[18] See also Hosmer and Lemeshow, p. 162.

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 (specificity)**	**Forecast**
		H1 HF0	**H0 HF1**	**H1 HF1**	**HF 1**
		Age of Female Partner, 18-30			
Hosmer Lemeshow Model	0.0%	0.0%	0.0%	0.0%	0.0%
Minimum Squared Error Model	0.0%	0.0%	0.0%	0.0%	0.0%
Actual		-	-	100.0%	100.0%
		Age of Female Partner, 31-40			
Hosmer Lemeshow Model	0.0%	0%	0%	0%	0%
Minimum Squared Error Model	0.0%	0%	0%	0%	0%
Actual		-	-	100%	100%
		Age of Female Partner, 41+			
Hosmer Lemeshow Model	0.0%	0%	0%	0%	0%
Minimum Squared Error Model	0.0%	0%	0%	0%	0%
Actual		-	-	100%	100%
		Total			
Hosmer Lemeshow Model	0.0%	0%	0%	0%	0%
Minimum Squared Error Model	0.0%	0%	0%	0%	0%
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)

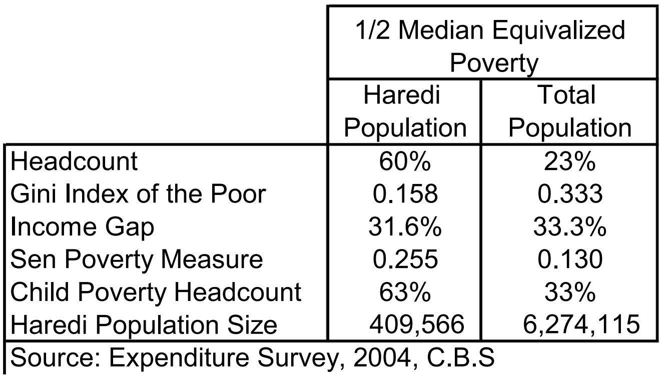
	1997	1998	1999	2000	2001	2002	2003	2004
Source Survey (Soc.S)								
Based on the respondents' declaration	-	-	-	-	-	469,017 0%	512,442 0%	658,669 0%
Optimal BIT	-	-	-	-	-	401,182 -14%	344,930 -33%	439,990 -33%
Education based model						171,511 -63%	165,034 -68%	215,966 -67%
Target Survey (HES)			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

*Percentages indicate deviations from the population size based on the respondents' declarations
Source: Social Survey (Soc.S), Household Expenditure Survey HES, Election results (E) Israeli Central Bureau of Statistics.

4.3 Possible Uses of the BIT Process

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[19]



	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

The proposed method may thus be used effectively to improve the targeting of poverty policy, and the efficiency of allocating funds for poverty alleviation. Furthermore it allows for monitoring progress in policy implementation. While until recently poverty estimates of this population group could only be roughly approximated, the present methodology improves the accuracy of poverty measurement for this group. Such improved estimates are crucial the more expensive the policy measures are and the longer the implementation lag of the policy. Consequently the method is useful in reducing waste of public funds in the pursuit of poverty reduction.

[19] Poverty in the Haredi population, based on an earlier version of the BIT-methodology is analyzed in detail in Gottlieb (2006).

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 No. 97.05., p. 1-36.

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

Dahan Momi, 1998, "The Ultra-Orthodox Jews and Municipal Authority, Part 1 – Income Distribution in Jerusalem", in Hebrew, The Jerusalem Institute for Israel Studies, Research Series No. 79, Jerusalem, p. 1-50.

Degani Avi and Rina Degani, 2000, "The Demand for Housing in the Haredi Sector", Institute for Spatial Analysis Ltd., September, p. 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.

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

Gottlieb Daniel, 2006, "Poor and Trapped in Ideology: A Case-Study of Poverty in the Jewish Ultra-Orthodox Society in Israel", (in Hebrew) forthcoming, Van Leer Institute, Jerusalem, 1-44.

Gottlieb Daniel 2006b, "On the Optimal Cutoff Value in Logistic Probability Models", forthcoming.

Gottlieb Daniel and Nitsa Kasir, 2004, "Poverty in Israel and a Strategy Proposal for its Reduction", in Hebrew, The Bank of Israel, www.bankisrael.gov.il, July, 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, March, 1-54.

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, July, Central Bureau of Statistics – Demography Sector.

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

Liviatan Oded, 2003, "The Effect of "Redistribution" on Poverty Incidence and Intensity," (Hebrew) November, Discussion Paper, Research Department, Bank of Israel, p. 1-49.

Mayshar Yoram, 2004, "Potential Income as a Measure of Poverty in Israel", forthcoming in the "Israel Economic Review", (in Hebrew), p. 1-30.

National Insurance Institute, 2000, "Annual Report", Jerusalem, Israel

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).