Session Number: Parallel Session 4D
Time: Tuesday, August 24, PM

Paper Prepared for the 31st General Conference of
The International Association for Research in Income and Wealth

St. Gallen, Switzerland, August 22-28, 2010

Who Does not Respond in the Social Survey: an Exercise in OLS and Gini regressions

Yolanda Golan and Shlomo Yitzhaki

For additional information please contact:

Name: Shlomo Yitzhaki
Affiliation: Central Bureau of Statistics, Israel

Email Address: shlomo.yitzhaki@huji.ac.il

Preliminary Draft
Comments Invited

Who Does not Respond in the Social Survey: an Exercise in OLS and Gini regressions

By

Yolanda Golan* and Shlomo Yitzhaki**

Central Bureau of Statistics
Jerusalem, Israel

Abstract

The main purpose of this paper is to analyze patterns of non-response in the social survey and to evaluate its effect on potential biases on satisfaction from life. An additional purpose is to apply the method of mixed regression, which combines the method of Ordinary Least Squares with Gini regression in the same estimation procedure in order to ensure that the conclusions reached do not depend on the regression methodology. The main conclusion is that young persons and ultra religious groups tend to have a lower participation in the survey and a high satisfaction from life. This in turn tends to bias satisfaction from life downward.

Keywords: non-response, Gini, OLS, satisfaction
JEL Classification: C39, C80

* Central Bureau of Statistics, Email address: yolandag@cbs.gov.il
** Corresponding author. Central Bureau of Statistics and Hebrew University.
Email address: yitzhaki@cbs.gov.il

1. Introduction

Social surveys, which include questions about subjective well-being, are making their way into the main stream of official statistics.[1] Recently, the Stiglitz' commission (2009) has recommended to augment summary statistics that include data on income distribution and subjective well-being into the traditional national-income accounts. In its early development, this kind of surveys were conducted by private or not for profit institutions, but as it is making its way to official statistics, we should expect the formation of some international guidelines, to increase the comparability of subjective surveys conducted by official bureau of statistics. Also, we should expect an increase in methodologies and data that are available only for the use of national statistical agencies such as administrative or census registrars.

Surveys that are dealing with subjective issues (hereafter S surveys) are different from the regular households' surveys. The first difference is that they have to be conducted at the personal and not at the household level. The second difference is related to the subject matter. Unlike surveys that collect factual data, S surveys also collect data on feelings and opinions. It is reasonable to assume that there will be some individuals that will be sensitive to report their opinion, especially if the questionnaire is conducted by a government official. Privacy concerns and the fear from an intrusive government seem to be among the factors that are contributing to the increase in non-response that were observed in many western countries. (De Leeuw and De Heer, 2002). Non response may be more severe among minorities and excluded groups (Feskens et al. 2007).[2]

The major statistical problem with non-response is that if the non-response is not random, then it may cause the estimates to be systematically biased, so that one does not get the true values and the true changes in values in the target population. Other issues are concerned with increasing costs and frustration on behalf of interviewers.

Social surveys, which concentrate on subjective feelings, may seem more intrusive than surveys that are concerned with solid facts that seem objective and known not only to the interviewed.

The purpose of this paper is to investigate patterns of non-response in the social survey which is conducted by the Israeli Central Bureau of Statistics. The survey is

[1] See Helliwell (2010) for a recent survey and the OECD conference on that subject.

[2] A third difference is concerned with the reliability of reports on subjective issues. See Schimmack *et. al.*(2009) for a recent contribution on this issue.

conducted each year since 2002, and it is in the field for a year. The sample is drawn from the population registrar. We refer to the registrar as the sampling framework. This is done several months prior to the interviewing stage, which is conducted by a face to face interview, using Computer Assisted Personal Interviews (CAPI).

In general, there are two ways of investigating patterns of non-response: One way is to analyze the characteristics of those who do not respond. We will refer to this method as the direct way of investigation. The alternative way is to rely on the process that is conducted by Statistical Bureaus in order to decrease random perturbations of the estimates and to correct for biases caused by non-response. This process is based of creating a weighting scheme attached to each observation so that each demographic group in the population is represented according to its weight in the population. By investigating the weighting scheme, one can learn about non-response, because the bigger the weight attached to an observation, the less its characteristic is represented in the sample. We will refer to this way of investigation as the indirect way, because one investigates non-response from the characteristics of those who responded.

Both methods are not perfect and each one has its own drawbacks: The direct way may suffer from errors in the population registrar and in the classification of the reasons for non-response. For example, the population registrar includes individuals that may be outside the country. Failing to contact the person does not distinguish between a person who does not respond because he avoids any connection with the interviewer, or because the person is outside the country for a long period of time. The major advantage of the indirect way is that the sample size of the respondents is bigger than the sample size of the non-respondents, and it includes more variables. Also, it is conducted after the interviewing stage is completed, so that it overcomes the lag in updating the population registrar. Therefore, in this paper we will rely on both methods.

The structure of the paper is the following: Section 2 presents the data and the results of the direct way, Section 3 presents the indirect way, while Section 4 presents the methodology used – the mixed Ordinary Least Squares and Gini regression method. Section 5 presents the results. Section 6 searches for a reasonable explanation, while Section 7 concludes.

2. The data

The Social Survey is conducted by the Israeli Central Bureau of Statistics (hereafter ICBS) since 2002. It comprises of a basic questionnaire that is asked every year, and an additional topic, to be conducted in a sporadic way. The Statistical Ordinance makes the response to the questionnaire mandatory. However, no person was prosecuted if he or she refuses to participate.[3] Since, non-respondents make a small portion of the sample, and the data on non-respondents is limited, we rely on two sources of data in order to analyze the implication of non-response.

The sample is drawn from the population registrar, about six months prior to the year in which the survey is conducted. The population registrar includes all the population of Israel. However, according to rough estimates, about 10 % of the population in the registrar is not living in the country. Based on other official records, like social security records, the population registrar is improved by the ICBS prior to the sampling but it is clear that the sampling framework is contaminated by records of individuals who do not belong to the target population of the survey. Hence, relying on the sampling framework may produce biased estimates of non-response. The population registrar includes demographic data only. For the purpose of this investigation, we have added to the registrar the earned income reported to the tax authorities. The earned income added is the earned income of the individual and it does not include income from capital nor government transfers from the National Insurance Institute.

Table 2.1 describes the field reports accumulated over the period 2004-2008. Overall, about 22 % of the individuals that were selected for the sample were not interviewed. However, one has to differentiate between those who were not supposed to be interviewed because of errors in the framework or administrative reasons and those that refused to be interviewed or the interview was not conducted because of other reasons. As can be seen from the Table, the failure to interview is higher among the immigrants, the elderly, the non-working population, and slightly higher among males, and the young. Comparison with tax data enabled us to estimate the participation rates and average earned income according to labor market type of employment. It can be seen that employees and self-employed are represented more among the participants than among the non-participants. However, the patterns are

[3] Romanov and Nir (2010) present an excellent review of the considerations in handling non-response in the ICBS.

different: among the employees the participants have a higher average income while among the self-employed we observe an opposite pattern. In general, it seems that the major difference between respondents and non-respondents is in participation in the labor market.

Table 2.1: The characteristics of Respondents and Non-respondents – 2004-2008

		Respondents	Non-Respondents
Total	**Obs.**	29,774	8,187
		78.4%	21.6%
Sex	**Males**	48.4%	52.0%
	Females	51.6%	48.0%
Age	**20-24**	11.9%	13.0%
	25-44	41.7%	39.7%
	45-64	30.2%	22.6%
	65+	16.2%	24.7%
	Average	45.1	47.9
Population Group	**Jews**	81.9%	81.1%
	Others	18.1%	18.9%
Immigrants 1990+		14.2%	17.2%
% Employees		56.4%	35.2%
Average Earned Income (New Shekel, monthly)		7,290	5,953
% Self-Employed		7.2%	3.6%
Average Earned Income (New Shekels, Monthly)		5,623	5,857
% Not working		36.4%	61.2%

Table 2.2 presents the reasons recorded by the interviewer for failing to interview. The observations are divided to two ethnic groups: Jews and others, mainly, Moslems. The reason for separating the groups is because Schechtman *et. al* (2008) have found in the Household's Expenditure Survey that the non-Jewish population tends to have a significantly higher response rate than the Jewish one, and the effect of an increase in income on response rate tend to be with different sign.[4]

Non-response is classified into three categories: temporary reasons include being absent from home, failure to find an appropriate time, persons who are outside the

[4] There can be several hypothetical explanations that may explain this result but there is no way to verify them. The range of the different explanations covers objective differences like the fact that the Arab population is less mobile than the Jewish one, or other differences such as cultural differences, different attitude toward a representative of the government. Data limitations do not allow verifying which explanation is appropriate.

country for more than a year and those who passed away;[5] Permanent reasons include refusal, language difficulty, and being in an institution; Administrative reasons include failure to find the person, located but weren't surveyed and does not belong to the population.

In both ethnic groups, non-response tends to be higher among the young and the elderly. Among the Jewish population, non-response tends to also be higher among males. We observe among the Jewish population, for the employees (self-employed) group, the higher the income, the higher (lower) the response rate. Among the non-Jewish population we also observe that both employees and self-employed with relatively high income tend to participate in the sample.

Table 2.2: Non-Participation according to personal characteristics and administrative classification

(a) Jewish

	Total		Sex		Age				Average Earned Income (New Shekels, Monthly)	
	Obs.		Males	Females	20-24	25-44	45-64	65+	Employees	Self-Employed
Respondents	**24,390**		**48.2%**	**51.8%**	**11.4%**	**38.8%**	**32.2%**	**17.7%**	**7,780**	**5,945**
Not Responded	**6,630**	**100%**	**52.5%**	**47.5%**	**11.5%**	**37.3%**	**23.4%**	**27.9%**	**6,412**	**6,327**
Type of Non-Response classification:										
Temporary	1,272	19.2%	62.6%	37.4%	22.9%	45.2%	23.6%	8.3%	7,067	7,858
Permanent	2,301	34.7%	43.8%	56.2%	6.9%	28.8%	25.4%	38.9%	6,856	6,322
Administrative	3,057	46.1%	54.8%	45.2%	10.0%	40.4%	21.8%	27.8%	5,453	4,438

(b) Non-Jewish

	Total		Sex		Age				Average Earned Income (New Shekels, Monthly	
	Obs.		Males	Females	20-24	25-44	45-64	65+	Employees	Self-Employed
Respondents	**5,384**		**49.0%**	**51.0%**	**13.7%**	**52.6%**	**25.3%**	**8.5%**	**4,361**	**3,508**
Not Responded	**1,557**	**100%**	**49.7%**	**50.3%**	**16.4%**	**51.4%**	**20.7%**	**11.5%**	**3,897**	**3,050**
Type of Non-Response classification:										
Temporary	354	22.7%	54.5%	45.5%	24.9%	54.8%	16.4%	4.0%	3,901	3,639
Permanent	355	22.8%	41.1%	58.9%	9.3%	47.6%	23.9%	19.2%	4,431	2,888
Administrative	848	54.5%	51.4%	48.6%	15.9%	51.6%	21.1%	11.5%	3,597	2,515

When one compares the reasons for non-response it seems that administrative reasons tend to be recorded more for the non-Jewish population than for the Jewish one, while

[5] Including cases of persons who passed away in the category of temporary reasons should be interpreted as assuming that there is a lag in updating the population registrar.

among the Jews permanent reason tend to be of a higher proportion. However, since this classification is done by different interviewers and refer to different populations one cannot tell the reason for the differences.

To sum up: the direct way of investigation reveals that the tendency not to respond is higher among the young, the elderly, and those who do not participate in the labor market. We do not observe, as (Feskens et al. 2007) found, that minorities tend to have a lower participation rate nor as Schechtman *et. al*. (2008) has found that minorities tend to have a higher response rate.

3. The Indirect way of Analyzing Non-Response

The indirect way of analyzing the effect of non-response is to use the sample of the respondents and the weighting scheme in order to analyze the effect of non-response. The advantages of this method over the direct way are the following: the weighting scheme is based on an updated framework. That is, while the sample is drawn about six month prior to the interviewing stage, the weights are derived after the interviewing stage is completed, and therefore the framework used is an updated one. The second advantage is that one can use both the variables in the framework and the responses of the respondents in the analysis. The third advantage is the possibility of separating the contribution of different attributes. The disadvantage of the method is that we can't classify non-response according to reasons and hence we can't separate refusals from administrative errors. We start with simple tabulations and later we use multiple regression methods.

The simplest way to see the effect of non-response is to compare the mean or the distribution of variables using non-weighted versus weighted observations. This way we can learn about the quantitative effect of the weighting scheme on the expected value of a variable of interest.

Table 3.1 presents the average of satisfaction from life, weighted and non-weighted. Satisfaction is classified into four discrete categories: (1) very satisfied, (2) satisfied, (3) not so satisfied and (4) not satisfied at all. As a result, the lower the value, the higher is the satisfaction. As can be seen, in most cases, using the weights does not change the average in a noticeable way, implying that non respondents tend to be, on average, equally satisfied with life than the respondents.[6]

[6] The fact that the differences are negligible raises the suspicion that average satisfaction from life is used as a constraint in creating the weighting scheme. We were assured that this is not the case.

Table 3.1: Average satisfaction according to ethnic group*

All			
	Weighted	**Sample**	**Ratio**
2004	1.9426	1.9372	1.003
2005	1.9525	1.9522	1.001
2006	1.9225	1.9324	0.995
2007	1.8819	1.8867	0.997
All Years	**1.9251**	**1.9272**	**0.999**

Jewish			
	Weighted	**Sample**	**Ratio**
2004	1.8949	1.8957	1.000
2005	1.9083	1.9120	0.998
2006	1.8781	1.8938	0.992
2007	1.8388	1.8488	0.995
All Years	**1.8804**	**1.8878**	**0.996**
Non Jewish			
	Weighted	**Sample**	**Ratio**
2004	2.1344	2.1354	1.000
2005	2.1273	2.1226	1.002
2006	2.0984	2.0949	1.002
2007	2.0530	2.0379	1.007
All Years	**2.1023**	**2.0965**	**1.003**

* The average satisfaction is based on individuals that belong to the same category who did respond.

This conclusion seems to contradict the findings of the direct method that participants in the labor market have a higher tendency to respond because we expect participants in the labor market to be more satisfied. One possible explanation is that there are several sources of non response that neutralize the effect of each other. For example, as can be seen the bias in average satisfaction due to non-response is upward while for the non-Jewish population the bias is in the opposite direction.

To further investigate this result, it is worth to look at the relationship between the degree of religiosity and non-response. In the questionnaire, the degree of religiosity among the Jewish population is divided into five categories, while among the non-Jewish one it is divided into four categories. We conducted a separate tabulation for the Jewish and non-Jewish population.

The first three columns of Table 3.2 present the share of each group in the sample and in the population among the Jewish population. As can be seen, the ultra religious population is under-represented in the sample. This means that non-response among the ultra-religious population is higher than the non-response among the rest of the population. Column 4 presents the average satisfaction reported by each group. As

can be seen, the ultra-religious group tends not to participate more than the others but also tend to report higher satisfaction than the others. That religiosity tends to increase life satisfaction is well documented in the literature. See among others, (Luttmer, 2005, Table 1, p-975). We are not aware of reference in the literature to two other unique properties related to religiosity: lower participation in the labor market, and lower response rate in surveys.[7]

Table 3.2: Non-Response According to Religiosity – Jewish Population*

Category	Observations	% of observations	% of weights	Average Satisfaction
Ultra religious	1,713	7.06%	7.52%	1.43
Religious	2,286	9.43%	9.44%	1.80
Traditional but religious	3,156	13.02%	13.10%	1.96
Traditional but no so religious	6,178	25.48%	25.57%	1.94
Non religious, secular	10,850	44.75%	44.37%	1.92
Unknown	64	0.26%	0.27%	1.83
Total	**24,247**	**100%**	**100%**	

* 143 observations with unknown satisfaction were omitted.

Overall, we can conclude that the higher tendency of the ultra-religious population not to participate decreases the average satisfaction among the Jewish population by 0.003 points. This means that correcting for the non-participation of the ultra-religious group does not fully explain the difference in satisfaction between respondents and non-respondents.

Table 3.3 Replicates Table 3.2 for the Non-Jewish population. The pattern of non-participation according to religiosity is a bit different. There is no tendency among the religious groups to participate less than other groups, especially the non-religious group.

[7] The former property is a well known one in Israel, while the latter is documented in Schechtman *et. al*. (2008).

Table 3.3: Non-Response According to Religiosity – Non-Jewish Population*

Category	Observations	% of observations	% of weights	Average Satisfaction
Very religious	335	6.89%	6.89%	1.96
Religious	2,101	43.22%	44.49%	2.07
Not so religious	1,238	25.47%	25.51%	2.18
Non religious at all	1,177	24.21%	22.93%	2.12
Unknown	10	0.21%	0.17%	2.10
Total	**4,861**	**100%**	**100%**	

* 513 observations are missing because either religion is not reported, or they define themselves as atheists and 10 observations with unknown satisfaction.

The non-Jewish population is less satisfied with life than the Jewish counterpart, but the ranking of groups' satisfaction is similar. Not correcting for non-response tends to increase satisfaction with life although marginally so.

Appendix A.1 presents additional classifications intended to find out groups that can contribute to a bias in average satisfaction. We looked at classifications according to age, health status, and participation in the labor market. The only group that seems to contribute to a noticeable bias is the group of young persons.

Overall, we may say that non-participation tends to bias the satisfaction reported by the Jewish population downward, and this finding cannot be fully explained by the lower participation rate of the ultra religious group. Another group that contributes to downward bias in satisfaction is the group of young persons.

4. Mixing Gini and OLS in the same Regression

Analyzing non-response by a regression method has the advantage of enabling control over different properties. In this paper we use a new regression technique which is based on mixing Ordinary Least Squares (OLS) and Gini regression (Schechtman, Yitzhaki and Pudalov (2010). The basic idea is the following: Yitzhaki (1996) has shown that the regression coefficients in a simple OLS or Gini regression can be interpreted as weighted average of slopes defined between adjacent explanatory variable observations. Yitzhaki and Schechtman (2004) have shown that the weighting scheme of both methods can be derived from the Lorenz curve of the independent variable. The bottom line implication of those observations is that the OLS and Gini estimators of the regression coefficients do not rely on the linearity assumption of the regression curve. Schechtman *et. al*. (2008) have used the concept of a statistical linear approximation to a regression curve, that is, estimating a linear

model without assuming that the model is truly linear. Schechtman, Yitzhaki and Pudalov (2010) have extended these ideas to the multiple regression frameworks. The aim of this section is to briefly present the basic derivation of estimators within the framework of mixed OLS and Gini regression. We refer to those regressions as covariance-based regressions because the estimators of the regression coefficients in a multiple regression framework are derived by solving a set of linear equations that are composed of simple regression coefficients that play the role of the parameters in those equations. The presentation is restricted to population parameters. All estimators are sample's analogues of the population parameters.

Let $(Y, X_1,\ldots,X_K)$ be continuous random variables that follow a multivariate distribution with finite second moments. For every choice of constants, α, β_1, $\ldots\beta_K$ define the random variable ε by the following identity

(4.1) $Y \equiv \alpha+ \beta_1X_1 +\ldots+ \beta_KX_K +\varepsilon$.

At this stage, α, $\beta_1,\ldots$, β_K are arbitrary constants ($\beta_1,\ldots,\beta_K$ will later stand for the multiple regression coefficients, while α will be a location parameter). The random variable ε is defined as a slack variable, intended to fulfill identity (4.1). The symbol $\equiv$ is used to indicate that at this stage there are no assumptions imposed on ε and all its properties are determined by the properties of the distribution of $(Y, X_1,\ldots, X_K)$. Identity (4.1) is a tautology, which means that no assumption has been imposed on the regression curve.

Let $T_1,\ldots,T_K$ be K random variables. The covariances between Y and these variables define a set of identities as follows:

(4.2)

$$\begin{aligned}
&\operatorname{cov}(Y, T_1) \equiv\beta_1 \operatorname{cov}(X_1,T_1) +\ldots+ \beta_K \operatorname{cov}(X_K,T_1)+ \operatorname{cov}(\varepsilon , T_1)\\
&\operatorname{cov}(Y,T_k) \equiv\beta_1 \operatorname{cov}(X_1, T_k) +\ldots+ \beta_K \operatorname{cov}(X_K,T_k)+ \operatorname{cov}(\varepsilon , T_k)\\
&\operatorname{cov}(Y,T_K)\equiv\beta_1 \operatorname{cov}(X_1, T_K) +\ldots+ \beta_K \operatorname{cov}(X_K,T_K)+ \operatorname{cov}(\varepsilon , T_K)
\end{aligned}$$

Dividing each line by the appropriate covariance, subject to the assumption that $\operatorname{cov}(X_k,T_k) \neq 0$, (k=1,…,K) we get:

(4.3)

$$\begin{aligned}
&\beta_{01} \equiv \beta_1\, 1 +\ldots+ \beta_K\, \beta_{K1} + \beta_{\varepsilon 1}\\
&\beta_{0k} \equiv \beta_1\, 1 +\ldots+ \beta_K\, \beta_{Kk} + \beta_{\varepsilon k}\\
&\cdot\\
&\beta_{0K} \equiv \beta_1\, \beta_{1K} +\ldots+ \beta_K\, 1 + \beta_{\varepsilon K}
\end{aligned}$$

Where the index 0 indicates the dependent variable,

$\beta_{\varepsilon j} = \dfrac{\mathrm{cov}(\varepsilon, T_j)}{\mathrm{cov}(X_j, T_j)}$ and $\beta_{kj} = \dfrac{\mathrm{cov}(X_k, T_j)}{\mathrm{cov}(X_j, T_j)}$ are a general formula for the regression coefficients in the simple regressions of X_k on T_j, k,j=1,…,K. Two special cases are the **OLS** (iff $T_j = X_j$), and the Gini (iff $T_j = F(X_j)$) Provided that the rank of the matrix of the coefficients composed of the β_{kj} 's is K we get the following "solution" of the identities in (4.3):

$$(4.4) \qquad \begin{pmatrix} \beta_1 \\ \\ \beta_K \end{pmatrix} \equiv \begin{pmatrix} 1 & \beta_{21} & \beta_{K1} \\ & & \\ \beta_{1K} & \beta_{2K} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \beta_{01} - \beta_{\varepsilon 1} \\ \\ \beta_{0K} - \beta_{\varepsilon K} \end{pmatrix} \equiv \mathbf{A}^{-1}[\boldsymbol{\beta}_0 - \boldsymbol{\beta}_\varepsilon] .$$

Where A^{-1} is a KxK matrix, while the $\boldsymbol{\beta}$'s are Kx1 vectors. The set of identities (4.4) is the basic structure of the identities that hold in an arbitrary model.

So far no assumption has actually been imposed, except that $\mathrm{cov}(X_k, T_k) \neq 0$, k=1,…,K, and that the rank of the matrix A is equal to K.

We now impose a set of restrictions. We impose them on the data in the sample. The restrictions hold in the sample by construction, and therefore cannot be verified nor tested without additional information.

The set of restrictions to be imposed, referred to as "orthogonality conditions" is given by

(4.5) $\beta_{\varepsilon k} = 0$, for k=1,…,K.

One possible interpretation of (4.5) can be that it represents first order conditions for an optimization with respect to a target function. This is the case for a specific choice of the variables T_k for example, if $T_k = X_k$, then we are in the OLS regression case. Alternatively, one can follow DeLaubenfels' (2006) geometric interpretation that the inner products of the vectors of explanatory variables and the residual are zero. That is, the explanatory vectors are orthogonal to the residual. In both cases it should be remembered that those conditions are *imposed* on the data and there is no a-priori reason to believe that they exist in the population.

The consequence of imposing the orthogonality conditions is that (4.4) now turns from an identity to a solution of a set of linear equations, so that β_k (k=1,…,K) cease to be arbitrary constants but become the solutions of a set of linear equations.

Formally, using the restriction (4.5), the identities of (4.4) turn into equations (4.6):

$$(4.6) \qquad \begin{pmatrix} \beta_1 \\ \\ \beta_K \end{pmatrix} = \begin{pmatrix} 1 & \beta_{21} & \beta_{K1} \\ & & \\ \beta_{1K} & & 1 \end{pmatrix}^{-1} \begin{pmatrix} \beta_{01} \\ \\ \beta_{0K} \end{pmatrix} = \mathbf{A}^{-1}\,\boldsymbol{\beta}_0 \quad .$$

The structure given in (4.6) is general, and it corresponds to all members of the covariance-based regressions, depending on the choice of T_k, k=1,…,K. Special cases that are relevant to this paper include:[8]

(a) $T_k = X_k$ for all k, k=1,…,K. Then it is easy to see that (4.6) represents the OLS.

(b) $T_k = F(X_k)$ for all k, k=1,…,K. Then (4.6) represents the semi-parametric Gini regression.

Several additional properties of (4.6) are worth mentioning.

By choosing T_k one is choosing the weighting scheme used in the regression, which is actually a choice of the variability measure used (variance in OLS (a), Gini or extended Gini in the regressions defined in (b) and (c) respectively). As a result, this choice determines the metric used (Euclidean in the case of OLS, city block in the case of Gini) and the "orthogonality conditions" applied. In the case of OLS the orthogonality condition is $cov(X_k, \varepsilon) = 0$, under the Gini regression it is $cov(F(X_k),\varepsilon) = 0$, etc…

Each of the K equations in (4.4) can be defined with different T_k so that one can have mixed regression methods: some equations can be defined as based on GMD, others on OLS etc…The advantage of a mixed method is that it enables the user to check the robustness of each **imposed** linear normal equation with respect to different regression methodologies, so that only the linear approximation of the regression curve that is not seriously affected by the choice of the methodology will be leading to a robust conclusion with respect to its sign and magnitude.

Having derived the regression coefficients**,** we turn to the constant term, α. To see whether the residuals are symmetrically distributed around the regression line, one can set the constant term so that the regression line passes either through the mean or through the medians of the observations. Comparisons between the two estimates

[8] There are other members of this family, such as extended Gini regression and instrumental variable estimation but they are irrelevant to this paper

yield a quantitative evaluation on the quality of the fit of the regression line. To do that: define a residual term, ε' , as:

(4.7) $\varepsilon_i^{'} = y_i - \sum \beta_j x_{ji}$.

Then if one wants the regression line to pass through the mean then one solves for α as:

(4.8) $\alpha = E\{\varepsilon'\}$.

On the other hand, if one wants the linear approximation to pass through the median, then one has to set α as the solution for

(4.9) $\underset{\alpha}{\text{Min}}\ E\{|\varepsilon' - \alpha|\}$.

The estimators are sample's values of the population parameters, corrected for the degrees of freedom. Standard errors are calculated using the Jackknife method.

Having estimated the coefficients we turn to the quality of the fit of the linear approximation of the regression curve. Under OLS regime, the R^2 can be interpreted as indicating a measure of correlation between the fitted and the realization of the dependent variable, and as one minus the ratio of the variance of the residual to the variance of the dependent variable. The Gini Mean Difference, (hereafter Gini) method has two correlation coefficients between two random variables, and the regression methodology used in this paper does not minimize the Gini of the residuals (Olkin and Yitzhaki, 1992). Therefore, we substitute the R^2 by three indicators: The (Gini) correlations between the fitted and the realizations of the dependent variable, and one minus the ratio of the Gini of the residuals to the Gini of the dependent variable.

Formally:

(4.10) $\Gamma_{\hat{y}y} = \dfrac{\text{cov}(y, F(\hat{y}))}{\text{cov}(y, F(y))}$ and $\Gamma_{y\hat{y}} = \dfrac{\text{cov}(\hat{y}, F(y))}{\text{cov}(\hat{y}, F(\hat{y}))}$

Where $\hat{y}$ is the linear approximation while F() represents the cumulative distribution. The ratio of the Ginis is:

(4.11) $GR = 1 - \dfrac{\text{cov}(e, F(e))}{\text{cov}(y, F(y))}$.

However, it is important to note that the Gini and the OLS are based on different metrics, so that further research is needed in order to make the concepts of the quality of the fit, comparable.

5. Empirical Results

In this section the weights that are derived in order to adjust the sample to the marginal distributions of key demographic properties of the population are the target of our investigation. The weights are produced by imposing several hundreds of linear constraints on the sample, so that key demographic properties of the population are preserved.

The dependent variable is the weight assigned to each observation. The higher the weight assigned the higher the degree of non-participation in the survey. Non-participation can occur because the respondent was not found, because he or she was not at home or that he or she refused to participate. For the issue of whether the sample is representative, it does not matter what was the reason for failing to participate.

The explanatory variables include age, ethnic group, gender, household size, education level and income. Religiosity was not used in the regressions because of the different categories of Jewish and non-Jewish population, and because unlike other explanatory variables, that potentially could have been used to improve the sampling process, there is no easy way to evaluate this variable prior to the interview.

The explanatory variables include several binary variables like education, gender, and ethnic group. For binary variables it does not matter whether one uses OLS or Gini regression.

In the regression we used two alternative ways to represent income. One was based on administrative source and it is the before-tax earned income of the individual. We refer to this income as Earned Income. Note, that it does not include income of other members of the household nor income from capital or transfers from the government. On the other hand, it includes the income of those who refused to answer the question about income. Earned income is measured in relative terms, that is, each income is divided by the average income in the sample for that year.

The other income used is the income reported by the individual in the survey about before tax income of the whole household. The respondent was asked to choose among ten different ranges of income of the household. Then, the mid-range income was divided by the number of persons in the household, the results were grouped into three new discrete categories: (1) up to 2,000 NIS per person; (2) between 2,001-4,000 NIS per person and (3) above 4,001 NIS per person. For our purpose, we multiplied the income per capita by the number of persons in the household. We refer

to this income as Household Income (HI). We stress the difference between the two different representations of income because it turned out that the way income is represented in the sample is crucial to the conclusions.

Table 5.1 presents the estimates of the mixed OLS and Gini regressions using the Earned Income: On the left-hand side are the OLS estimates while on the extreme right-hand side are the estimates of the Gini regression. Column 1-8 present the estimates of the mixed regressions, with the letter O represents an OLS weighting scheme while G represents the Gini weighting scheme.

The basic regression is for the largest group, which is composed of Jewish women, with above secondary school education but without a B. A. degree.

Comparison of the OLS regression coefficients with column (1) and the Gini regression with column (8) reveals that whenever the explanatory variable is binary, then it does not matter which regression method is used for that variable, as long as the continuous variables remain at the same regression method. Therefore, the difference between the estimates produced by the two methods should be attributed to the three non-binary variables: age, household's size and earned income.

The regression coefficient of age is negative, indicating that for a linear approximation, the higher the age the higher the response rate. However, the magnitude of its impact is about 20 percent higher under OLS regime than under Gini, which is a hint that it is caused by extreme observations, either the young or the elderly. It seems that roughly, we can attribute half of the difference to the direct impact of applying the Gini weighting scheme to age, and another half of the change should be attributed to the covariance with earnings.[9] The sign of the age coefficient is in agreement with the results derived in Section 2, based on the administrative records of the survey department.

The impact of household size is positive which means that the larger the household's size the lower the participation. This finding negates the finding in Schechtman *et. al*. (2008) that the larger the household, the larger the participation rate. The latter was found in the Household's Incomes and Expenditures survey (hereafter HIES). One possible explanation is that in the social survey the interviewer has to locate the individual while in the Household's survey, the participation is of the household. The

[9] Note that it is not meaningful to compare the standard errors of the Gini and OLS estimates because they are not statistically independent.

larger the household size, the higher the probability of establishing a contact with the household.

The impact of earned income on participation seems to be the most important factor in the regression. Whenever the OLS weighting scheme is applied to this variable then the estimate is not lower than minus five, while applying the Gini methodology, then the estimate is not bigger than minus 26. This indicates that the higher the income the higher the participation. This also seems to be in agreement with the findings in the direct method reported in Section 2. It may also be the result of the tendency for higher non-participation among the ultra-religious, which also tends to have lower income. Also, the effect of the correlation of other explanatory variables on the estimate of the coefficient of this variable is negligible. This finding is similar to the one found in Schechtman *et. al*. (2008) concerning participation in the HIES.

The rest of the variables are binary, so that the estimates are not directly affected by the methodology applied to them, but they are affected by the co-variation with other explanatory variables, especially of Earned Income.

The role of education on participation rate seems to differ between the methodologies. According to OLS, the higher the degree held the higher the response rate, but in some cases that are closer to the base group, the differences are not significant. On the other hand, under Gini regime for earnings, we get that high levels of educations, holding a B. A. degree or M. A. degree worsen the response rate. However, for low levels of education (elementary school) both methods agree that low level of education reduce the participation rate.

Being a male improves participation relative to the reference group in a non-significant way under OLS but significantly reduce it under Gini.

Being non-Jewish reduces participation rate under both methods. Again, this result is the opposite of the conclusion reached by Schechtman *et. al*. (2008) that participation rate of non-Jews is significantly higher than the participation rate of Jews. However, this result confirms Feskens *et. al*. (2007) that non response may be more severe among minorities and excluded groups

The constant term was estimated in two ways: one is the usual way of imposing the restriction that the regression line passes through the means, (Equation (4.8)), and the other is to force the regression line to pass through the median, as is the case the Least Absolute Deviation (LAD) regression (Equation (4.9)). In both methods the mean constant term is higher than the median constant term indicating that the distribution

of the residuals is skewed, having a larger tail of positive errors than negative ones. Moreover, the OLS constant term are higher than the Gini's counterpart, which is another indication that the distribution of the residuals is skewed, since the OLS is more sensitive to extreme observations than the Gini regression.

The quality of the fit of the regressions seems similar: while $R^2 = 0.06$, $\Gamma_{\hat{y}y} \cdot \Gamma_{y\hat{y}} = 0.29 \circ 0.25 \approx 0.07$. However, the interpretation of comparison between concepts that are based on different metrics is not clear. All that one can say is that it seems that there is no significant gain in the explanatory power of the regressions under the different regimes.

In the regression, we have omitted one variable with a potential of having an important effect on participation, which is health status. In Appendix A.2 we have reproduced Table 5.1 including health status as an explanatory variable. As can be seen the inclusion of this variable did not change the main conclusions.

A key variable for determining our conclusions is the treatment of the earned income variable. Hence, it is worth to dwell a bit on this variable.

Table 5.2 replicates Table 5.1 with one major difference. Instead of using the earned income that was taken from the administrative file, the income of the household reported in the survey is used. This difference is causing the following changes: (a). There are 4,093 observations with a missing response on income in the survey. Naturally, those observations did not participate in the regression. (b). The income reported in the survey includes all sources of income, in particular transfers from the government. (c). The income in the survey is a result of two stages of grouping, an issue that discussed earlier. Comparison of the OLS column in Table 5.2 with the Gini column reveals that all the signs of the coefficients agree in the two columns so that there is no qualitative difference between the results reported according to the methodologies, and even the magnitudes of the coefficients do not seem to deviate from each other. It is interesting to note that the quality of the fit did not change. Appendix A.2 replicates table 5.2 with health being included as an explanatory variable. Again, there is no noticeable change in the tables.

Having found that the way income is included, and the methodology of the regression may affect the conclusions with respect to participation of different groups deserves further investigation. In Section 6 we search for an explanation of the finding.

Table 5.1: Multiple Gini and OLS Regressions: Dependent Variable, the weight attached to an observation.

Regression Coefficient	OLS		1		2		3		4		5		6		7		8	GINI
Age	-1.19 (0.06)	O	-1.19	G	-1.12	O	-1.12	O	-1.24	G	-1.05	G	-1.05	O	-1.14	G	-0.96	-0.96 (0.07)
Household size	11.19 (0.52)	O	11.19	O	11.37	G	13.29	O	12.59	G	13.47	O	13.03	G	15.38	G	15.87	15.87 (0.60)
Earned Income	-4.32 (0.42)	O	-4.32	O	-4.31	O	-4.40	G	-27.13	O	-4.40	G	-26.58	G	-27.36	G	-26.86	-26.86 (0.87)
Elementary/ middle school or other certification	23.32 (3.29)	G	23.32	O	22.73	O	22.45	O	10.44	G	21.94	G	9.19	G	9.21	O	8.02	8.02 (3.44)
Secondary school without matriculation	1.73 (3.10)	G	1.73	O	1.87	O	1.22	O	-6.01	G	1.33	G	-5.47	G	-6.73	O	-6.25	-6.25 (3.04)
Secondary school with matriculation	10.79 (3.11)	G	10.79	O	11.53	O	11.28	O	2.92	G	11.91	G	5.04	G	3.53	O	5.51	5.51 (3.28)
BA degree	-16.13 (3.29)	G	-16.13	O	-15.87	O	-15.68	O	-0.08	G	-15.44	G	0.25	G	0.62	O	0.94	0.94 (3.25)
MA+ degree	-4.61 (3.67)	G	-4.61	O	-5.00	O	-4.28	O	21.79	G	-4.60	G	20.16	G	22.38	O	20.89	20.89 (4.04)
Jewish Male	-0.15 (2.12)	G	-0.15	O	-0.07	O	-0.23	O	16.03	G	-0.17	G	15.84	G	16.01	O	15.83	15.83 (2.12)
Non-Jewish Male	13.97 (3.72)	G	13.97	O	14.37	O	11.95	O	18.42	G	12.25	G	19.36	G	15.76	O	16.54	16.54 (4.64)
Non-Jewish Female	15.63 (3.74)	G	15.63	O	16.09	O	13.81	O	8.15	G	14.16	G	9.54	G	5.69	O	6.89	6.89 (4.83)
α(mean)	612.43		612.43		608.36		601.92		625.95		598.35		614.93		612.05		601.49	601.49
α(median)	593.67		593.67		589.82		583.36		608.54		579.88		597.48		594.97		584.41	584.41

$R^2 = 0.06$; $\Gamma_{y\hat{y}} = 0.29$; $\Gamma_{\hat{y}y} = 0.25$; GR = 0.01
Number of observations: 28,029

Table 5.2: Multiple Gini and OLS Regressions: Income reported by the interviewed

Regression Coefficient	OLS		1		2		3		4		5		6		7		8	Gini
Age	-1.15 (0.07)	O	-1.15	G	-1.11	O	-1.08	O	-1.14	O	-1.07	G	-1.11	G	-1.05	G	-1.05	-1.05 (0.10)
Household size	12.48 (0.73)	O	12.48	O	12.57	G	16.35	O	10.86	G	14.67	O	10.92	G	16.44	G	14.74	14.74 (0.92)
Survey's Income	-1.56 (0.40)	O	-1.56	O	-1.56	O	-2.90	G	-0.15	G	-1.57	G	-0.14	O	-2.90	G	-1.57	-1.57 0.5
Elementary/ middle school or other certification	24.49 (3.56)	G	24.49	O	24.18	O	22.22	O	26.27	G	23.92	G	26.02	G	21.97	O	23.72	23.72 (3.72)
Secondary school without matriculation	-0.34 (3.32)	G	-0.34	O	-0.29	O	-1.87	O	0.86	G	-0.73	G	0.90	G	-1.85	O	-0.70	-0.70 (3.26)
Secondary school with matriculation	11.17 (3.35)	G	11.17	O	11.50	O	11.66	O	11.34	G	11.78	G	11.62	G	11.91	O	11.99	11.99 (3.52)
BA degree	-17.00 (3.49)	G	-17.00	O	-16.87	O	-14.94	O	-18.72	G	-16.59	G	-18.62	G	-14.83	O	-16.50	-16.50 (3.33)
MA+ degree	-6.44 (3.85)	G	-6.44	O	-6.63	O	-4.55	O	-8.22	G	-6.23	G	-8.38	G	-4.68	O	-6.34	-6.34 (3.75)
Jewish Male	-1.13 (2.20)	G	-1.13	O	-1.14	O	-0.30	O	-2.27	G	-1.35	G	-2.29	G	-0.30	O	-1.36	-1.36 (2.06)
Non-Jewish Male	18.16 (4.83)	G	18.16	O	18.40	O	12.98	O	19.92	G	14.87	G	20.12	G	13.13	O	14.99	14.99 (6.55)
Non-Jewish Female	34.28 (5.08)	G	34.28	O	34.58	O	28.62	O	36.53	G	30.98	G	36.80	G	28.81	O	31.14	31.14 (7.00)
α(mean)	594.10		594.10		592.07		585.33		591.15		583.36		589.46		583.73		582.05	582.05
α(median)	576.55		576.55		574.43		568.59		573.05		565.78		571.33		566.95		564.42	564.42

$R^2 = 0.06$; $\Gamma\hat{y}y = 0.25$; $\Gamma y\hat{y} = 0.25$; GR =0.03
Number of observations: 23,936

6. A Search for an explanation

We have seen in the last section that if one uses earned income from administrative sources then the signs of several regression coefficients of other explanatory variables may disagree between the two methods, while if one uses the income reported in the survey, then the two methods produce similar estimates. There are three major differences between the two incomes: The earned income variables includes 4,093 additional observations, of those with a missing income variable in the survey; the earned income variable includes actual earned income while the income in the survey was grouped into rough categories; On the other hand, the income variable in the survey includes income from all sources and not only earned income. In this section we will try to find out the effect of the differences between the variables.

Figure 6.1 presents the density function of earned income. Before plotting the density function three observations with very large incomes were deleted. As can be seen, it still includes some very extreme observations, with the highest income being about 60 times the average income. Those extreme incomes overshadow the whole distribution. In general earned income is skewed.

Figure 6.1: The density function of earned income

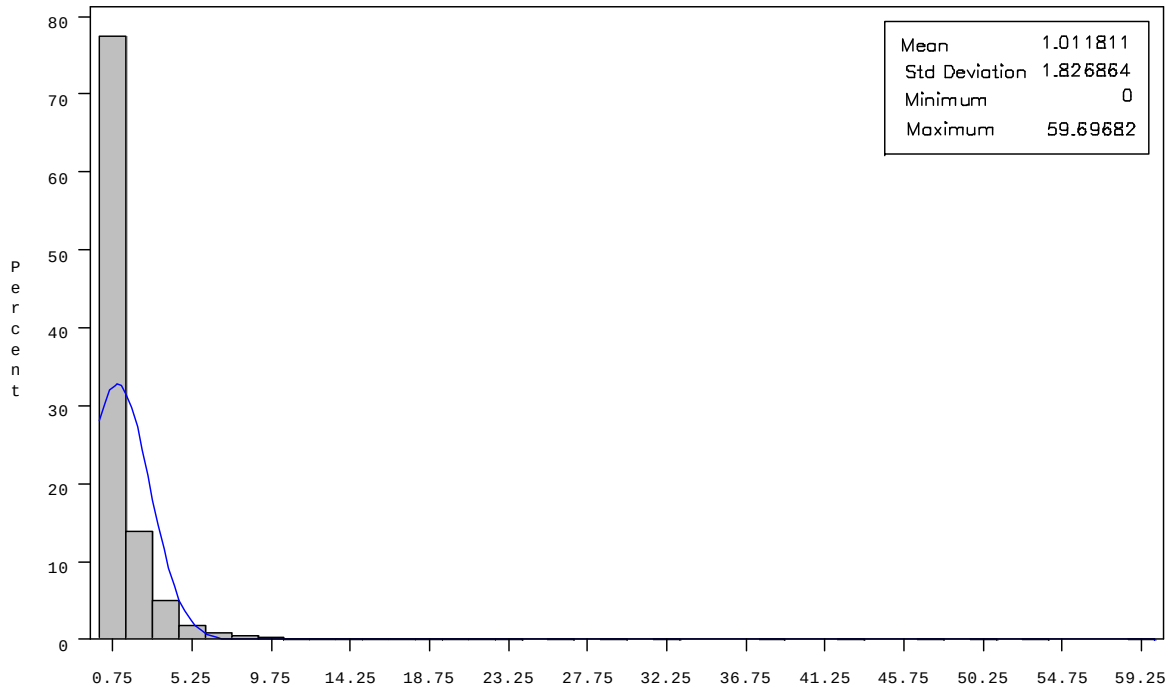
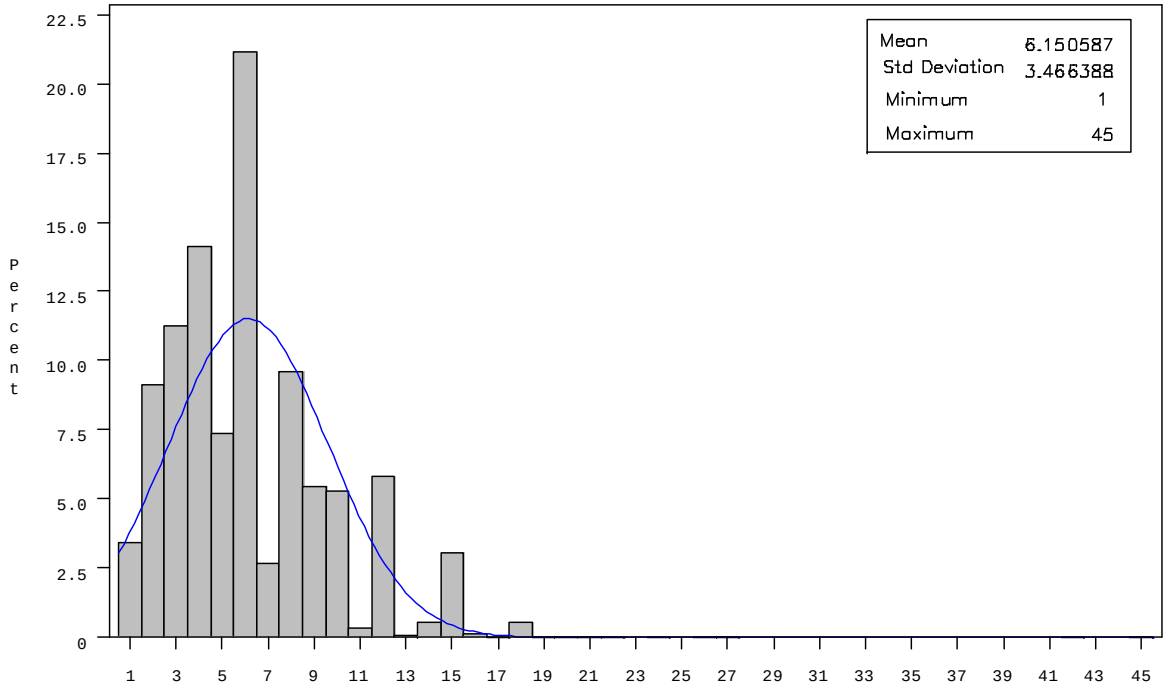


Figure 6.2 presents the density function of the household's income reported in the survey. As can be seen, it is less skewed than the distribution of earned income, the grouping of observations makes it less asymmetric so that it is almost like a truncated normal. One possible conclusion is that decreasing the asymmetry of the distribution of income reduces the difference between the estimates derived by the two methodologies.

Figure 6.2: The density function of the income in the survey



To see, whether the omitted observations caused the difference between the estimates of the two methods, we reran the regression with earned income omitting three extreme observations of earned income. As can be seen from Table 6.1, the difference in the effect of earned income is still very big while the effect of having a B.A. degree is still with a negating signs, although the differences between the estimates produced by the two methods have somewhat reduced.

Table 6.1: Multiple Regressions: 3 observations were omitted.

Regression Coefficient	OLS	Gini
Age	-1.23 (0.07)	-1.07 (0.10)
Household size	11.59 (0.58)	16.28 (0.64)
Earned Income	-8.32 (0.61)	-27.38 (0.85)
Elementary/ middle school or other certification	20.17 (3.60)	7.81 (3.80)
Secondary school without matriculation	-1.27 (3.37)	-7.97 (3.29)
Secondary school with matriculation	10.01 (3.41)	5.52 (3.59)
BA degree	-13.63 (3.40)	0.70 (3.33)
MA+ degree	0.98 (3.94)	21.17 (3.82)
Jewish Male	3.92 (2.26)	17.66 (2.14)
Non-Jewish Male	20.43 (4.89)	19.81 (6.68)
Non-Jewish Female	32.60 (5.13)	21.46 (7.72)
α(mean)	614.58	604.29
α(median)	596.37	587.87

$R^2 = 0.07$; $\Gamma\hat{y}y = 0.22$; $\Gamma y\hat{y} = 0.21$; GR = 0.01
Number of observations: 23,933

Table 6.2 replicates Table 6.1 with one major difference: all observations with no earned income were omitted from the regression. This means that we omitted non-participants in the labor market. Comparisons of the two columns indicates that there is no disagreement with respect to the signs of the regression coefficients although one can observe quantitatively large differences between some estimates: The impact of earned income is different -3 in the OLS, -10 in the Gini, the effect of a B.A. degree is -6 and significant under the OLS, -0.07 and insignificant under the Gini.

Table 6.2: Multiple Regressions without observations with zero earned income

Regression Coefficient	OLS	Gini
Age	-1.39 (0.10)	-1.18 (0.10)
Household size	7.96 (0.66)	9.87 (0.72)
Earned Income	-2.93 (0.62)	-10.42 (1.03)
Elementary/ middle school or other certification	-1.16 (4.70)	-6.35 (4.85)
Secondary school without matriculation	-5.77 (3.77)	-8.92 (3.75)
Secondary school with matriculation	8.22 (3.77)	6.82 (3.95)
BA degree	-6.22 (3.72)	-0.07 (3.63)
MA+ degree	1.54 (4.29)	11.38 (4.36)
Jewish Male	14.04 (2.56)	21.53 (2.57)
Non-Jewish Male	25.10 (5.30)	25.27 (7.03)
Non-Jewish Female	-49.47 (8.07)	-52.89 (12.01)
A(mean)	606.74	598.87
A(median)	593.05	585.37

$R^2 = 0.04$; $\Gamma\hat{y}y = 0.17$; $\Gamma y\hat{y} = 0.17$; GR = 0.01
Number of observations: 15,135 (8,798 observations were omitted).

Based on the comparison between Table 6.1 and 6.2 it seems that the difference between the results produced by the two methodologies is affected by whether one includes in the regression observations of individuals with no earned income. If one omits those observations, then the two methods produce similar results. The major change that occurs is that the effect of education turned to be insignificant. An alternative way of getting similar results by both methods is by using the income definition reported in the survey. Appendices A.3 and A.4 report the results of two sensitivity tests: in A.3 we re-estimated the regressions with earned income, omitting observations which do not report income in the survey. There is no meaningful change in the estimates. In A.4 we estimated the regression coefficient among those with zero earned income, using the income reported in the sample. One does not observe major changes in the estimates between the two methodologies. .

7. Conclusions

In general, the effect of non-response on average satisfaction reported in the social survey in Israel turned out to bias average satisfaction downward. However, this bias is relatively small. This result can be attributed to two factors: on one hand the groups with a lower participation in the labor market tend also to have a lower participation rate in the survey. The most satisfied groups are the ultra-religious Jewish group and the young who also have a lower participation rate both in the labor market and in the survey. On the other hand, the higher the income the higher the participation rate in the survey. The result is a small bias downward. In general non-response occurs mainly among the elderly and the young, and it tends to decline with an increase in income. It is also important to report that we do not observe low response among minorities.

In this paper we also applied a mixed Gini and OLS regression, so that we avoided conclusions that are due to the use of one methodology. It turned out that using different methodologies can sometimes result in contradicting signs of regression coefficients. This phenomenon should bother us because it means that the regression methodology used can reverse our conclusions. We have traced this phenomenon as to whether one includes in the sample participants and non-participants in the labor market. Our guess is that this result is due to nonlinearity of the regression curve with respect to earned income when both participants and non-participants in the labor market are included in the regression. However, we can't exclude other possible explanations, such as grouping of the income variable. The advantage of the mixed regression methodology is that it enables us to find out the variable or the action that can change the sign of the regression coefficients and as a result to reverse the conclusions. Further research is needed to find out whether this fragility of the regression-based research is limited to extreme cases.

Acknowledgment: A SAS program that can estimate the mixed regression is written by Alexandra Katzenelenbogen. The program will be sent upon request. We are also grateful to Dmitri Romanov and Moses Shayo for helpful discussions.

References:

Davis, P. S. And T. L. Fisher (2009). Measurement Issues Associated with Using Survey Data Matched with Administrative Data fron the Social Security Administration, *Social Security Bulletin*, 69, 2, 1-12.

De Leeuw, E.D. & De Heer, W. (2002). Trends in Household Survey nonresponse: A Longitudinal and International Comparison. In *Survey Nonresponse*, In: R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (Eds). *Survey nonresponse*. New York: John Wiley, pp. 41-54.

DeLaubenfels, R. (2006). The victory of least squares and orthogonality in Statistics. *The American Statistician*, 60, 4, (November), 315-321.

Feskens, Remco; Joop Hox; Gerty Lensvelt-Mulders and Hans Schmeets (2007). Nonreponse Among Ethnic Minorities: A multivariate Analysis. *Journal of Official Statistics*, 23, 3, 387-408.

Helliwell, J. P. (2010). Measuring and Understanding Subjective Well-Being, National Bureau of Economic Research, Working paper no. 15887, (April).

Luttmer, E. F. P. (2005). Neighbors, as Negatives: Relative Earnings and Well-Being. *Quarterly Journal of Economics*, 120, 3, (august), 963-1002.

Olkin, Ingram and S. Yitzhaki (1992). Gini Regression Analysis. *International Statistical Review*, 60, 2, August, 185-196.

Romanov, D. and M. Nir (2010). Get It or Drop It? Cost-Benefit Analysis of Attempts to Interview in Household Surveys, *Journal of Official Statistics*, 26, 1, 165-191.

Schechtman, E; S. Yitzhaki and Y. Artzev (2008). Who Does Not Respond in the Household Expenditure Survey: An Exercise in Extended Gini Regressions, *Journal of Business & Economic Statistics*, 26, 3, July, 329-344.

Schechtman, E; S. Yitzhaki, T. Pudalov (2010). Gini's multiple regressions: two approaches and their interaction. Draft, Http://SSRN.com

Schmmack, U., P. Krause, G. G. Wagner, and J. Schupp (2009). Stability and Change in Well-Being: An Experimentally Enhanced Latent State-Trait-Error Analysis, *Social Indicators Research*, 95, 19-31.

Stiglitz, J. E., A. Sen and J. P. Fitoussi (2009). Report by the Commission on the Measurement of Economic and Social Progress, http://www.stiglitz-sen-fitoussi.fr/documents/rapport_anglais.pdf

Yitzhaki, S. (1996). On Using Linear Regression in Welfare Economics, *Journal of Business & Economic Statistics*, 14, 4, October, 478-86.

Yitzhaki, S. and E. Schechtman (2004). The Gini Instrumental Variable, or the "double instrumental variable" estimator, *Metron,*, LXII, 3, 287-313.

Appendices:

A.1 Appendix or Section 3

The following tables are intended to find out the effect of non-response on satisfaction according to different classifications of the population. Table A.1 presents classification according to age groups, table A.2 presents classification according to participation in the labor market, while table A.3 presents the classification according to health status. As can be seen the group that biases the average satisfaction the most is the group of young persons. Non-participation of young persons biases average satisfaction from life by 0.005 points. Note, however that since a person can be included in several categories there is no point in adding up biases caused by non-participation according to different classifications.

Table A.1: Non-Response According to Age

Category	Observations	% of Observations	% of weights	Average satisfaction*
20-24	3,539	11.95%	12.62%	1.69
25-44	12,393	41.84%	42.68%	1.87
45-64	8,927	30.14%	29.74%	2.00
65+	4,762	16.08%	14.96%	2.14
Total	**29,621**	**100.00%**	**100.00%**	

Table A.2: Non-Response According to Participation in the labor market

Category	Observations	% of Observations	% of weights	Average satisfaction*
Working	18,132	61.21%	59.28%	1.86
Not Working	11,489	38.79%	40.72%	2.01
Total	**29,621**	**100.00%**	**100.00%**	

Table A.3: Non-Response According to Health status

Category	Observations	% of Observations	% of weights	Average satisfaction*
Very good	12,821	43.28%	43.67%	1.64
Good	10,352	34.95%	34.48%	1.99
Bad	4,544	15.34%	15.39%	2.29
Very bad	1,875	6.33%	6.36%	2.62
Unknown	29	10.0%	10.0%	1.95
Total	**29,621**	**100.00%**	**100.00%**	

* 153 observations with unknown satisfaction were not included.

Appendix A.2: The effect of adding health status

The following two regressions are intended to find out whether adding health status as an explanatory variable would affect the results. The Evaluation of health is classified into five categories: (0) don't know; (1) very good; (2) good; (3) bad; (4) very bad. The difference between the regressions is that the first regression used the earned income and the second regression used the survey's income.

As can be seen there is no major changes in the values of the regression coefficients.

Table A.4: Multiple Regressions - The variables are: Age, Household size, Evaluation of health, Earned Income, Education, Gender and Religion.

Regression Coefficient	OLS		1		2		3		4		5		6		7		8		9		10	Gini
Age	-1.58 (0.07)	O	-1.58	G	-1.47	O	-1.51	O	-1.55	O	-1.45	O	-1.34	G	-1.18	G	-1.13	G	-1.40	G	-1.11	-1.11 (0.10)
Household size	11.16 (0.53)	O	11.16	O	11.33	G	13.34	O	11.14	O	12.38	G	15.22	O	12.77	G	15.69	G	13.50	G	15.68	15.68 (0.60)
Evaluation of health	15.73 (1.32)	O	15.73	O	14.82	O	15.99	G	14.75	O	8.66	G	8.28	G	5.89	O	7.06	G	14.19	G	6.32	6.32 (1.55)
Earned Income	-3.791 (0.42)	O	-3.79	O	-3.81	O	-3.87	O	-3.82	G	-25.78	G	-26.09	G	-25.52	G	-25.72	O	-3.92	G	-25.80	-25.80 (0.86)
Elementary/ middle school or other certification	17.17 (3.34)	G	17.17	O	16.72	O	16.49	O	17.42	O	6.82	G	6.01	G	6.08	G	5.03	G	16.37	O	5.19	5.19 (3.47)
Secondary school without matriculation	0.174 (3.09)	G	0.17	O	0.36	O	-0.27	O	0.23	O	-6.59	G	-7.19	G	-6.01	G	-6.72	G	-0.06	O	-6.70	-6.70 (3.04)
Secondary school with matriculation	10.39 (3.10)	G	10.39	O	11.19	O	10.95	O	10.39	O	2.99	G	3.65	G	5.01	G	5.56	G	11.63	O	5.54	5.54 (3.27)
BA degree	-13.63 (3.28)	G	-13.63	O	-13.52	O	-13.06	O	-13.81	O	0.47	G	1.21	G	0.40	G	1.34	G	-13.15	O	1.25	1.25 (3.26)
MA+ degree	-2.914 (3.67)	G	-2.91	O	-3.41	O	-2.59	O	-3.00	O	21.71	G	22.29	G	20.09	G	20.82	G	-3.09	O	20.82	20.82 (4.00)
Jewish Male	1.36 (2.05)	G	1.36	O	1.30	O	1.35	O	1.25	O	16.14	G	16.18	G	15.69	G	15.85	G	1.18	O	15.80	15.80 (2.06)
Non-Jewish Male	19.46 (4.61)	G	19.46	O	19.94	O	16.33	O	19.46	O	23.75	G	19.69	G	24.83	G	20.55	G	16.68	O	20.56	20.56 (6.35)
Non-Jewish Female	29.14 (4.81)	G	29.14	O	29.83	O	25.92	O	29.21	O	21.83	G	17.60	G	23.62	G	19.06	G	26.51	O	19.09	19.09 (6.73)
α(mean)	601.14		601.14		597.48		590.03		601.93		619.40		605.56		610.90		596.21		587.60		596.82	596.82
α(median)	582.95		582.95		579.33		571.72		583.69		602.31		588.73		593.73		579.40		569.52		579.98	579.98

$R^2 = 0.07$; $\Gamma_{y\hat{y}} = 0.29$; $\Gamma_{\hat{y}y} = 0.25$; GR = 0.009; Number of observations: 28,029

Table A.5: Multiple Regressions - The variables are: Age, Household size, Evaluation of health, Survey's Income, Education, Gender and Religion.

Regression Coefficient	OLS		1		2		3		4		5		6		7		8		9		10	Gini
Age	-1.66 (0.08)	O	-1.66	G	-1.61	O	-1.57	O	-1.63	O	-1.68	O	-1.57	G	-1.61	G	-1.56	G	-1.51	G	-1.54	-1.54 (0.10)
Household size	12.53 (0.74)	O	12.53	O	12.63	G	16.46	O	12.56	O	10.68	G	14.58	O	10.81	G	14.61	G	16.61	G	14.67	14.67 (0.92)
Evaluation of health	17.92 (1.45)	O	17.92	O	17.46	O	17.41	G	16.96	O	18.96	G	17.34	G	17.59	O	18.07	G	16.03	G	17.04	17.04 (1.64)
Survey's Income	-0.93 (0.41)	O	-0.93	O	-0.95	O	-2.30	O	-0.98	G	0.70	G	-0.82	G	0.63	G	-0.78	O	-2.38	G	-0.84	-0.84 (0.52)
Elementary /middle school or other certification	18.82 (3.64)	G	18.82	O	18.57	O	16.62	O	19.01	O	20.60	G	18.54	G	20.58	G	18.19	G	16.61	O	18.37	18.37 (3.86)
Secondary school without matriculation	-0.37 (3.38)	G	-0.37	O	-0.31	O	-1.91	O	-0.35	O	0.95	G	-0.62	G	1.01	G	-0.60	G	-1.86	O	-0.60	-0.60 (3.32)
Secondary school with matriculation	12.70 (3.41)	G	12.70	O	13.06	O	13.20	O	12.70	O	12.89	G	13.33	G	13.19	G	13.56	G	13.48	O	13.56	13.56 (3.58)
BA degree	-14.51 (3.56)	G	-14.51	O	-14.44	O	-12.51	O	-14.65	O	-16.29	G	-14.36	G	-16.37	G	-14.17	G	-12.58	O	-14.31	-14.31 (3.32)
MA+ degree	-4.66 (3.92)	G	-4.66	O	-4.89	O	-2.78	O	-4.72	O	-6.59	G	-4.67	G	-6.82	G	-4.76	G	-3.00	O	-4.80	-4.80 (3.68)
Jewish Male	1.36 (2.24)	G	1.36	O	1.30	O	2.14	O	1.25	O	0.20	G	0.95	G	0.04	G	1.03	G	1.98	O	0.92	0.92 (2.08)
Non-Jewish Male	18.22 (4.92)	G	18.22	O	18.45	O	12.94	O	18.15	O	20.27	G	15.05	G	20.36	G	15.27	G	12.98	O	15.16	15.16 (6.75)
Non-Jewish Female	32.83 (5.17)	G	32.83	O	33.17	O	27.11	O	32.84	O	35.35	G	29.76	G	35.62	G	29.96	G	27.33	O	29.94	29.94 (7.78)
α(mean)	596.47		596.47		594.69	O	588.06		597.35		592.22		585.92		591.68		583.81		587.55		584.75	584.75
α(median)	578.44		578.44		576.61	O	570.14		579.22		573.58		567.55		572.85		565.51		569.65		566.35	566.35

$R^2 = 0.07$; $\Gamma_{y\hat{y}} = 0.19$; $\Gamma_{\hat{y}y} = 0.23$; GR = 0.02; Number of observations 23,936

Appendix A.3: The effect of omitting observations with no response about income

The following regression is intended to find out whether omitting observations according to the number of observation in the survey, will change the values of the estimates in the regression. As can be seen there is no major changes in the values of the regression coefficients.

Table A.6: Multiple Regressions: Observations included are only those who responded in the survey.

Regression Coefficient	OLS	1		2		3		4		5		6		7		8		Gini
Age	-1.21 (0.07)	O	-1.21	G	-1.16	O	-1.14	O	-1.30	O	-1.202	G	-1.13	G	-1.10	G	-1.05	-1.05 (0.10)
Household size	11.31 (0.58)	O	11.31	O	11.44	G	13.51	O	12.87	G	15.85	O	13.23	G	13.62	G	16.25	16.25 (0.67)
Earned Income	-4.18 (0.43)	O	-4.18	O	-4.17	O	-4.26	G	-26.77	G	-27.08	G	-26.25	O	-4.26	G	-26.62	-26.62 (0.95)
Elementary/ middle school or other certification	22.39 (3.60)	G	22.39	O	21.95	O	21.78	O	10.09	G	9.16	G	8.95	G	21.45	O	8.11	8.11 (3.73)
Secondary school without matriculation	0.12 (3.37)	G	0.12	O	0.18	O	-0.29	O	-7.38	G	-8.01	G	-7.01	G	-0.25	O	-7.69	-7.69 (3.29)
Secondary school with matriculation	11.59 (3.42)	G	11.59	O	12.06	O	12.18	O	4.30	G	5.03	G	5.96	G	12.53	O	6.54	6.54 (3.63)
BA degree	-16.39 (3.54)	G	-16.39	O	-16.20	O	-15.92	O	-0.10	G	0.67	G	0.14	G	-15.79	O	0.90	0.90 (3.48)
MA+ degree	-3.18 (3.92)	G	-3.18	O	-3.44	O	-2.88	O	23.07	G	23.71	G	21.65	G	-3.07	O	22.45	22.45 (4.30)
Jewish Male	0.88 (2.24)	G	0.88	O	0.87	O	0.71	O	17.55	G	17.47	G	17.15	G	0.70	O	17.11	17.11 (2.26)
Non-Jewish Male	19.76 (4.90)	G	19.76	O	20.08	O	16.47	O	23.31	G	18.88	G	24.26	G	16.67	O	19.63	19.63 (6.65)
Non-Jewish Female	34.27 (5.14)	G	34.27	O	34.69	O	30.93	O	25.08	G	20.47	G	26.63	G	31.21	O	21.75	21.75 (7.15)
α(mean)	611.61		611.61		608.78		600.83		626.13		611.65		616.69		598.65		602.87	602.87
α(median)	592.97		592.97		590.22		582.30		608.83		594.56		599.43		580.06		585.86	585.86

$R^2 = 0.06$; $\Gamma_{y\hat{y}} = 0.29$; $\Gamma_{\hat{y}y} = 0.25$; GR = 0.007; Number of observations 23,936

Appendix A.4: The effect of the observations without earned income, but with survey's income.

The following regression is intended to find out how the explanatory variables behave when we use the survey's income instead of the zero earned income.

Table A.7: Multiple regressions: 8,798 observations with zero earned income.

Regression Coefficient	OLS	Gini
Age	-2.14	-1.91
	(0.11)	(0.10)
Household size	19.22	28.23
	(1.37)	(1.77)
Survey's Income	-3.18	-5.78
	(0.87)	(1.05)
Elementary/ middle school or other certification	6.63	3.49
	(5.90)	(5.95)
Secondary school without matriculation	-8.65	-11.24
	(6.26)	(6.16)
Secondary school with matriculation	-0.23	1.38
	(6.49)	(6.98)
BA degree	-29.52	-26.35
	(7.66)	(7.66)
MA+ degree	-9.28	-6.72
	(7.71)	(7.81)
Jewish Male	-18.70	-18.35
	(4.16)	(3.85)
Non-Jewish Male	39.64	26.21
	(9.85)	(14.61)
Non-Jewish Female	11.20	-2.46
	(7.36)	(8.98)
α(mean)	702.74	675.54
α(median)	685.46	658.60

$R^2 = 0.15$; $\Gamma_{y\hat{y}} = 0.43$; $\Gamma_{\hat{y}y} = 0.41$; GR = 0.08
Number of observations 8,798