WORKING PAPER SERIES
סדרת ניירות עבודה

No. 12 מס׳

Who Does Not Respond in the Household Expenditure Survey: An Exercise in Extended Gini Regressions

Edna Schechtman, Shlomo Yitzhaki* and Yevgeny Artsev

מי לא משיב לשאלוני סקר הו״מ: תרגיל ברגרסיות של ג׳יני מורחב

עדנה שכטמן, שלמה יצחקי* ויבגני ארצב

אלול תשס״ה, ספטמבר September 2005

* Corresponding author: Shlomo Yitzhaki, Central Bureau of Statistics, 66 Kanfei Nesharim St., Jerusalem, 95464, E-mail: yitzhaki@cbs.gov.il

Published by the Central Bureau of Statistics, 66 Kanfei Nesharim St.,

Corner Bachi St., P.O.B 34525, Jerusalem 91342, Israel

Tel. 972-2-6592666; Fax: 72-2-6521340

Internet Site: www.cbs.gov.il

E-Mail: info@cbs.gov.il

הוצאת הלשכה המרכזית לסטטיסטיקה, רח' כנפי נשרים 66, פינת רח' בקי,

ת"ד 34525, ירושלים 91342

טל': 02-6592666; פקס: 02-6521340

אתר הלמ"ס באינטרנט: www.cbs.gov.il

דואר אלקטרוני: info@cbs.gov.il

מי לא משיב לשאלוני סקר הו״מ: תרגיל ברגרסיות של ג׳יני מורחב

תקציר

מטרת העבודה היא להציג גישה חדשה של רגרסיה אי-פרמטרית, אשר מבוססת על מדידת פיזור לפי ג׳יני מורחב ((Extended Gini (EG ומאפשרת לעקוב אחר צורתו של עקום הרגרסיה. הגישה מבוססת על אמידה של סדרה של קירובים ליניאריים לעקום הרגרסיה, דבר שמאפשר להדגיש חלקים שונים לאורך משתנה בלתי תלוי מסוים, ללא שינוי ההתייחסות למשתנים בלתי תלויים אחרים.

הגישה מבוססת על שימוש במדדי פיזור התלויים בפרמטר אחד בלבד, v. ככל ש-v גבוה יותר כן מודגש במדידת הפיזור החלק התחתון של ההתפלגות. ממדד הפיזור גוזרים מקדמי מתאם ומקדם רגרסיה – כך שמכל מדד פיזור ניתן ליצור אומדן חלופי לעקום הרגרסיה. עד כמה שידוע לנו – גישה זו היא הגישה היחידה המאפשרת לשקלל את שיפועי עקום הרגרסיה על פי תחומים של המשתנה הבלתי תלוי.

ההבדל בין האומדנים נובע מצורת שקלול של שיפועי עקום הרגרסיה. באמצעות בחינת דפוסי השינוי של מקדמי הרגרסיה ניתן לנתח את צורת העקמומיות של עקום הרגרסיה.

יישום השיטה נעשה על בחינה של דפוסי אי-השבה על שאלוני סקר הוצאות משקי הבית. מוקד העניין הוא בבדיקת קיום הקשר בין אי-ההשבה להכנסה של משק הבית וההשפעה של קשר זה על הטיות במדידת אי שוויון בהכנסות.

הניתוח האמפירי מגלה כי שיעור ההשבה עולה עם הגידול בהכנסה ועם גודל משק הבית. בנוסף, שיעור ההשבה בקרב האוכלוסייה הערבית גבוה יותר מאשר בקרב האוכלוסייה היהודית, ואילו הקהילה החרדית נוטה להשיב פחות מאשר שאר האוכלוסייה. שני הממצאים האחרונים תקפים, ללא תלות בהכנסה ובגודל משק הבית. בגלל שאי ההשבה קטן עם ההכנסה, אי התחשבות באי השבה תטה להגדיל את אומדן ההכנסה הממוצעת ולהקטין את אי השוויון בהכנסות במשק.

We are grateful to Shaul Lach, Malka Kantorowitz, Aryeh Reiter, and Dmitri Romanov for helpful comments and discussions. We thank the associate editor of the Journal of Business & Economic Statistics and two anonymous referees for many critical comments that improved the paper.

WHO DOES NOT RESPOND IN THE HOUSEHOLD EXPENDITURE SURVEY: AN EXERCISE IN EXTENDED GINI REGRESSIONS

Abstract

The aim of this paper is to suggest a new, nonparametric regression method, based on the Extended Gini (EG) measures of dispersion, which enables the user to follow the curvature of the regression curve. The method is capable of estimating a series of linear approximations of the regression curve, allowing the investigator to stress different sections along the range of one independent variable, while keeping the treatment of other independent variables intact. The method is based on the extended Gini family, which depends on one parameter, ν. The choice of this parameter enables the user to produce infinite alternative estimators of the regression curve. The difference between them lies in the weighting schemes applied to the slopes of the regression curve. By investigating the patterns of changes in those regression coefficients, the curvature of the regression curve can be traced.

As an application, we investigate nonresponse patterns in the survey of household expenditures in Israel. We will mainly be interested in whether nonresponse increases or decreases with income, and the kind of functional relationship one can find between income and nonresponse.

The empirical illustration shows that the higher the income, the larger the response rate, and the larger the household, the higher the response rate. Also, the Arab population tends to respond more than the Jewish one, while the ultra religious group tends to respond less than the rest of the population. Those last two results hold with and without adjustment for income and household size. The implications on the bias in the estimates are discussed.

Keywords: nonresponse, regression, Gini.

INTRODUCTION

The purpose of this paper is to present a new, nonparametric method for investigating the curvature of a regression curve. The method is capable of estimating a series of linear approximations of the regression curve, allowing the investigator to stress different sections along the range of one independent variable, while keeping the other independent variables intact. This property enables the researcher to learn whether the conditional regression curve is linear, convex or concave in one independent variable, given the weighting scheme applied to slopes of other independent variables. The main purpose of the method is descriptive. One major advantage is that it can treat each independent variable individually, concentrating on the (conditional) relationship between the dependent variable and one independent variable, given the other independent variables. All regressions rely on all observations, eliminating the need to arbitrarily define windows, or omit observations.

The methodology is illustrated by investigating the tendency not to respond to questionnaires on finances of the household in official surveys. The common wisdom with respect to this issue is that either or both rich and poor people tend to respond less than ordinary people. Since we have no firm priors with respect to the kind of relationship we expect to see, the need for a nonparametric method that can analyze the data arises.

The main conclusion of the empirical application is that nonresponse to the survey of family expenditures is a decreasing convex function of income, and almost reaches a plateau when high-income groups are stressed. Nonresponse tends to be negatively related to household size. The nonresponse rate differs among ethnic groups: the Arab population shows below average nonresponse rate, while the ultra religious Jewish group has above average nonresponse rate. This result holds even when the response rate is adjusted for income and household size.

Since the estimators in this paper are based on the sample's analogues of the population parameters, we will use capital letters to represent population parameters and small letters to represent the estimators.

The structure of the paper is the following: The first section presents the nonparametric regression coefficients in the simple regression case. Section 2 extends it to a multiple regression framework, and the third section presents the estimators and their standard errors. Readers who are not interested in the exact derivation of the

estimates and their standard errors can skip the third section. Section 4 presents the data and the research question, the fifth section presents the empirical results while the sixth evaluates the implication of nonresponse on measurement of inequality. Section 7 concludes.

SECTION 1: THE SIMPLE REGRESSION CASE

Let (Y,X) be a bi-variate random variable with expected values μ_Y and μ_X and finite variances σ_Y^2 and σ_X^2, respectively, and let g(x) = E{Y|X=x} be the regression curve. The error term at (Y_i, X_i) is defined as the deviation of Y_i from a linear approximation $\alpha + \beta X_i$, i.e., $\varepsilon_i = Y_i - \alpha - \beta X_i$, where α and β are parameters to be defined later. No assumptions are imposed on the error term, and no structure is imposed on the regression curve. In particular, this means that $E\{\varepsilon |X=x\} = g(x) - \alpha - \beta x$, which is equal to zero for all values of x only if g(x) is a linear function of x.

We are interested in estimating a linear approximation to the regression curve, g(x). We start with the parameter representing the slope, β, and only later the constant term is dealt with. The parameter representing the slope of the linear approximation of the regression curve will be referred to as the EGRC (Extended Gini Regression Coefficient). The slope will be defined as a weighted average of the derivatives of g(x), with the weights being derived from the extended Gini variability index. In some sense, the approach in this paper can be viewed as similar to the one presented in Angrist, Chernozhukov and Fernández-Val (2004), who analyzed the linear approximation of a misspecified model in a quantile regression approach.

The extended Gini variability index is a member of a family of indices defined by

$$G(X,\nu) = -(\nu+1)COV(X,[1-F(X)]^{\nu}),\ \nu>-1,\ \nu \neq 0.$$

By determining ν the investigator introduces his preference concerning the measurement of variability of the independent variable. The role of ν in the extended Gini variability index is to reflect the investigator's attitude toward variability. The higher ν, the more stress is put on the lower portion of the distribution of the independent variable. In the extreme case ($\nu \to \infty$) the investigator cares only about the lowest part of the cumulative distribution, as if he is guided by the max-min criterion. If $\nu = 1$ then the investigator measures variability according to Gini's mean difference, implying a symmetric weighting scheme around the median. If $\nu \to 0$, the investigator

does not care about variability; the range $-1 \leq \nu < 0$ reflects giving higher weights to the upper side of the distribution of the independent variable, while $\nu \rightarrow -1$ implies an investigator whose attitude to variability follows the max-max strategy, that is, caring about variability around the highest part of the distribution only. It is worth noting that when $-1 \leq \nu < 0$, the index of variability is negative. (See Donaldson and Weymark 1983; Yitzhaki 1983; and Chakravarty 1988, chap. 3, pp. 82-102, for description of the properties of the extended Gini index. Garner (1993), Lerman and Yitzhaki (1994), and Wodon and Yitzhaki (2002) are examples of its decomposition and use in welfare economics; see Araar and Duclos (2003) for a possible extension; see Davidson and Duclos (1997) for statistical inference, and Millimet and Slottje (2002) for an application in environmental economics). However, in the above-mentioned literature, the parameter is restricted to $\nu > 0$. Schechtman and Yitzhaki (1987, 1999, 2003) define and investigate the properties of the equivalents of the covariance and the correlation. The decomposition of the extended Gini of a sum of random variables into the contributions of the extended Gini's of the individual random variables, and the (equivalent of) correlations among them can also be found there. Olkin and Yitzhaki (1992) define the simple Gini regression coefficients and investigate their properties.

Definition and properties of Extended Gini Regression Coefficient (following proposition 3 in Yitzhaki 1996).

(a) The extended Gini regression coefficients (EGRC) are defined as:

$$\beta_{yx}(\nu) = \frac{-(\nu+1)\mathrm{COV}(Y,[1-F_x(X)]^{\nu})}{-(\nu+1)\mathrm{COV}(X,[1-F_x(X)]^{\nu})}, \quad \nu > -1;\ \nu \neq 0, \tag{1}$$

where ν is a parameter, determined by the investigator in order to determine the weighting scheme. (The factor $-(\nu+1)$ cancels out. It is presented here to emphasize that both numerator and denominator are based on extended Gini's). The subscript yx will be omitted in the simple regression case.

(b) Equation (2) presents the EGRC as a function of the derivatives of g(x) and shows how ν determines the weighting scheme. That is,

$$\beta(\nu) = \int W(x,\nu)\, g'(x)\, dx, \tag{2}$$

with $W(x,\nu) \geq 0$ and $\int_{-\infty}^{\infty} W(x,\nu)dx = 1$, where

Yitzhaki (1996) presents the weights for specific distributions of the independent

$$W(x,\nu) = \frac{[1-F_x(x)]-[1-F_x(x)]^{(\nu+1)}}{\int_{-\infty}^{\infty}[[1-F_x(t)]-[1-F_x(t)]^{(\nu+1)}]dt} \quad . \tag{3}$$

variable.

(c) Let $x_{[i]}$ be the i-th order statistic, and let y_i be the observation of y that accompanies $x_{[i]}$. Then, the estimators of β(ν), for all ν, can be expressed as weighted averages of slopes defined by pairs of adjacent observations:

$$b(\nu) = \sum_{i=1}^{n-1} w_i \, b_i \,, \tag{4}$$

where $b_i = \frac{y_{i+1} - y_i}{x_{[i+1]} - x_{[i]}}$ $(i = 1, ..., n-1)$; $w_i > 0$, $\Sigma w_i = 1$, and

$$w_i = w_i(x,\nu) = \frac{[n^{\nu}(n-i)-(n-i)^{\nu+1}]\Delta x_i}{\sum_{k=1}^{n-1}[n^{\nu}(n-k)-(n-k)^{\nu+1}]\Delta x_k}. \tag{5}$$

where $\Delta x_i = x_{[i]} - x_{[i-1]}$.

(d) The estimators b(ν) can be expressed as ratios of U-statistics. [Generally speaking, a U-statistic is an unbiased estimator for the parameter, which is created by forming an average of symmetric functions, called kernels. The interested reader is referred to Randles and Wolfe 1979, chap. 3, for details]. As such, b(ν) are consistent estimators of β (ν); for large samples, the distributions of the estimators converge to the normal distribution under regularity conditions.

(e) Suppose that $E(Y|X) = \alpha + \beta X$ and $Var(Y|X) = \sigma^2 < \infty$, then: (1) following property (b), β (ν) = β for all ν, and (2) all extended Gini estimators b(ν) (that is, regardless of ν) are consistent estimators of the same β.

Proofs: See Yitzhaki (1996), Proposition 3.

The properties above show that all EGRC can be expressed as weighted averages of slopes defined between adjacent observations, with the weighting scheme being determined by the ν attached to the independent variable, and by the distribution of the independent variable.

By changing ν and re-estimating the model, the investigator can learn about the curvature of the regression curve. The higher ν, the higher is the weight that is given to the slopes of the regression curve at the lower end of the range of the independent variable. In case the curve is linear, the estimates of the regression coefficient, for all ν,

do not differ significantly. If b(ν) turns out to be a declining (increasing) function of ν, then the regression curve is convex (concave). But of course, it may be that it does not show a specific pattern. Also, since it is based on all observations, it is clearly not able to detect small local deviations. (In the conclusions section, we offer an extension to deal with this issue as well). Note that unlike the case of using windows to estimate the slopes of the regression curves at different ranges, in the EG regressions all observations participate in all the regressions, and the only difference is in the weighting scheme applied.

Although it was not produced through an optimization, each extended Gini regression produces a normal equation. To see this, define

$$\varepsilon = \varepsilon(\nu) = Y - \alpha - \beta(\nu)X . \tag{6}$$

Then,

$$COV(\varepsilon, [1-F(X)]^{\nu}) = 0 \quad . \tag{7}$$

Equation (7) can be proved by plugging (6) and (1) into (7). By the same reasoning, it can be shown that the sample's version of Equation (7) holds.

We now move to define the constant term. Unlike other regression methods, the constant term is not estimated simultaneously with the regression coefficients, but follows it. Hence the constant term need not be selected according to the methodology used to derive the regression coefficients. The constant term depends on the function of the residuals that is being minimized. To do that one first derives an error term without taking into account the constant term. Minimizing the sum of squared deviations of the error term from a constant yields an estimated linear approximation, which passes through the means of the variables, while minimizing the sum of the absolute deviations of the error term from a constant forms a constant term so that the estimated linear approximation will pass through the medians, etc.

Finally, we would like to mention the following:

(a) The OLS can be presented in the same way, except that the weighting scheme is derived from the variance of the independent variable (Yitzhaki 1998).

(b) As far as we know, the EG regression is the only regression method in which the investigator controls the weights attached to each part of the distribution of the *independent variable*. For example, under quantile regression (Koenker and Bassett 1978), the weighting scheme is applied to the residual e_i. (To see this, note that $e_i = Y_i -$

$\hat{\alpha} - \hat{\beta} X_i$ is the residual, and hence the optimization must be applied to a function of the residuals. Therefore, the quantile is a quantile of the residual).

(c) Numerically, the extended Gini regression coefficient, when $\nu = 1$, is identical to the Durbin's (1954) suggested estimator, which is based on using the rank of the independent variable as an instrumental variable. In the sample, the rank is given by the empirical distribution multiplied by the sample size, hence the identity. However, the motivation, the distribution of the estimators, and other properties are totally different. One can take advantage of this analogy and extend it to all EGRCs, and use it to calculate the estimates (but not the standard errors) of the EGRC using standard regression software. Note, however, that this interpretation does not apply without qualification to the multiple regression case (as discussed at the end of the next section).

SECTION 2: THE MULTIPLE REGRESSION CASE

The target of this section is to develop an extension of the simple regression coefficients suggested in Yitzhaki (1996) to the multiple regression case.

Let $(Y,X_1,\ldots,X_K)$ be a (K+1)-variate random variable with expected values $(\mu_Y, \mu_1, \ldots, \mu_K)$ and a finite variance-covariance matrix Σ. Let $g(x) = E\{Y|X_1 = x_1, \ldots, X_K = x_K\}$ be the regression curve. Similar to the simple regression case, the error term at $(Yi, X_{1i}, X_{2i},\ldots,X_{ki})$ is defined as the deviation of Yi from the linear approximation $\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$, i.e., $\varepsilon_i = Y_i - \alpha - \beta_1 X_1 - \ldots - \beta_k X_k$. Again, no assumptions are imposed on the error term, and the regression curve need not be a linear function of the independent variables.

As before, an investigator is interested in estimating a linear approximation of the regression curve. Consider a first order Taylor expansion around zero of the regression curve. By construction, the expansion is linear.

The slopes of the linear approximation can be written as:

$$\begin{pmatrix} \frac{dy}{dx_1} \\ \frac{dy}{dx_i} \\ \frac{dy}{dx_K} \end{pmatrix} = \begin{pmatrix} \frac{\partial g}{\partial x_1} + & \frac{\partial g}{\partial x_2}\frac{dx_2}{dx_1} + & \ldots + & \frac{\partial g}{\partial x_K}\frac{dx_K}{dx_1} \\ \ldots & \ldots & \ldots & \ldots \\ \frac{\partial g}{\partial x_1}\frac{dx_1}{dx_i} + & \ldots & \ldots + & \frac{\partial g}{\partial x_K}\frac{dx_K}{dx_i} \\ \frac{\partial g}{\partial x_1}\frac{dx_1}{dx_K} + & \frac{\partial g}{\partial x_2}\frac{dx_2}{dx_K} + & \ldots + & \frac{\partial g}{\partial x_K} \end{pmatrix} \tag{8}$$

Using the simple regression coefficients developed in Section 1 to represent the simple slopes in the Taylor expansion implies that $\frac{dy}{dx_k} = \beta_{0k}(\nu_k)$ and $\frac{dx_j}{dx_k} = \beta_{jk}(\nu_k)$, where the subscript 0 refers to the dependent variable, and k=1,…, K, indicate the independent variables. Having done that, we can rewrite (8) as:

$$\begin{pmatrix} \beta_{01}(\nu_1) \\ \dots \\ \beta_{0i}(\nu_i) \\ \beta_{0K}(\nu_K) \end{pmatrix} = \begin{pmatrix} 1 & \beta_{21}(\nu_1) & \dots & \beta_{K1}(\nu_1) \\ & & & \\ & & & \\ \beta_{1K}(\nu_K) & \beta_{2K}(\nu_K) & \dots & 1 \end{pmatrix} \begin{pmatrix} \frac{\partial g}{\partial x_1} \\ \frac{\partial g}{\partial x_i} \\ \\ \frac{\partial g}{\partial x_K} \end{pmatrix}. \tag{9}$$

Using Equation (9) one can solve for the estimators of the partial derivatives $\frac{\partial g}{\partial x_k}$ which will be equal to:

$$\begin{pmatrix} \frac{\partial g}{\partial x_1} \\ \frac{\partial g}{\partial x_i} \\ \\ \frac{\partial g}{\partial x_K} \end{pmatrix} = \begin{pmatrix} 1 & \beta_{21}(\nu_1) & \dots & \beta_{K1}(\nu_1) \\ & & & \\ & & & \\ \beta_{1K}(\nu_K) & \beta_{2K}(\nu_K) & \dots & 1 \end{pmatrix}^{-1} \begin{pmatrix} \beta_{01}(\nu_1) \\ \dots \\ \beta_{0i}(\nu_i) \\ \beta_{0K}(\nu_K) \end{pmatrix}. \tag{10}$$

Note that in (10), the vector on the right-hand side depends on all the ν_i. Also, the denominator of each row k in the matrix (before inverting it) is $G(X_k, \nu_k) = G_k(X) = -(\nu_k + 1)\,\mathrm{COV}(X_k, [1 - F_k(X)]^{\nu_k})$ (i.e. the denominator in each row is the extended Gini of the appropriate variable).

We refer to the estimators as *implied* partial derivatives because we do not argue that they represent the derivatives at a given point, but if one accepts the notion of a linear approximation, and accepts simple regression coefficients as representing weighted averages differentials, then for consistency, one has to accept the partial regression coefficients as the solution offered in (10).

If (8) represents slopes of a truly linear model, then all the coefficients at the right-hand side of (10) are constants, and the left hand side must represent by construction the partial derivative of the regression curve. On the other hand, if the regression curve is not linear, then by changing ν_i, one can trace the change in $\partial g / \partial x_i$, other things being equal, by changing the weighting scheme attached to the slopes of

variable i. Note that by other things being equal, it is meant that all rows, except row *i* in the matrix of regression coefficients in (10), remain unaffected, and all elements in the vector of simple regression coefficients of the dependent variable on the independent variables, except element *i* do not change. This is a unique property of the EG, which is due to the fact that there are two covariances and two correlations between each pair of random variables (see comment at the end of this section). Therefore, β_{ij} can be changed without affecting β_{ji}. We will discuss further adjustments of the estimators to represent changing derivatives along the range of the independent variable in the conclusions section.

Since we are allowed to multiply each row by a constant (in our case, the constant is the extended Gini of the independent variable), the matrix can be presented in a way, which is similar to the variance-covariance matrix in OLS, with Gini's and co-Gini's replacing the variances and covariances, respectively.

Like the simple regression, the multiple regression procedure, although it is not based on an optimization procedure, generates equivalents to the OLS's normal equations. By defining the error term, and substituting for the multiple regression coefficients, it can be shown that

$$COV(\varepsilon, [1-F_k(X)]^{\nu_k}) = 0 \text{ for k=1,…, K.} \qquad (11)$$

The first step in this section was to define the linear approximation of the regression in the population. We now move to the estimation procedure. Before we turn to the estimation stage, we shall rewrite the parameters in matrix notation. Consider Y as a dependent variable and let X_1,..., X_K be the independent variables. Let V be a (nxK) matrix of power functions of (one minus) the cumulative distributions of X_1,..., X_K (in deviations from their expected values), multiplied by $-(\nu_k+1)$. That is, a typical element in V is

$V_{ik} = -(\nu_k + 1)\{[1-F_k(x_{ik})]^{\nu_K} - \frac{1}{\nu_k + 1}\}$. The vector of regression coefficients, $\beta(\nu)$, is defined by

$$\beta(\nu) = [V'X]^{-1} V'Y, \qquad (12)$$

where $\beta(\nu) = \{\beta_1(\nu_1),..., \beta_K(\nu_K)\}$ is a (K x 1) column vector, V is an (n x K) matrix defined above, Y is an (n x 1) column vector of the dependent variable, and X is an (n x K) matrix of the deviations of the independent variables from their expected values. The vector V'Y is a column vector, the elements of which are

$-(\nu_k+1)COV(Y,[1-F(X_k)]^{\nu_k})$, that is, the EG covariances between the dependent variable and the independent variables. The matrix A = V'X is a matrix with the elements $-(\nu_k+1)COV(X_j,[1-F(X_k)]^{\nu_k})$, that is - its elements are EG equivalents of variances and covariances of the independent variables: the diagonal elements are the EG's of the independent variables, while the off-diagonal elements are EG - covariances between pairs of independent variables. It is assumed that rank [V'X] equals K, the number of independent variables. The requirement that one can invert the matrix [V'X] implies a restriction on the choice of the independent variables that does not exist in OLS. If the ν's assigned to different independent variables are all equal, then no independent variable can be a monotonic transformation of another independent variable, because it will imply identical rows in the matrix (which depends on X via F(X)). However, if the ν's are different, then one independent variable can be a monotonic transformation of another independent variable without causing colinearity in the [V'X] matrix. For example, X and e^X cannot appear simultaneously as independent variables in EG regressions if they have the same ν, but they can participate with different ν's.

We now turn to the estimation step. First one estimates the regression coefficient, and given the estimated regression coefficient, one moves to estimate the constant term. The natural estimators of the regression coefficients are based on replacing the cumulative distributions by the empirical distributions (which are calculated using ranks):

$$b(\nu) = [v'x]^{-1}v'y \quad , \qquad (12')$$

where v is a matrix with elements $-(\nu_k+1)[n^{-\nu_k}(n-r(x_{i_k}))^{\nu_k}-1/(\nu_k+1)]$, and $r(x_{ik})$ is the rank of x_{ik} among $x_{1k},\ldots,x_{nk}$. As in the simple regression case, an orthogonality condition is satisfied as given in the following lemma:

Lemma 2.1: Define the vector $\varepsilon(\nu) = Y - X\beta(\nu)$. Then, $V'\varepsilon(\nu) = o$ where o is a vector of zeros.

Proof: $V'\varepsilon(\nu) = V'Y - V'X\beta(\nu) = V'Y - V'X[V'X]^{-1}V'Y = o$.

This property holds in the sample as $v'e(\nu) = o$, where $e(\nu) = y - x\,b(\nu)$.

As pointed out in the simple regression case, $b(\nu)=\{b_1(\nu_1),..,b_K(\nu_K)\}$ is identical to an instrumental variable estimator in an OLS regression, with a power function of rank X serving as the instrumental variable. This interpretation provides an alternative method for calculating b(ν) by using any standard regression software. Note, however,

that (a) since no assumptions are imposed on the model, the sampling distribution of b(ν) still needs to be investigated for the present setup (as detailed in Section 3), and (b) each independent variable is substituted by only one instrumental variable, which makes the reliance on a standard regression package problematic.

It is worth emphasizing an important property that is unique to the EG multiple regression. The EG has two correlations defined between each pair of variables so that in contrast to the OLS, a symmetric correlation is not imposed on the independent variables. When the parameter ν of one independent variable is changed, only one of the two correlations defined with any other independent variable is affected. This property allows us to refer to the partial regression coefficients as representing partial derivatives because changing ν for one independent variable may only change its own slope, and one set of the asymmetric EG correlation coefficients that accompany it.

The constant term will be estimated in a way, which is identical to the simple regression case. That is, the investigator can minimize a function of the error terms. The exact function used determines whether the regression passes through the mean or the median. Section 3 derives the asymptotic properties of the estimators. Readers who are mainly interested in the application can skip Section 3.

SECTION 3: THE ASYMPTOTIC BEHAVIOR OF ESTIMATORS

For simplicity, the presentation of the asymptotic properties will be restricted to the two independent variables case, with different values of ν. All the results can be extended to the K-variable case.

A natural way to estimate the regression coefficients is based on replacing the cumulative distributions by the empirical distributions.

Let the estimated equation be:

$$y = b(\nu_1, \nu_2)_{01.2}\, x_1 + b(\nu_1, \nu_2)_{02.1}\, x_2 + e(\nu_1, \nu_2) \quad , \qquad (13)$$

where $b_{0i.j}$ represents the nonparametric EG regression coefficient of the implied partial effect of X_i on the conditional mean of Y, given that X_j is in the model, but held fixed. The constant term is not needed for estimating (13) and it will be estimated later, according to the requirements on where the linear approximation should pass (mean, median, quantile). For simplicity of presentation we will omit, whenever possible, the parameters ν_1 and ν_2, keeping in mind that the partial regression coefficients, b, are functions of those parameters. The vector b, expressed in matrix notation in (12'), can be explicitly written as (for K=2):

$$b = b(\nu_1, \nu_2) = \begin{pmatrix} b_{01.2} \\ b_{02.1} \end{pmatrix} = \frac{1}{D} \begin{pmatrix} c_{22}c_{01} - c_{21}c_{02} \\ c_{11}c_{02} - c_{01}c_{12} \end{pmatrix} \quad (14)$$

where $c_{0i} = -(\nu_i+1)\ \mathrm{cov}(y, [1-r_i/n]^{\nu_i})$, (i=1,2); $c_{ij} = -(\nu_j+1)\ \mathrm{cov}(x_i, [1-r_j/n]^{\nu_j})$; $D = c_{11}c_{22} - c_{12}c_{21}$ and r_i is the vector of ranks of x_i.

In what follows, we introduce an alternative estimator of β, based on functions of U-statistics. We show that its limiting distribution is normal, and that the differences between the b of (14) and the estimators based on U-statistics are negligible; hence, the limiting distribution of b is also normal. The advantages of using the U-statistics as estimators are that they provide unbiased estimators of the individual parameters, they have minimum variance among all unbiased estimators, and their asymptotic distributions are well known (Randles and Wolfe 1979, chap. 3). Also, the functions of U-statistics are consistent estimators of the respective parameters, and their limiting distributions are normal (under regularity conditions). The structure of this section is as follows: first, a kernel is found for the relevant parameters (defined below). Then, a U-statistic based on the kernel is obtained for each of the parameters. Using the above U-statistics, an estimator b_U is suggested for β, based on a function of dependent U-statistics, and it is shown that it is a consistent estimator for β; while the next step is to find its asymptotic distribution. This is done using asymptotic results from U-statistics theory. Finally, it is shown that $\sqrt{n}\, b_U$ and $\sqrt{n}\, b$ have the same limiting distribution. Once the asymptotic distribution is known to be normal, and the asymptotic variance can be calculated using jackknife, for example- (see Shao and Tu 1996), inference can be drawn (confidence intervals and hypothesis tests) for each parameter.

Let

$$\theta = -(\nu+1)\ \mathrm{COV}(Y, [1-F(X)]^{\nu}) . \quad (15)$$

The parameter θ is the EG equivalent of the covariance between Y and X, and is a typical element of V'Y (see (12)) and of V'X (when X replaces Y).

The following theorems are proved only for the case where ν is an integer.

Theorem 3.1: Let $(X_1,Y_1),\ldots,(X_{\nu+1},Y_{\nu+1})$ be a random sample of size (ν+1) from a continuous bi-variate distribution $F_{X,Y}$ with finite second moments. Let

$$h((x_1, y_1),\ldots,(x_{\nu+1}, y_{\nu+1})) = \bar{y}_{\nu+1} - y_{x_{(1)}} \quad , \quad (16)$$

where $\bar{y}_{\nu+1}$ is the average of $y_1,..,y_{\nu+1}$ and $y_{x_{(1)}}$ is the y that belongs to $x_{(1)}$, the minimum of $x_1,...,x_{\nu+1}$. Then $h((x_1,y_1),...,(x_{\nu+1},y_{\nu+1}))$ is a symmetric kernel of degree $\nu+1$ for the parameter θ of (15). That is - $h((x_1,y_1),...,(x_{\nu+1},y_{\nu+1}))$ is an unbiased estimator of the parameter θ, based on $\nu+1$ observations. (See Randles and Wolfe 1979, p. 61).

Proof: The parameter θ of (15) can be expressed as follows:

$$\theta = -(\nu+1)\,COV(Y, [1-F(X)]^{\nu}) = (\nu+1)\,E\{Y\}\,E\{[1-F(X)]^{\nu}\} - (\nu+1)\,E\{Y\,[1-F(X)]^{\nu}\}$$

$$= \mu_Y - (\nu+1)\,E\{Y\,[1-F(X)]^{\nu}\}\,.$$

Therefore, we need to show that $E\{Y_{X_{(1)}}\} = (\nu+1)\,E\{Y\,[1-F(X)]^{\nu}\}$.

Claim: $E\{Y_{X_{(1)}}|X_{(1)} = x\} = E\{Y|X = x\}$.

The proof of the claim is restricted to the discrete case.

Proof of the Claim:

$$E\{Y_{X_{(1)}}|\,X_{(1)} = x\} = \sum_{i=1}^{\nu+1} y_i\,P(Y_{X_{(1)}} = y_i \mid X_{(1)} = x) =$$

$$= \sum_{j=1}^{\nu+1}\sum_{i=1}^{\nu+1} y_i\,P(Y_{X_{(1)}} = y_i|\,X_{(1)} = x,\, X_j = X_{(1)})\,P(X_j = X_{(1)}) =$$

$$= \sum_{j=1}^{\nu+1}\sum_{i=1}^{\nu+1} y_i\,P(Y_j = y_i|\,X_j = x)\,1/(\nu+1) =$$

$$\sum_{j=1}^{\nu+1} E\,(Y_j \mid X_j = x)\,1/(\nu+1) = E(Y|X=x)\,.$$

Using the claim,

$$E\{Y_{X_{(1)}}\} = E_{X_{(1)}}\{\,E\,(Y_{X_{(1)}}|\,X_{(1)} = x\,\} =$$

$$\int E(Y_{X_{(1)}}|X_{(1)}=x)\,f_{X_{(1)}}(x)\,dx = (\nu+1)\int E(Y_{X_{(1)}}|X_{(1)}=x)\,[1-F(x)]^{\nu}f(x)\,dx$$

$$= (\nu+1)\int E(Y|X=x)\,[1-F(x)]^{\nu}f(x)\,dx$$

$$= (\nu+1)\iint y\,f(y|x)\,[1-F(x)]^{\nu}f(x)\,dy\,dx$$

$$= (\nu+1)\iint y\,[1-F(x)]^{\nu}f(x,y)\,dy\,dx = (\nu+1)\,E\{\,Y\,[1-F(X)]^{\nu}\,\}\,.$$

The symmetry of $h((x_1,y_1),...,(x_{\nu+1},y_{\nu+1}))$ is obvious. QED.

Let $h((x_1,y_1),...,(x_{\nu+1},y_{\nu+1})) = \bar{y}_{\nu+1} - y_{x(1)}$, as in (16) and let

$$U = \frac{1}{\binom{n}{\nu+1}} \sum_{i_1<} \sum_{\ldots<} \sum_{i_{\nu+1}} h((X_{i_1}, Y_{i_1}),\ldots,(X_{i_{\nu+1}}, Y_{i_{\nu+1}}))$$

$$= \frac{1}{\binom{n}{\nu+1}} \sum_{i_1<\ldots<i_{\nu+1}} \cdots \sum \left(\frac{\sum_{j=1}^{\nu+1} Y_{i_j}}{\nu+1} - Y_{\min(X_{i_1},\ldots,X_{i_{\nu+1}})} \right),$$

where $(i_1,\ldots,i_{\nu+1})$ is a permutation of $(\nu+1)$ indices chosen from $(1,\ldots,n)$. Then, U is a U-statistic for the parameter θ, and is therefore an unbiased and consistent estimator of θ (Randles and Wolfe 1979, corollary 3.2.5).

Using combinatorial arguments, U can be simplified and written as a linear combination of concomitants of the order statistics as follows:

$$U = \frac{1}{\binom{n}{\nu+1}} \sum_{i=1}^{n} \left[\frac{1}{\nu+1} \binom{n-1}{\nu} - \binom{n-i}{\nu} \right] Y_{X_{(i)}} \quad . \tag{17}$$

(Note that if $\nu > (n-i)$, then $\binom{n-i}{\nu} = 0$).

Following theorem 3.1, all the elements of (12) can be estimated by U-statistics and hence, for k=2, β can be estimated by a vector of size 2, whose elements are functions of several (dependent) U-statistics.

Let

$$\mu' = (\mu_1, \mu_2,\ldots, \mu_6) = (C_{01}, C_{02}, C_{11}, C_{12}, C_{21}, C_{22})$$

be the vector of the parameters, where $C_{0i} = COV\ (Y, [1-F(X_i)]^{\nu_i})$ and

$C_{ij} = COV\ (X_i, [1-F(X_j)]^{\nu_j})$, and let

$$U' = (U_1, U_2,\ldots,U_6) = (c_{01}, c_{02}, c_{11}, c_{12}, c_{21}, c_{22})$$

be the corresponding vector of U-statistics (whose elements are given in equation 14). The estimator of β, based on functions of U-statistics, can be written as:

$$b_U = \begin{pmatrix} b_{U01.2} \\ b_{U02.1} \end{pmatrix} = \frac{1}{D} \begin{pmatrix} U_{22}\,U_{01} - U_{21}\,U_{02} \\ U_{11}\,U_{02} - U_{01}\,U_{12} \end{pmatrix} \tag{18}$$

where U_{0i} is the U-statistic based on the kernel $\bar{y}_{\nu_i+1} - y_{x_i(1)}$, (i=1,2), and U_{ij} is the U-

statistic based on the kernel $\bar{x}_{i,\nu_j+1} - x_{i,x_j(1)}$, where $\bar{x}_{i,\nu_j+1}$ is the average of (ν_j+1) observations of X_i, and $x_{i,x_j(1)}$ is the value of X_i, which belongs to the smallest value of X_j (out of the ν_j+1 values), and $D = U_{11} U_{22} - U_{12} U_{21}$.

Using the above notation, we can obtain the following results:

Theorem 3.2

Let $(y_i, x_{i1}, x_{i2}, i=1,2,..,n)$ be a sample drawn from a continuous multivariate distribution with finite second moments and such that $\mu_3 \mu_6 - \mu_4 \mu_5 \neq 0$. Then, b_U in (18) is a consistent estimator of β in (12) (i.e., each component of b_U is a consistent estimator of the respective element of β).

Proof: Since each U-statistic converges in quadratic mean, and thus in probability, to the parameter it estimates, it follows by Slutzky's theorem (Randles and Wolfe 1979, Theorem A.3.1.3) that

$$\frac{U_6 U_1 - U_5 U_2}{U_3 U_6 - U_4 U_5} \text{ converges in probability to } \frac{\mu_6 \mu_1 - \mu_5 \mu_2}{\mu_3 \mu_6 - \mu_4 \mu_5}$$

and thus the former is a consistent estimator of the latter. QED.

Theorem 3.3, due to Hoeffding (1948, Theorem 7.1) and Theorem 3.4, due to Serfling (1980, Theorem 3.3.A) are needed for the derivation of the asymptotic distribution of b_U.

Theorem 3.3

Under the assumptions of Theorem 3.2, the vector U has an asymptotic normal distribution with mean μ and a variance-covariance matrix $d_n^2\Sigma$, where $d_n = 2/(\sqrt{n})$. That is, $\sqrt{n}\,(U-\mu) \xrightarrow{D} N(0, 4\Sigma)$.

Theorem 3.4

Let $U_n = (U_{1n},..,U_{6n})$ be asymptotically normally distributed with mean vector μ and a variance-covariance matrix Σ. Let $g(U) = (g_1(U), g_2(U))$ be a vector-valued function for which each component function $g_i(U)$ is real-valued and has a nonzero differential $g(\mu;t)$, $t =(t_1,..,t_6)$ at $U=\mu$. Let

$$M = [\,(\frac{\partial g_i}{\partial U_j})_{U=\mu}\,] \quad i = 1,2;\ j = 1,..,6 \quad .$$

Then $b_U = g(U_n)$ is $AN(g(\mu)\ ,\ d_n^2 M\,\Sigma M')$, where AN is asymptotic normal and $d_n \to 0$ as $n \to \infty$.

Proof:

By Theorem 3.3 the vector U is AN(μ , $d_n^2\Sigma$), with Σ a variance-covariance matrix and $d_n = 2/(\sqrt{n}) \to 0$ as $n \to \infty$. Let

$$g(U) = \begin{pmatrix} g_1(U) \\ g_2(U) \end{pmatrix} = \frac{1}{D} \begin{pmatrix} U_6 U_1 - U_5 U_2 \\ U_3 U_2 - U_1 U_4 \end{pmatrix},$$

where $D = U_3U_6 - U_4U_5$, is the vector-valued function. The proof follows Serfling (1980) Theorem 3.3.A. The explicit form of Σ is omitted since it is complicated. (The matrix Σ involves variances of the individual U-statistics, as well as covariances between each pair of U-statistics. Using Slutzky's Theorem, a consistent estimate of Σ can be obtained by replacing each parameter by its consistent estimate). A practical way to estimate it, using jackknife, is given in Yitzhaki (1991). Theorem 3.4 states that $\sqrt{n}\, b_U$ is asymptotically normal. In order to obtain the same result for $\sqrt{n}\, b$, the estimator based on replacing the cumulative distribution by the empirical one, it is required to show that $\sqrt{n}(b-b_U) \to 0$ as $n \to \infty$. This is shown in the Appendix.

SECTION 4: AN APPLICATION: WHO DOES NOT RESPOND TO QUESTIONNAIRES?

Surveys suffer from nonreporting, even if refusal to respond is illegal, as is the case in official surveys in Israel. If nonreporting is correlated with income, then the estimates of the mean income and the index of income inequality may be biased. Nonreporting can occur for various reasons; some of them depend on the individual (refusal, not-at-home, etc.), while others may be due to problems at the collecting agency (the interviewer did not find the dwelling, did not approach the respondent at a convenient time, errors and omissions at the agency, etc.). In this paper we do not investigate the causes of nonresponse. We will be interested in describing it as a function of several demographic variables (which can be used later in designing the sample) and one major variable, income. In general, the experience concerning nonreporting is that the propensity not to respond is a U-shaped function with respect to income, because the rich tend not to participate, while the poor and the young can not be found easily at home. A recent study by Mistiaen and Ravallion (2003) presents a model in which compliance can either decrease or increase with income, and also be of an inverted U-shape. Moreover, adding other arguments such as the ability to find the members of

the households at home, finding the address, viewing participation as a democratic value, etc., can lead to almost all kinds of patterns. Mistiaen and Ravallion (2003) find that the nonresponse problem is not ignorable, and that there is a highly negative significant income effect on compliance. Deaton (2003) raises the plausible conjecture that richer households are less likely to participate in surveys, in order to explain the gap between growth estimates based on households' surveys and those that are based on national accounts. (Comprehensive studies, dealing with almost all aspects of nonresponse are detailed in Groves and Couper (1998) and Groves, Dillman, Eltinge and Little (2002)). The main conclusion from reading the literature is that nonresponse is a serious issue that may bias the estimates, but we do not have enough knowledge to justify making the assumptions needed for running OLS or other parametric regressions.

Investigating the magnitude and the effect of nonresponse on the results is a bit complicated since one is dealing with missing observations. The direct way to learn about the problem is to analyze the properties of non-respondents from the scatter information known about them, like the location of the dwelling, and other direct or indirect information that can be taken from the files used for the sampling. Such an approach suffers from two major problems: (a) the information that one can gather is not sufficient (b) the response rate in the sample we are dealing with is around ninety percent, so that the sample of nonresponse is relatively small. The main idea in this empirical illustration is to use the sample of the respondents to learn about the effect of nonresponse. For this purpose, we rely on a common procedure used in many official statistical offices.

To overcome biases that are caused by the sample being a nonrepresentative one and to reduce standard errors, many statistical agencies adjust the distribution of the sample to fit known marginal distributions of current demographic estimates that are based on the census. The outcome of this adjustment is a weighting scheme: a weight is attached to each observation. (A necessary condition to be able to perform such an adjustment is having a detailed census data. Also, there are other reasons for using those procedures, among them is to insure that different samples, performed by different units of the agency, report the same demographic structure so that official statistics will not be blamed by the media of publishing contradicting estimates. This may explain why the adjustment to given margins is performed mainly by producers of official statistics). For a survey of the different methodologies used to construct weighting schemes, see the

survey by Kalton and Flores-Cervantes (2003). A detailed description of the method used in Israel is offered in Kantorowitz (2002). For the purpose of this paper, it is sufficient to say that the above-mentioned procedures change the weight of each observation, so that it adds up to given marginal demographic and geographic distributions.

The sample we are dealing with is a sample of dwellings. It is a stratified sample according to geographical areas and types of dwelling, but the probability of each dwelling to be included in the sample is the same. Since the probability of each dwelling to be included in the sample is the same, the expected value of the weight of each observation is equal to the ratio of the overall population to the sample size. When nonresponse occurs in a certain group, it will be underrepresented in the sample, so that the weight that will be assigned to those who responded in that group will be higher than its expected value in case of equal tendency to respond.

From the point of view of our investigation, the important fact is that if the propensity of nonresponse is equal among all potential respondents, then the expected weights of all observations will be equal. Moreover, the higher the nonresponse, the higher the weight assigned to this type of observation. Hence, the weight attached to an observation can serve as an indicator of nonresponse, and will serve as the dependent variable in our analysis. If nonreporting is random, then we should expect the weight to be uncorrelated with other characteristics of the population. If the slope of the regression curve of weight on income is positive then we conclude that nonreporting increases with income. If it is positive, but declining when high incomes are stressed, then we conclude that nonreporting increases with income but the propensity not to respond declines with income.

The weighting scheme of the sample is produced by an algorithm for calibration, with several hundreds of constraints imposed, and is intended to make the sample representative (Kantorowitz 2002). In particular, a constraint is imposed on the maximal weight assigned to each observation, so that standard errors do not increase unnecessarily. The constraints insure that the reported age structure, geographic distributions, household sizes will add up to given margins of the distributions of the population.

In some sense our purpose is to summarize the effect of the several hundreds of constraints imposed on weights to add up to given demographic margins, into the effect on the variable of interest, which is income. It is important to note that the income is not

involved at all in the derivation of the weights. Hence, there is no built-in correlation (i.e., spurious correlation) between the weight of each observation and income.

The survey of household expenditures in Israel is conducted every year since 1997. Since in some years observations from East Jerusalem were missing, we have omitted those observations from all years to get an identical coverage. The probability of each household to be included in the sample is equal. Hence, if the propensity of the population to respond and the surveying process are not correlated with demographic properties, then the expected values of all weights should be equal. The sum of the weights represents the whole population of the country. Since the size of the sample relative to the overall population changes over the years, the average value of the weights and the slope of the regression can change between years. (It is as if one multiplies the weights in each year by a different constant). We did not correct for this problem, since its effect is small and it is important only when comparing results from different years. For the sake of simplicity, we preferred to concentrate and present the results for the last available year, and checked whether the main conclusions reached are sensitive to the selected year.

Table 1 provides descriptive statistics of the weights according to different years and ethnic groupings. The population is divided into three groups: Two minority groups - Arabs and ultra religious Jews (an ultra religious household is defined according to the school attended by the head of the household), and the rest – referred to as the majority group. The reason for this distinction is that experienced enumerators reported that those groups tend to have different patterns of nonresponse.

In general, one can observe from Table 1 that for all years the average weight of the Arab population is the lowest, meaning that they have fewer cases of nonresponse. The maximum weight attached to an observation should be viewed with caution because some of the programs assigning the weights may restrict the weight not to be greater than a certain value. It is worth noting that although the order of average weights according to the groups remains the same over the years (except for 1999), the order of the standard deviations of the weights changes.

Table 1. Descriptive Statistics of Household Weights by Ethnic Grouping.

Year	Ethnic Group	n	Weight			
			Average	Max	Min	Std.Dev.
1997	Majority	4,942	283	1,196	20	166
	Arabs	529	267	1,051	19	199
	Ultra religious Jews	90	398	1,127	45	245
1998	Majority	5,068	286	1,196	18	169
	Arabs	606	256	1,049	24	176
	Ultra religious Jews	98	321	766	26	166
1999	Majority	5,114	291	1,134	13	154
	Arabs	597	269	1,129	14	169
	Ultra religious Jews	105	292	639	20	115
2000	Majority	5,146	301	1,195	22	170
	Arabs	629	260	959	43	142
	Ultra religious Jews	89	310	1,017	33	148
2001	Majority	5,049	314	1,185	18	152
	Arabs	662	285	1,902	31	171
	Ultra religious Jews	76	341	834	104	145

NOTE: Source: HES 1997-2001, excluding the observations of East Jerusalem in 1997-1999.

Since the groups differ in household size, which may affect the probability of finding someone at home, Table 2 presents the average weights according to household size. It can be seen that for the majority, household of size 1 has the highest weight, and the rest are similar (year 2001 is different). This may be a result of small households not being at home while the elderly, although being at home, do not have the patience to complete the questionnaire. (For Arabs, there is no obvious pattern. Nothing can be said about the ultra religious Jews, since the sample sizes are quite small).

Table 2. Mean of Household Weights by Ethnic Grouping and Household Size.

Year	Ethnic Group		Household Size 1	2	3	4	5+
	Majority	Mean	325	278	278	276	266
		N	*840*	*1,199*	*807*	*910*	*1,186*
1997	Arabs	Mean	288	246	255	236	281
		N	*12*	*48*	*54*	*102*	*313*
	Ultra religious Jews	Mean	273	363	294	315	488
		N	*2*	*18*	*16*	*13*	*41*
	Majority	Mean	330	280	290	276	268
		N	*869*	*1,281*	*788*	*964*	*1,166*
1998	Arabs	Mean	316	338	241	252	247
		N	*13*	*49*	*70*	*98*	*376*
	Ultra religious Jews	Mean	235	247	385	329	325
		N	*3*	*12*	*13*	*12*	*58*
	Majority	Mean	333	291	276	297	266
		N	*892*	*1258*	*878*	*919*	*1167*
1999	Arabs	Mean	230	248	321	302	256
		N	*15*	*49*	*64*	*93*	*376*
	Ultra religious Jews	Mean	176	293	291	333	288
		N	*3*	*20*	*11*	*12*	*59*
	Majority	Mean	351	291	310	280	282
		N	*875*	*1,296*	*867*	*965*	*1,143*
2000	Arabs	Mean	323	223	272	381	237
		N	*17*	*58*	*64*	*78*	*412*
	Ultra religious Jews	Mean	239	327	254	336	301
		N	*3*	*21*	*1*	*13*	*51*
	Majority	Mean	363	294	326	291	308
		N	*886*	*1,325*	*838*	*949*	*1,051*
2001	Arabs	Mean	521	261	290	304	271
		N	*20*	*64*	*74*	*116*	*388*
	Ultra religious Jews	Mean	393	346	483	366	319
		N	*3*	*16*	*4*	*8*	*45*

NOTE: Source: HES 1997-2001, excluding the observations of East Jerusalem in 1997-1999.

To summarize: The dependent variable is the weight assigned to each observation by a calibration procedure, intended to represent the entire population. The sample is a stratified sample, but the probability of each dwelling and each person living in a dwelling to be included in the sample is equal. Hence, if the propensity not to be included in the sample is equal, either because of nonresponse or errors on behalf of the agency, the expected weight assigned to each observation should be equal. It may differ between years, if the ratio of sample size to population changes between years. The

weight is treated as an indicator of nonresponse. Having described the dependent variable, we now move to describe the results concerning the regression coefficients.

SECTION 5: EMPIRICAL RESULTS

We turn first to simple regression coefficients. Those simple regression coefficients are used in finance (Gregory-Allen and Shalit 1999; Shalit and Yitzhaki 2002), while a variation of them has been used for almost twenty years in analyzing the income elasticity of consumption goods (see the survey in Wodon and Yitzhaki 2002). We have estimated the regression coefficients for all the years. Since they present a stable picture, only the results for the year 2001 are presented here.

Table 3. Regression Coefficients of Household Weight on Gross Income per Household, by Gini Parameter (ν).

	ν for Gross Income:			
Coefficient	3	2	1	-0.5
b	-0.0025*	-0.0022*	-0.0017*	-0.0007*
SE(b)	*0.0003*	*0.0003*	*0.0002*	*0.0001*
a (mean)	348.3	342.8	335.7	321.0
a (median)	320.7	314.9	307.8	293.2

NOTE: Source: HES 2001.

* indicates a value significantly different than 0 (at α=0.05).

Table 3 presents the regression coefficients of weight on household income, for different values of ν. The higher the parameter ν, the more the regression stresses the slopes of the regression curve at the lower end of the income distribution. (It is worth noting that those weights are solely determined by the distribution of the independent variable, and ν. They should not be confused with the dependent variable of the regression, which is the weight assigned to an observation). As can be seen, the regression coefficients are negative, which means that the higher the income - the lower the weight assigned to observations, implying that nonresponse declines with income. Even when high-income groups are stressed (ν = -0.5) we still have a significant negative regression coefficient. The interpretation of this finding is that we have a monotonic relationship between nonresponse and income. We have checked the pattern for the years 1997–2000 and found the same pattern. To save space the results are not presented.

The rest of Table 3 presents the two versions of the constant term. One presents the constant term when the regression line passes through the median while the other -

through the mean. It is interesting to note that the difference between the two is around 28 with a(mean) higher than a(median). This is an indication that the error term tends to be asymmetric. It is not clear to us whether this kind of result, i.e. that the difference in the constant terms is independent of the slopes is a coincidence or it is a property of the extended Gini regression procedure.

Table 4 presents the simple regression coefficients of weight on household size. As in the regression on income, the larger the family size the higher the value of the regression coefficient, and in all cases, the signs of the regression coefficients are negative. This means that nonresponse is higher among small households. Note that as before, the constant term of the regression passing through the mean is larger than the constant term of the regression passing through the median, but again, the difference between the two constants is around 28.

Table 4. Regression Coefficients of Household Weight on Household Size, by Gini Parameter (ν).

Coefficient	ν for Household Size:			
	3.0	2.0	1.0	-0.5
b	-12.2*	-10.6*	-8.5*	-3.5*
SE(b)	*1.4*	*1.3*	*1.1*	*1.2*
a (mean)	354.1	347.5	339.3	321.4
a (median)	326.7	320.0	311.1	293.2

NOTE: Source: HES 2001.

* indicates a value significantly different than 0 (at α=0.05).

It's interesting that although the differences in the regression coefficients are relatively large, the differences in the standard errors are relatively small. Further research and additional results from different data sets are needed to form an opinion regarding this issue.

Table 5 presents the multiple regression coefficients, with gross income, household size and dummy variables for being a member of a minority group (Arabs, ultra religious Jews) as the independent variables. The parameter ν is set to 1 (symmetric around the median) for household size and minority groups (represented by dummy variables), and it varies for income only. As can be seen, the regression coefficients of weight on income decline in absolute value, as ν declines (i.e., stressing higher incomes), by up to 0.001, so that the patterns detected in the simple regression continue to hold. However, the sign of the regression coefficients of household size remains the same

(negative), indicating that larger households respond in greater proportion to the questionnaires. Since the only difference between the regressions is the change in the parameter of income, the decline in absolute value of the coefficient of household size should be attributed to a change in the pattern of association between income and household size. The higher the stress on high-income groups, the lower is the absolute value of the effect of household size on nonresponse. Also, the magnitude of the regression coefficient of household size has changed from (–8.5) in the simple regression case, to (–4.0), which may be an indicator of the magnitude of association between income and household size. Given income and household size, Arabs tend to respond in higher proportion than the majority group, but the more we stress high income, the lower the effect (this may be due to small sample size in the upper range of incomes). One possible interpretation is that the higher the income, the lower the difference in response rates between the majority group and Arabs. On the other hand, the effect of stressing high-income range on ultra religious Jews is the opposite. The more high incomes are stressed, the lower is the response rate. Since it is a group with a low response rate on average, and seems to be motivated by an ideology, it is reasonable to conclude that the difference in response rate between this group and the rest of the population increases with income. However, the high standard errors show that only when high income is stressed, the dummy for ultra religious Jews is significant. As before, the difference between the constant terms is approximately 28.

Table 5. Multiple Regression Coefficients of Household Weight (0) on Gross Income per Household (1), Household Size (2), and Ethnic Grouping Dummy Variables (3, 4) (**), by Gini Parameter (ν).

Regression Coefficient	ν for Gross Income: 3.0	2.0	1.0	-0.5
b_{01}	-0.0026*	-0.0021*	-0.0016*	-0.0006*
	(0.000)	*(0.000)*	*(0.000)*	*(0.000)*
b_{02}	-1.41	-2.54	-3.95*	-6.32*
	(1.44)	*(1.39)*	*(1.35)*	*(1.31)*
b_{03}	-40.8*	-36.0*	-30.1*	-20.1*
	(8.1)	*(8.0)*	*(7.9)*	*(7.8)*
b_{04}	18. 8	23.3	28.9	38.3*
	(17.4)	*(17.2)*	*(17.1)*	*(16.9)*
a (mean)	357.9	354.6	350.5	343.6
a (median)	330.2	326.6	322.6	316.1

NOTE: Source: HES 2001.

* Indicates a value significantly different than 0 (α =0.05)

** The Dummy variable No. 3 has the following values: 1="Arab", 0="Other"; The Dummy variable No. 4 has the following values: 1="Ultra religious Jew", 0="Other".

All in all we can conclude that given religion and household size, the lower the income the lower the response rate, and given income and congregation, the smaller the household, the lower the response rate. The indications of the simple descriptive statistics that Arabs tend to better respond than the majority, and ultra religious Jews respond less than the rest of the population remained intact.

SECTION 6: THE EFFECT OF WEIGHTING ON MEAN INCOME AND INEQUALITY

We turn now to answer directly the research question raised by Deaton's (2003) conjecture concerning the impact of nonresponse on reported average income and on the measurement of poverty. Having found that there is a nonlinear systematic relationship between income and the response rate, one wonders whether this relationship can seriously bias measures of inequality like the Gini coefficient.

Table 6 presents the weighted mean of income (and extended Gini's) and the simple (nonweighted) mean of income (and extended Gini's). Presumably, the former represent unbiased estimates of the population parameters, while the latter represent the biased estimates, due to nonresponse. As expected from the regression results, weighted mean incomes are lower than nonweighted, which means that nonresponse tends to increase average income by a magnitude of up to 10%. To verify this conclusion, Table 7 presents the weighted and nonweighted mean incomes by the different demographic groups and household sizes. In only two out of fifteen cases we get that nonweighted average income is lower than weighted average income, and in these cases the sample sizes and the differences are small.

The second and more complicated issue is the effect of nonresponse on inequality and poverty measures. First, one has to distinguish between absolute and relative poverty lines. If the poverty line is an absolute one, then we get a clear answer: an under representation of the number of poor people is in contrast to Deaton's conjecture. On the other hand, if the poverty line is relative, or if we are concerned with the effect on inequality, then the direction of the bias is not clear, because nonresponse will affect both the numerator and the denominator of the inequality index.

Table 6. Mean Values and Extended Gini Coefficients of Gross Household Income.

Year	N	With Weighting				Without Weighting			
		Mean Income	ν of Gross Income			Mean Income	ν of Gross Income		
			3.0	2.0	1.0		3.0	2.0	1.0
2001	5,787	14,110	0.62	0.55	0.42	14,758	0.61	0.55	0.42
			(0.005)	*(0.005)*	*(0.005)*		*(0.005)*	*(0.005)*	*(0.006)*
2000	5,864	13,273	0.61	0.54	0.41	13,978	0.61	0.54	0.40
			(0.005)	*(0.005)*	*(0.005)*		*(0.004)*	*(0.004)*	*(0.005)*
1999	5,816	12,837	0.62	0.55	0.41	13,300	0.61	0.54	0.41
			(0.005)	*(0.005)*	*(0.005)*		*(0.005)*	*(0.005)*	*(0.006)*
1998	5,772	11,336	0.61	0.54	0.41	12,383	0.61	0.54	0.40
			(0.005)	*(0.005)*	*(0.006)*		*(0.005)*	*(0.005)*	*(0.005)*
1997	5,561	10,724	0.62	0.55	0.41	11,494	0.61	0.53	0.40
			(0.006)	*(0.006)*	*(0.007)*		*(0.005)*	*(0.005)*	*(0.005)*

NOTE: The case ν= -0.5 is omitted because we are assuming inequality aversion.

Source: HES 1997-2001, excluding the observations of East Jerusalem in 1997-1999.

To see this, let us evaluate the effect on the Gini index of inequality. The Gini coefficient can be written as:

$$G = \frac{2COV(Y, F(Y))}{\mu},$$

where Y is income, F(Y) is the cumulative distribution of income, and μ is mean income. We have already established that nonresponse tends to bias mean income upward, leading to a downward bias in inequality measurement. However, omitting an observation can increase or decrease the numerator. Lower participation rates by the poor may increase or decrease the numerator, depending on the shape of the distribution, while it will certainly increase the denominator. By stochastic dominance considerations, it can be shown that omitting an observation of a poor person will shift the cumulative distribution to the right. Hence, higher nonresponse rates among the poor will increase the mean income and will also tend to increase social welfare indicators such as μ (1-G). (The explanation to this argument is that $\mu_1 (1-G_1) > \mu_2 (1-G_2)$ is a necessary condition for distribution 1 to dominate distribution 2 according to second degree stochastic dominance – see Yitzhaki 1982). This means that the bias in the mean imposes a constraint on the magnitude of the bias in the Gini coefficient. However, it does not impose a constraint on the direction of the bias.

Instead of trying to solve the problem theoretically, we approach the question by comparing the effect of weighting on inequality. In any case, since nonresponse is not a

simple function of income, and it can affect the numerator in both directions, we cannot expect a clear theoretical answer.

The rest of Table 6 presents the extended Gini's of gross income, with the case ν =1 being the simple Gini, which is the most relevant. In all years, the simple Gini inequality index calculated with weights is higher than the Gini index without weights. Hence, nonresponse tends to bias inequality downward. However, in three years (out of five) the effect is relatively small (less than one standard error), and in one year, 1997, the difference is about 3.5%, which is quite large. Turning to the extended Gini, we see that in all cases the extended Gini, which relies on a weighted sample, is higher than the extended Gini based on the biased (nonweighted) sample. Hence, we may safely conclude that nonresponse tends to bias inequality downward.

Table 7. Estimates of Sample and Population-Adjusted Means of Gross Income per Household, by Household Size and Ethnic Grouping.

Ethnic Group		Household size					Total
		1	2	3	4	5+	
Majority	N	886	1,325	838	949	1,051	5,049
	Nonweighted Mean	7,136	12,613	17,471	19,477	20,943	15,482
	Weighted Mean	6,846	12,447	16,929	18,934	20,027	14,762
Arabs	N	20	64	74	116	388	662
	Nonweighted Mean	3,846	6,192	7,961	9,545	10,916	9,676
	Weighted Mean	3,598	5,755	7,579	8,861	10,608	9,121
Ultra religious Jews	N	3	16	4	8	45	76
	Nonweighted Mean	5,829	10,681	14,683	7,374	11,648	10,925
	Weighted Mean	4,385	10,794	11,180	6,740	11,776	10,617
Total	N	909	1,405	916	1,073	1,484	5,787
	Nonweighted Mean	7,060	12,299	16,690	18,313	18,040	14,758
	Weighted Mean	6,736	12,153	16,214	17,690	17,530	14,110

NOTE: Source: HES 2001.

SECTION 7: CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER RESEARCH

This paper presents a descriptive method that enables the researcher to trace the curvature of the regression curve by changing the weights assigned to different sections of the distributions of the independent variables. One major advantage of the method is that the researcher can use a different weighting scheme for each independent variable.

Although descriptive in nature, it can be turned into a standard analytical regression technique. By selecting the same weighting scheme for all independent

variables, one can have the structure of the OLS, with one simple modification: each variance is substituted by an extended Gini, and each Pearson's correlation is substituted by two appropriate extended Gini correlations. The only difference is that the method offers an infinite number of alternative regression coefficients. Clearly, the method enables the investigator to verify whether the results are sensitive to the specific weighting scheme used.

Turning to nonresponse, we have found that in the survey of household expenditure in Israel, nonresponse decreases with income, decreases with household size, and differs among ethnic groups. The Arab population tends to respond more than the majority, while the ultra religious Jewish population tends to respond less than the majority group. These results are in contrast with Deaton (2003) conjecture that high income groups tend to respond less to surveys. However, one should be aware that nonresponse is a survey-specific, not to mention the possibility of a country-specific phenomenon. Preliminary tests, which we conducted, have shown that the nonresponse in the income survey, which is a panel in the labor force survey, demonstrates a totally different response behavior. Also, additional variables may be relevant in explaining nonresponse behavior: schooling is a primary candidate.

In its present form, the regression method does not offer estimates of partial derivatives of the regression curve, but it seems that one can overcome this deficiency. Yitzhaki (2002) shows that if one divides the range of an independent variable into two sections, then the Gini regression coefficient (and OLS) can be presented as a weighted average of the two within-section regression coefficients, and a between section regression coefficient. The weights are the relative contribution of each section to intra and inter group Gini (variance – in OLS) of the independent variable. This decomposition can be easily expanded to an arbitrary number of sections. Further research is needed to apply this additional decomposition to get a piece-wise linear approximation to the regression curve that is based on a between-section component and within-section components of the approximation to allow the estimate of the partial derivative to vary over sections of the independent variables.

APPENDIX: PROOF FOR CONVERGENCE

This appendix shows that the difference between the two proposed estimators of β is negligible, and hence, b has a limiting normal distribution (as was shown for b_U). (The proof is limited to the case where ν is an integer).

The proof proceeds in two steps. In the first step, we express b_N as a linear combination of concomitants of order statistics; in the second step we show that the difference between the coefficients of $y_{x(i)}$, using the two presentations is negligible, for all i.

Let $\theta = -(\nu+1)\,\mathrm{COV}(Y, [1-F(X)]^{\nu})$ (see (15)) be the parameter of interest. As mentioned in (17), a U-statistic for estimating θ is

$$U = \frac{1}{\binom{n}{\nu+1}} \sum_{i=1}^{n} \left[\frac{1}{\nu+1} \binom{n-1}{\nu} - \binom{n-i}{\nu} \right] Y_{X_{(i)}} = \sum_{i=1}^{n} a_i Y_{X_{(i)}} .$$

(Note that if $\nu > (n-i)$, then $\binom{n-i}{\nu} = 0$).

Hence, the coefficient of $Y_{X(i)}$ is

$$\begin{aligned} a_i &= \frac{1}{\binom{n}{\nu+1}} \left[\frac{1}{\nu+1} \binom{n-1}{\nu} - \binom{n-i}{\nu} \right] \\ &= \frac{1}{n} - \frac{(n-i)!\,(n-\nu-1)!\,(\nu+1)}{(n-i-\nu)!\,n!} . \end{aligned} \qquad \text{(A.1)}$$

Using the semi-parametric approach, that is, replacing F by $\frac{r(x)}{n} = F_n$ (the empirical cdf), the estimator of θ (15) is given by

$$\begin{aligned} \hat{\theta}_N &= -(\nu+1)\,\mathrm{cov}\left(y, \left(1 - \frac{r(x)}{n}\right)^{\nu}\right) \\ &= \frac{-(\nu+1)}{(n-1)} \sum_{i=1}^{n} y_i \left(\left(1 - \frac{r_i}{n}\right)^{\nu} - \mathrm{AVE}\left(1 - \frac{r_i}{n}\right)^{\nu} \right) \\ &= \frac{-(\nu+1)}{n-1} \sum_{i=1}^{n} \left(\left[1 - \frac{i}{n}\right]^{\nu} - \mathrm{AVE}\left(1 - \frac{i}{n}\right)^{\nu} \right) y_{x_{(i)}} = \sum_{i=1}^{n} b_i y_{x_{(i)}} , \end{aligned}$$

where

$$\mathrm{AVE}\left(1-\frac{i}{n}\right)^{\nu} \text{ is the average of } \left(1-\frac{i}{n}\right)^{\nu} .$$

Hence, the coefficient of $y_{x(i)}$ is:

$$b_i = \frac{-(\nu+1)}{(n-1)}[(1-\frac{i}{n})^{\nu} - \mathrm{AVE}(1-\frac{i}{n})^{\nu}] .$$

Note that:

$$\mathrm{AVE}\left(1-\frac{i}{n}\right)^{\nu} = \frac{1}{n}\sum_{j=1}^{n-1}(\frac{j}{n})^{\nu} = \frac{1}{n^{\nu+1}}\sum_{j=1}^{n-1} j^{\nu} .$$

Using Rieman approximation, then

$$\int_0^1 x^{\nu}\,dx = \frac{1}{\nu+1},$$

implies that

$$\frac{1}{\nu+1} - \frac{1}{n} \le \frac{1}{n^{\nu+1}}\sum_{j=1}^{n-1} j^{\nu} \le \frac{1}{\nu+1} .$$

Using these bounds, the difference between the coefficients is

$$a_i - b_i = \frac{1}{n} - \frac{(n-i)(n-i-1)...(n-i-(\nu-1))(\nu+1)}{n(n-1)...(n-\nu)} + \frac{\nu+1}{n-1}\left(\frac{(n-i)^{\nu}}{n^{\nu}} - \frac{1}{\nu+1}\right) + O(\frac{1}{n^2}).$$

Using the common denominator $n^{\nu}(n-1)...(n-\nu)$, of order $n^{2\nu}$, it is easy to see that in the numerator, the highest power of n, which is $2\nu-1$, cancels out, so that the numerator will be of order $n^{2\nu-2}$. Therefore, the difference is $O(n^{-2})$ while each coefficient is of order $1/n$. This implies that $\sqrt{n}\,(a_i-b_i) \to 0$ as $n \to \infty$, which completes the proof.

References:

Angrist, J., Chernozhukov, V., and Fernándaz-Val, I. (2004), "Quantile Regression Under Misspecification, With an Application to the U. S. Wage Structure," NBER Working Paper Series 10428, (http://www.nber.org/papers/w10428).

Araar, A., and Duclos, J.Y. (2003), "An Atkinson-Gini Type Family of Social Evaluation Functions," *Economics Bulletin*, 3, 19, 1-16.

Chakravarty, S. R. (1988), *Ethical Social Index Numbers*, Berlin: Springer-Verlag.

Davidson, R., and Duclos, J.Y. (1997), "Statistical Inference for the Measurement of the Incidence of Taxes and Transfers," *Econometrica*, 65, 6, 1453-1465.

Deaton, A. (2003), "Measuring Poverty in a Growing World (or Measuring Growth in a Poor World)," Working Paper No. 9822 (July), National Bureau of Economic Research.

Donaldson, D., and Weymark, J.A. (1983), "Ethically Flexible Gini Indices for Income Distributions in the Continuum," *Journal of Economic Theory*, 29, 353-358.

Durbin, J. (1954), "Errors in Variables," *Review of International Statistical Institute*, 22, 23-32.

Garner, T.I. (1993), "Consumer Expenditures and Inequality: An Analysis Based on Decomposition of the Gini Coefficient," *Review of Economics and Statistics*, 75(1), 134-138.

Gregory-Allen, R., and Shalit, H. (1999), "The Estimation of Systematic Risk Under Differentiated Risk Aversion: A Mean-Extended Gini Approach," *Review of Quantitative Finance and Accounting*, 12, 135-157.

Groves, R. M., Dillman D. A., Eltinge J. L., and Little, J. A. (eds.) (2002), *Survey Non Response*, New York: John Wiley & Sons.

Groves, R. M., and Couper, M. P. (1998), *Non Response in Household Interview Surveys*, New York, John Wiley & Sons.

Hoeffding, W. (1948), "A Class of Statistics with Asymptotically Normal Distribution," *Annals of Mathematical Statistics*, 19, 293-325.

Kalton, G., and Flores-Cervantes, I. (2003), "Weighting Methods," *Journal of Official Statistics,* 19, 2, 81-97.

Kantorowitz, M. (2002). "Reducing Nonresponse Bias of Income Estimates by Integrating Data from Households' Expenditure and Income Surveys," Paper presented at the International Conference on Improving Surveys (ICIS), Copenhagen, Denmark

Koenker, R., and Bassett, G. Jr. (1978), "Regression Quantiles," *Econometrica*, 46, 33-50.

Lerman, R., and Yitzhaki, S. (1994), "The Effect of Marginal Changes in Income Sources on U. S. Income Inequality," *Public Finance Quarterly*, 22, 4, 403-417.

Millimet, D. L., and Slottje, D. (2002), "An Environmental Paglin-Gini," *Applied Economics Letters*, 9, 271-274.

Mistiaen, J. A., and Ravallion, M. (2003), "Survey Compliance and the Distribution of Income," Washington D. C. The World Bank, (June), Processed.

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

Randles, R. H., and Wolfe, D. A. (1979), *Introduction to the Theory of Nonparametric Statistics*, New York: John Wiley & Sons.

Schechtman, E., and Yitzhaki, S. (1987), "A Measure of Association Based on Gini's Mean Difference," *Communications in Statistics, Theory and Methods*, 16, 207-231.

_____ (1999), "On the Proper Bounds of the Gini Correlation," *Economics Letters*, 63, 2, 133-138.

_____ (2003), "A Family of Correlation Coefficients Based on Extended Gini," *Journal of Economic Inequality*, 1, 2, 129-146.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: John Wiley & Sons.

Shalit, H., and Yitzhaki, S. (2002), "Estimating Beta," *Review of Quantitative Finance and Accounting*, 18, 2, 95-118.

Shao, J., and Tu, D. (1996), *The Jackknife and Bootstrap*, Springer, New York.

Wodon, Q., and Yitzhaki, S. (2002), "Inequality and Social Welfare," in J. Klugman, ed., *PRSP Sourcebook*, the World Bank: Washington DC.

Yitzhaki, S. (1982), "Stochastic Dominance, Mean Variance, and the Gini's Mean Difference," *American Economic Review*, 72, 1, 178-185.

_____ (1983), "On an Extension of the Gini Inequality Index," *International Economic Review*, 24, 617-628.

_____ (1991), "Calculating Jackknife Variance Estimators for parameters of the Gini Method," *Journal of Business & Economic Statistics*, 9, No. 2, 235-239.

_____ (1996), "On Using Linear Regressions in Welfare Economics," *Journal of Business & Economic Statistics*, 14, 4, 478-486.

_____ (1998), “More Than a Dozen Alternative Ways of Spelling Gini,” *Research on Economic Inequality*, 8, 13-30.

_____ (2002), “Do We Need a Separate Poverty Measurement,” *European Journal of Political Economy*, 18, 61-85.