סדרת ניירות עבודה

WORKING PAPER SERIES

No.53

רגרסיית ג׳יני מרובה:

שתי גישות והאינטראקציה ביניהן

Gini's multiple regressions:

two approaches and their interaction

שלמה יצחקי*, עדנה שכטמן**, טאינה פודלוב***
Shlomo Yitzhaki*, Edna Schechtman**, Taina Pudalov***

November, 2010 נובמבר

* הלשכה המרכזית לסטטיסטיקה והאוניברסיטה העברית.
** אוניברסיטת בין גוריון.
*** הלשכה המרכזית לסטטיסטיקה.

*Central Bureau of Statistics and Hebrew University.
** Ben Gurion University.
*** Central Bureau of Statistics.

L H B E N

Published by the Central Bureau of Statistics, 66 Kanfe Nesharim St.,
Corner Bachi St., P.O.B 34525, Jerusalem 91342, Israel
Tel. 972-2-6592666; Fax: 972-2-6521340
Internet Site: www.cbs.gov.il
E-Mail: info@cbs.gov.il

We would like to thank Yevgeny Artzev, Zvi Gilula, James Heckman, Peter Lambert, Vadim Marmer and Gideon Schechtman for helpful comments and discussions on previous drafts of the paper.

Abstract

Two regressions can be interpreted as based on Gini's Mean Difference (GMD): a semiparametric approach, which relies on weighted average of slopes defined between adjacent observations and a minimization approach, which is based on minimization of the GMD of the residuals. The estimators obtained by the semiparametric approach have representations that resemble the OLS estimators. In addition they are robust with respect to extreme observations and monotonic transformations. The estimators obtained by the minimization approach do not have a closed form. The relationship between the estimators obtained by the two methods is studied in this paper. Combination of the methods provides tools for challenging the specification of the model. In particular it provides tools for assessing the linearity of the model. It can be applied to each explanatory variable individually and to several explanatory variables simultaneously without requiring replications. The semiparametric method and its relationship with the minimization approach are illustrated using consumption data. It is shown that the linearity of the Engel curve, and therefore the 'linear expenditures system' is not supported by the data.

Key Words: Gini's Mean Difference, Average Derivative, Linear Expenditure System, Monotonicity.

1. INTRODUCTION

Two regression methods can be interpreted as based on Gini's Mean Difference (GMD). One relies on a weighted average of slopes defined between adjacent observations (a semi-parametric approach) and the other is based on minimization of the GMD of the residuals.

The **semi-parametric approach** is based on estimating a regression coefficient that is a weighted average of slopes defined between adjacent observations (or all pairs of observations) of the regression curve. It resembles the OLS in the sense that the estimators can be explicitly presented and all the expressions used have parallels in OLS regression. The derivation of the estimators and their properties are discussed in detail in Schechtman *et al*. (2008b) (for a more general case) therefore only a review will be offered below. We note that this regression does not require a specification of the functional form of the model. It can be used whenever the investigator is interested in estimating average slopes or arc-elasticities without requiring a formal model. In this sense it resembles the method suggested by Härdle and Stoker (1989) and Rilstone (1991). Parameters and estimators derived according to this approach will be denoted by the subscript N.

The **minimization approach** is based on minimization of the GMD of the residuals. This approach requires the assumption of a linear model. It is similar to Least Absolute Deviation (LAD) regressions (Bassett and Koenker, 1978). Instead of minimizing the sum of absolute deviations of the residuals, the GMD of the residuals which is the mean of the absolute differences between all pairs of residuals is minimized. Similar to the case in LAD, the estimators can be derived numerically but there are no explicit expressions for them. Parameters and estimators derived following the minimization approach will be denoted by the subscript M.

Before we proceed we note that our underlying assumption throughout this paper in that the explanatory variables are random variables. This is not in line with the common practice which assumes that they are fixed. We feel that it is a more realistic assumption (but comes with a price tag attached to it – the analysis is more complicated).

In this paper we show that the combination of the two approaches mentioned above has several advantages over OLS and other regression techniques. For example, the GMD regression method offers tools to assess monotonicity and linearity. The tools are based on the fact that the

Gini regression coefficient is a function of Gini covariances (hereafter co-Ginis) between the dependent variable and the explanatory variables as well as of the co-Ginis between each pair of explanatory variables. Note that due to the asymmetric property of the co-Gini, there are two co-Ginis between each pair of variables.[1] More specifically, each of the two regression methods generates a co-Gini between the residuals generated by the fit and the explanatory variable (for each explanatory variable separately). The combination of the two co-Ginis can be used to evaluate the goodness of the linear fit for each explanatory variable individually, as well as for a subset or for the entire model, and does not require replications.

In addition we point out that

(i) The estimators of the Gini regression coefficients derived under the minimization approach can be interpreted as the solutions of the minimization of an average of all possible absolute deviations from all possible quantiles of the residual (Koenker and Bassett (1978)). Hence, one can view it as an extension of quantile-regressions. Similar to those regressions, the estimators cannot be explicitly derived and are derived numerically.

(ii) The parameters associated with the GMD (the equivalents of the variance and covariance) can be presented as areas enclosed between the Line of Independence and the Absolute Concentration Curve (ACC) (Yitzhaki and Olkin (1991); Yitzhaki (1998, 2003);Yitzhaki and Schechtman (2004)). The properties of the ACC curves enable one to check the monotonicity of the regression curve and to visually observe whether omitting observations that are located in a given section of the range of the explanatory variable could change the sign of the regression coefficient.[2]Also, the ACC can be used to suggest improvements of the model to better fit the data. However we do not investigate this issue here. We only point out several directions for further

[1] This paper relies heavily on the decomposition properties of the GMD (Yitzhaki, 2003). The properties of the decomposition of regression coefficients are listed in Yitzhaki and Schechtman (2009).

[2] Heckman, Urzua, and Vytlacil (2004) explain why one should refer to this property as uniformity. We will define uniformity (or monotonicity) of a regression curve as a regression curve that yields a regression coefficient with the same sign along different sections of the range of the explanatory variable.

research. To make it easier to read the effects of the sections on the GMD regression coefficient from the figures of the concentration curve, we present a variation of the concentration curve, called the LMA. The LMA is the vertical difference between the ACC when the two variables are considered independent (called the line of independence, LOI) and the ACC of the variable. In some sense it resembles the difference between moments and central moments.

It is worth emphasizing that one of the major advantages of the GMD regressions is in offering a complete framework for dealing with a multiple regression problem. First one estimates the slopes of the regression curve according to the semi-parametric approach without specifying a model. Then one uses the residuals from the fitted curve and tests whether they fulfill the necessary conditions for the minimization approach (which were obtained assuming linearity). If for any given explanatory variable the above conditions are fulfilled; that is, the hypothesis that the two regression coefficients are equal is not rejected, then one concludes that the regression curve is linear in this variable. This property is especially important in regressions with several explanatory variables. It enables the investigator to find a set of variables that allows linear predictions without having to commit herself to the linearity of the model as a whole. Provided that the linearity hypothesis is not rejected for all explanatory variables one can examine the properties of the residuals, such as their distribution, whether it is symmetric around the regression line or not, the serial correlation between them, etc, using the methodologies that will keep the analysis under the Gini framework (see for example Frick et. al. (2006) who developed ANOGI - the Gini equivalent of ANOVA). Although each stage could be done by alternative methods, we are not aware of any method that can offer a complete set that is governed by a unified framework and therefore offers a method to test the assumptions behind the regression with an internal consistency. Note also that there is no need for replications of observations, as is the case in the common tests for linearity.

To avoid complicated notation, capital letters are used to indicate population parameters while lower-case letters indicate sample statistics. For example, the term COV(,) is used to represent the covariance in the population while the term cov(,) represents the sample's value. Finally, F(X) is used to denote the cumulative distribution of X while r(x) denotes ranks in the

sample. Note that in what follows X (the explanatory variable) is treated as a random variable. Although it complicates the theory, it is a more realistic situation in economics.

The structure of the paper is as follows: Section 2 is devoted to a brief review of Gini's multiple regression. Section 2.1 presents the semi parametric approach and Section 2.2 presents the minimization approach. Section 3 concentrates on the relationship between the two types of regression methods while Section 4 relies on the properties developed in Section 3 to assess the linearity of the model. Section 5 shows the connection with concentration curves while Section 6 illustrates the methodology by assessing the linearity of consumption as a function of income and family size. Section 7 concludes and offers a direction for further research.

2. GINI'S MULTIPLE REGRESSIONS

The aim of this section is to briefly review the results for Gini's multiple regression. The results for the semi parametric method for the simple Gini regression are presented in Olkin and Yitzhaki (1992) and Yitzhaki (1996), and the more general case, the extended Gini multiple regression case, is presented in detail in Schechtman *et al*. (2008b). The GMD is a special case, with the extended Gini parameter $\nu=1$. Therefore we shall concentrate here on results that are relevant to the current paper and refer the interested reader to Schechtman *et al*. (2008b) for the general theory.

2.1 The Semi Parametric Approach

Let $(Y, X_1,\ldots,X_K)$ be a (K+1)-variate random variable with expected values $(\mu_Y, \mu_1, \ldots, \mu_K)$, respectively, and a finite variance-covariance matrix Σ. Assume that we have a general regression curve defined by

$$g(x_1,\ldots,x_K) = E\{Y| X_1=x_1,\ldots, X_K=x_K\}.$$

The investigator is interested in estimating a linear approximation of the regression curve. That is, she needs to estimate a set of slopes (the constant term will be determined later) which are conditional slopes: the slope of Y on X_i is conditional on the other X's in the model.

The steps taken are as follows. First, the linear approximation is defined. Then, the parameters (conditional slopes) are interpreted as the solutions of a set of linear equations which involve the (known) univariate slopes, and the last step is the estimation procedure, based on the data.

The resulting vector of regression coefficients of step 2, β_N, is given by

$$\beta_N = [E(V'X)]^{-1} E(V'Y), \qquad (2.1)$$

where $\beta_N = \{\beta_{N1},..., \beta_{NK}\}$ is a (K x 1) column vector of the (conditional) regression coefficients, V is an (nxK) matrix of the cumulative distributions of $X_1,..., X_K$ (in deviations from their expected values), Y is an (n x 1) vector of the dependent variable and X is an (n x K) matrix of the deviations of the explanatory variables from their expected values. The elements of E(V'Y) and E(V'X) are $COV(Y, F(X_k))$ and $COV(X_j, F(X_k))$, respectively. It is assumed that the rank of V'X equals K, the number of explanatory variables. This implies a restriction on the choice of the explanatory variables that does not exist in OLS: no explanatory variable can be a monotonic transformation of another explanatory variable, because if it does it will imply identical rows in the matrix V'X (which depends on X_i via $F_i(X)$). The details of the derivation are given in Schechtman *et al*. (2008b).

The natural estimators of the regression coefficients are based on replacing the cumulative distributions by the empirical distributions (which are calculated using ranks):

$$b_N = [v'x]^{-1}(v'y) \ , \qquad (2.2)$$

where v is a matrix with elements $[n^{-1}(r(x_{ik})) - 1/2]$, and $r(x_{ik})$ is the rank of x_{ik} among $x_{1k},\ldots,x_{nk}$. Schechtman *et al*. (2008b) prove that b_N is a consistent estimator of β_N and its limiting distribution is normal under regularity conditions. The proofs will not be repeated here.

Once the Gini regression coefficients are estimated, the constant term can be estimated by minimizing a function of the residuals. The exact function used determines whether the regression passes through the mean, the median, or any other quantile. The multiple regression procedure, although it is not based on an optimization procedure, generates equivalents to the OLS's normal equations. This property plays an important role in this paper, as will be shown in the next section. By defining the error term and substituting for the multiple regression coefficients, it can be shown that

$$COV(\varepsilon, F_k(X)) = 0 \text{ for } k=1,\ldots, K, \qquad (2.3)$$

as stated in the following Lemma.

Lemma 2.1: Define the vector $\varepsilon = Y - X\beta_N$. Then, $E(V'\varepsilon) = 0$ where 0 is a vector of zeros.

This property holds in the sample as $v'e = 0$, where $e = y - x b_N$.

Finally, because each variance and covariance in OLS regression is substituted in Gini regression by GMD and co-Gini, respectively, it is easy to verify that other concepts used in the OLS such as partial correlation coefficients can be translated into the Gini regression. Among

those concepts is R^2 of the regression, which can be considered as a measure to assess the share of the (square of the) GMD which is explained by the model. That is, the R^2 for the Gini semi-parametric regression is defined (Olkin and Yitzhaki (1992)) as one minus the square of the GMD of the error term *divided by* the square of the GMD of the dependent variable:[3]

$$GR^2 = 1 - [\ cov(e,r(e))/cov(y,r(y))]^2. \qquad (2.4)$$

However, because the decomposition of the GMD of a linear combination into the contributions of the different components is more complicated than the decomposition of the variance, the properties of GR^2 differ from the properties of the equivalent term in OLS (see Yitzhaki, 2003). For example, as will be seen in the next section, GR^2 will obtain its maximal value under the Gini minimization approach. Therefore, the GR^2 in the semi-parametric version would always be not greater than the GR^2 in the minimization approach. Equality holds when the model is linear in all the explanatory variables. (When the model is linear in all the explanatory variables, GR^2's of both methods are equal).

An additional measure of the quality of the fit of the model to the data in the GMD regression is to look at the Gini correlations between the dependent variable and the fitted model. Formally:

$$\Gamma_{Y\hat{Y}} = \frac{cov(Y,F(\hat{Y}))}{cov(Y,F(Y))} \quad \text{and} \quad \Gamma_{\hat{Y}Y} = \frac{cov(\hat{Y},F(Y))}{cov(\hat{Y},F(\hat{Y}))}\ , \qquad (2.5)$$

where $\hat{Y}$ is the predicted variable. As a result of the differences between the properties of the decomposition of the variance to those of the decomposition of the GMD, we substitute the R^2 of OLS by three measures: the one in (2.4) and the two in (2.5). Note, however, that in the OLS, the parallels to these three measures are numerically equal.

2.2 The Minimization Approach

This approach, which is based on minimization of the GMD of the error term, has already been developed in the literature and it is referred to as R-regression (Jaeckel (1972); Jurečková (1969, 1971); McKean and Hettmansperger (1978); Hettmansperger (1984)). Therefore its properties will not be repeated here.[4] For our argument, only the orthogonality condition (also known as the normal equations) is needed. Note that this method requires the specification of a model.

[3] In the empirical application we use GR= 1 - cov(e,r(e))/cov(y,r(y)).

[4] It is worth emphasizing that the connection between R-regression and GMD was not recognized in the literature mentioned above. Many of the properties of those regressions can be traced to the properties

Consider the following model:

$$Y = X\beta + \varepsilon \tag{2.6}$$

with the usual assumptions on ε, that is: the ε's are independent and have mean zero and a constant variance, and the additional assumption that X_i and ε_j are independent for all i,j. (Note that this assumption follows automatically if the X_i 's are considered non-random). The estimated equation is:

$$y = xb_M + e_M \tag{2.7}$$

where b_M is the estimator of the slope β_M using the minimization of GMD of the error term e_M. Using the covariance presentation of GMD and imposing the restriction that the mean of the residuals is zero enables us to show that minimizing GMD of the residuals is equivalent to minimizing

$$\sum_{i=1}^{n} r(e_{Mi})\, e_{Mi} \tag{2.8}$$

where $r(e_{M_i})$ is the vector of ranks of the residuals $e_{M_{ii}}$. The only property required for our present paper is that minimizing the GMD of the error term yields an orthogonality condition which is the equivalent of the OLS normal equation, and is given by

$$x'r_M = 0\,, \tag{2.9}$$

where r_M is the vector of the ranks of e_M, rescaled and shifted to have a zero mean.[5] (The i-th element of $x'r_M$ is $cov(x_i,r_M)$). Equation (2.9) says that the sample covariance between the rank of the residuals and the variate value of the explanatory variable is set to zero as a result of the minimization of GMD of the residuals. Finally, it is worth noting that Olkin and Yitzhaki (1992) show that (2.9) holds for the simple regression case. Here the statement in (2.9) is extended to the multiple regression case and it shows that the same relation exists, for each explanatory variable, in the multiple regression case.

---

of GMD. Bowie and Bradfield (1998) compare the robustness of several alternative estimation methods in the simple regression case and find the minimization of the GMD among the most robust methods.

[5] Because (2.6), the GMD of the residuals, is a piece-wise linear function, its partial derivative with respect to b_M may not exist because the derivative is a step function. In this case the solutions b_M to (2.7) form a segment on the real line and b_M is determined up to a range. The larger the sample the lower the probability that such an event occurs.

3. THE RELATIONSHIP BETWEEN THE ESTIMATORS OF THE TWO APPROACHES

As seen in the previous section, each approach yields a set of estimators, a set of residuals, and a set of "normal equations". The aim of this section is to identify the condition under which the two estimators actually estimate the same parameters in the population and to give necessary and sufficient conditions under which the two estimators are algebraically identical. Recall that b_M, the estimator obtained by the minimization of GMD of the residuals, estimates the vector of slopes β **under the linearity assumption**, while no model was required in order to derive b_N, which is based on weighted averages of slopes. Therefore in general the two approaches may yield different estimators. However, when the model **is linear**, the vector of (true) slopes of the regression curve under the semi parametric approach, β_N, is equal to the vector of slopes which was obtained under the linear assumption, β_M . The reason is because when the model is linear, the slopes are all equal along the regression curve and the weighted average of them is that same constant, therefore $\beta_N = \beta_M = \beta$. In other words, under the linearity assumption $\beta_N = \beta_M = \beta$.

Hence, both b_N and b_M estimate the same vector of slopes, namely β, and the first order conditions of both methods should hold with the same set of residuals. (The last fact will be our basic tool for assessing the linearity of the model, as will be discussed in the next section). To see that, let

$$Y = X\beta + \varepsilon, \tag{3.1}$$

where X and ε are independent.

Then, by (2.1)

$$\beta_N = [E(V'X)]^{-1} E(V'Y) = \beta + [E(V'X)]^{-1} E(V'\varepsilon) \quad . \tag{3.2}$$

The fact that $\beta_N = \beta$ implies that $E\{V'\varepsilon\} = 0$, which is the first-order condition for the semi-parametric approach (Lemma (2.1)).

By substituting $y = xb_M + e_M$ into (2.2) the following relationship holds between b_N and b_M in the sample:

$$b_N = b_M + (r'x)^{-1} r'e_M \, . \tag{3.3}$$

The following proposition gives necessary and sufficient conditions for b_N to be algebraically equal to b_M in the sample.

Proposition 3.1: Let $(y_i, x_{1i},..., x_{Ki})$, i=1,...,n, be a sample of size n from a continuous multivariate distribution with finite second moments. Then

(a) $r'e_M = 0$ iff $b_M = b_N$.

(b) $x'r_N = 0$ iff $b_M = b_N$, where r_N is the vector of ranks of e_N.

Proof:

(a) Trivial. Follows directly from (3.3).

(b) Assume $x'r_N = 0$. It can be seen from (2.6) and the first-order condition for minimization, that b_M solves the set of equations $x'r_M = 0$.[6] Also, by the assumption $x'r_N = 0$, there exists b_N which solves the same equation. Since b_M may be non-unique and an entire interval may be obtained as a solution, one should qualify the proposition to say that there exists a b_M which is a solution to the minimization problem and fulfills the proposition. Therefore, $b_M = b_N$. On the other hand, if $b_M = b_N$ then, clearly, $x'r_N = 0$.

Note that proposition 3.1 holds for each explanatory variable separately. That is, we may find that the model is linear with respect to some explanatory variables, and non-linear with respect to others.

The second term in (3.3) is equal to the semi-parametric estimator of a regression in which the dependent variable is the residuals of the minimization approach. Therefore one can view (3.3) as running the regression in two stages: in the first stage the minimization approach is applied to the data. Then, the semi-parametric approach is applied to the residuals obtained by the minimization approach. If the regression is linear in each explanatory variable, then the second stage estimate of the first stage residuals with respect to each explanatory variable is equal to zero. If, on the other hand, the regression is not linear in one of the explanatory variables, then the second-stage estimate will deviate from zero. The order can be changed. One can run the semi-parametric regression first, and then use the residuals to test whether the first order conditions of the minimization approach are fulfilled using the set of residuals of the semi-parametric approach. This order is suggested in the next section.

4. ASSESSING THE GOODNESS OF FIT OF THE LINEAR MODEL.

[6] Note that r_M (and r_N) are functions of e, the error term.

In what follows, we treat each explanatory variable X_k separately. For simplicity, $F_k(X_k)$ will be denoted by $F(X_k)$. When the model is linear with respect to an explanatory variable, the semi-parametric approach and the Gini minimization approach estimate the same vector of parameters, β. Lemma 2.1 and the assumptions of the linear model in the minimization approach imply that $COV(\varepsilon, F(X_k))$ and $COV(X_k, F(\varepsilon))$ are equal to zero for each X_k. That is, if the specification of the model is correct then the following relationships hold in the population:

$$COV(\varepsilon, F(X_k)) = 0 = COV(X_k, F(\varepsilon)). \quad (k=1,\ldots,K) \qquad (4.1)$$

The left-hand side of (4.1) is the population version of the "normal equation" obtained by the semi-parametric approach, while the right-hand side is the population version of the "normal equation" obtained by the minimization. The proposed method will take advantage of the fact that two covariances are involved, which is special to the Gini regression approach.

In estimating a Gini regression coefficient, one sample covariance is *set to zero by construction*, according to the approach taken, but the other sample covariance can be used for the test. For example, by running the semi parametric regression, the sample covariance of the left-hand side of (4.1) is set to zero by construction (with e_N as the residuals). Hence, one can test for linearity by testing (against a broad alternative) whether $COV(X_k, F(\varepsilon)) = 0$, using e_N. Note that this last covariance is set to equal zero under the minimization approach, when using e_M as the residuals. Alternatively, one can run R-regression and reverse the procedure, as was mentioned at the end of the previous section. Starting with a semi-parametric regression to construct a linearity test has several advantages:

(1) The semi-parametric regression does not require specification of the model.

(2) Unlike the minimization approach, there is no problem of non-uniqueness of the estimated regression coefficient.

(3) The estimators of the semi-parametric approach can be written explicitly using OLS – like terminology.

(4) The point estimators of the semi-parametric approach can be calculated easily using the instrumental variable approach therefore standard regression software can be used.[7]

For these reasons the following procedure is suggested for assessing the linearity in X_k:

[7] The semi parametric estimators can be viewed as OLS instrumental variable (IV) estimators, with the rank of each variable being used as an IV. However, note that the assumptions that are assumed here are entirely different (see Yitzhaki and Schechtman (2004)), therefore the inference can not be drawn from there.

Step 1: Use the semi-parametric approach to estimate the Gini regression coefficients. Obtain the residuals e_N and the normal equation $cov(e_N,r_k)=0$, where r_k is the vector of ranks of X_k.

Step 2: Use e_N to test H_0: $COV(X_k, F(\varepsilon)) = 0$. H_0 states that the normal equation of the minimization approach holds for the residual of the semi-parametric approach. If H_0 is rejected, then one can conclude that the model is not linear in X_k. Recall that in the sample $cov(x_k, r(e_M)) = 0$ by construction (where $r(e_M)$ is the rank of the residuals according to the minimization approach).

A test of H_0 will be based on a U-statistic, obtained by replacing the cumulative distribution by the ranks. Its consistency and asymptotic distribution under H_0 are given in Schechtman and Yitzhaki (1987). Because equation (4.1) holds for each X_k separately, one can use the proposed test for each explanatory variable separately.[8] However, if one wishes to test for linearity of several X's simultaneously, then one should run the regression twice – once for the full model and then for the reduced model (excluding the relevant X's) and compare the Gini's of the residuals of the two models. If one is interested in the model as a whole, then one would replace X_k by $\hat{Y}$ in (4.1). That is, one would use $\hat{Y}$ from the semi-parametric approach and test whether $COV(\hat{Y}, F(\varepsilon)) =0$, where $\hat{Y}$ is the predicted value of Y. This test examines whether the same set of residuals and predicted values can serve as solutions to both methods.

A comparison of the proposed test with alternative tests of linearity (see, for example, Lewbel (1995) and Banks, Blundell and Lewbel (1997)) is beyond the scope of this paper. However it is worthwhile to mention that most of the parametric tests for linearity require replications or near replications (see Neill and Johnson (1984) for a review). The nonparametric tests are based on kernel regression estimators and smoothing splines. For example, Kozek (1990) derives a test for linearity which is based on a comparison between a non-parametric estimate of the regression function and the estimates of the linear forms, with β's being replaced by their OLS estimators. The test is based on the maximal deviation between the two estimators of the regression function.

Eubank and Spiegelman (1990) propose tests that are constructed from non-parametric regression fits to the residuals from the linear regression.

[8] A similar approach would be to test whether $b_N = b_M$. The advantage of relying on covariances is that only one set of regression coefficients has to be estimated.

Finally, it is worth mentioning that other assumptions imposed on the regression can be tested by using the GMD. For example, D'Agostino's (1971, 1972) test for normality of the residuals is based on the statistic $cov(e_N, r(e_N))/nS$ (where S is the sample standard deviation of the residuals and n is the sample size), whose numerator is the GMD of the residuals.

5. GINI REGRESSIONS AND CONCENTRATION CURVES

Yitzhaki (1990) suggests a method for examining whether a monotonic transformation of a variable can change the sign of the OLS regression coefficient. It turns out that similar conditions hold for the possibility of changing the sign of a Gini regression coefficient. Moreover, the application of the method for the Gini regression is more convenient because one can see the contributions of the different sections of the inspected variable to the sign of the estimate geometrically. The suggested methodology relies on an advantage of using the GMD as an index of variability: its connection with the Lorenz curve, which provides more information about the structure of the variability. A similar curve, the concentration curve (to be defined below) can be used to shed further light on a key parameter of Gini regressions — the co-Gini, that is: the covariance between a variable and the cumulative distribution of another variable. For discussions on the properties of concentration curves see Kakwani (1980), Yitzhaki and Olkin (1991), and Yitzhaki (1998, 2003). Yitzhaki and Schechtman (2004) present its properties with respect to the monotonicity of the instrumental variables, both for simple OLS and Gini regressions. The concentration curve can supply a visual way to determine whether the slopes of the regression curve, defined by adjacent observations, are somewhat equal or whether one can see a pattern in them. One can see visually whether omitting extreme observations can change the sign of the covariance or whether a monotonic transformation can change the sign of a regression coefficient. Also, it can serve as a visual indicator as to whether the model is well specified. Another important use of the concentration curve in econometrics is that it enables to see whether the relationship between two random variables is monotonic along the entire range of the variables, which is a crucial property whenever heterogeneous reactions to a treatment are assumed (See Heckman, 2001; Heckman, Urzua and Vytlacil, 2004). In this paper a regression curve is considered monotone (or uniformity according to Heckman, Urzua and Vytlacil, 2004) if the sign of the slope of the curve does not change over the entire range of the explanatory variable.

In what follows the concentration curve is defined and its use in Gini regression is illustrated.

Given two variables X and Y, the absolute concentration curve (hereafter ACC) [9] of Y with respect to X is defined as:

$$\theta_{Y,X}(p) = \int_{-\infty}^{Z_P} H(x)\ f(x)dx$$
$$\text{where } p = \int_{-\infty}^{Z_P} f(x)dx, \qquad (5.1)$$

where $H(x) = \mu_Y - E\{Y \mid X=x\}$ and f(x) is the density function of X. The intermediate variable, $Z_{p,}$ is the inverse of the cumulative distribution of X at p. [10] That is, $Z_{p,}$ is the p-th percentile of the distribution of X. θ is the vertical axis and p is the horizontal axis. The properties of the concentration curve are:

a. The concentration curve starts at $(\theta,p) = (0,0)$ and ends at (0,1).

b. If X and Y are independent then $\theta_{Y,X}(p) = 0$ for all p.

c. The area between the horizontal axis and the concentration curve is equal to COV(Y,F(X)). That is

$$\text{COV}(Y, F(X)) = \int_{0}^{1} \theta_{Y,X}(p)dp\ . \qquad (5.2)$$

d. If $\theta_{Y,X}(p) \geq (\leq)\ 0$ for all p, then $\text{COV}(Y, T(X)) \geq (\leq)\ 0$ for every monotonic transformation T(x) with $T' \geq 0$.

e. If $\theta_{Y,X}(p)$ intersects the horizontal axis then one can always find :

[9] The literature concerning concentration curves defines the concentration curve in percentage terms. In regression analysis one has to define it in absolute terms. To distinguish between the two versions, the term "absolute" is added.

[10] The usual definition of H(x) is $H(x) = E\{Y| X=x\}$. For the special case of analyzing the regression coefficient it is convenient to define a curve as the vertical difference between the diagonal and the Absolute Concentration curve. In Olkin and Yitzhaki (1991) terminology this new curve is the vertical difference between the LOI and the Absolute Concentration curve. This curve is sometimes referred to as LMA, Line Minus Absolute concentration curve.

e.1. Two monotonic non-decreasing transformations $T_1(X)$ and $T_2(X)$ so that $COV(Y,T_1(X))$ $COV(Y,T_2(X)) < 0$ (Yitzhaki, 1990). Note, however, that those transformations can change the sign of the OLS regression coefficient, but they can't change the sign of the Gini semi-parametric regression coefficient. (See Yitzhaki, 1998).

e.2. Two monotonic non-decreasing transformations $T_3(Y)$ and $T_4(Y)$ so that $COV(T_3(Y),X)$ $COV(T_4(Y),X) < 0$.

e.3. Two monotonic non-decreasing transformations $T_5(Y)$ and $T_6(Y)$ so that

$COV(T_5(Y), F(X))\ COV(T_6(Y), F(X)) < 0$.

Properties (e.1) to (e.3) distinguish between the OLS and GMD regressions: (e.1) refers to a transformation of X, and it can only change the sign of the OLS regression coefficient, while (e.2) and (e.3) refer to transformations of Y and can change both the OLS and GMD regression coefficients

Property (b) implies that if the residuals ε and an explanatory variable X are statistically independent then the two ACCs that can be defined between the residuals and X (intended to describe the structure of $cov(\varepsilon, F(X))$ and $cov(X,F(\varepsilon))$ should be identical to the horizontal axis. In the sample they should oscillate randomly around the horizontal axis. Property (c) implies that by estimating a Gini regression, the area between the appropriate concentration curve and the horizontal axis is set to zero. (Minimization of the GMD of the residuals equates the area between the concentration curve of X with respect to e_M and the horizontal axis to zero, while the semi-parametric regression equates the area between the concentration curve of e_N with respect to X and the horizontal axis to zero). Once one area is equal to zero, the specification test is based on testing whether the area between the other concentration curve and the horizontal axis equals to zero as well. Hence, by plotting the appropriate curve one can see what causes the rejection or the failure to reject of the specification hypothesis by seeing whether the ACC has any pattern and in case of rejection, which observation(s) are responsible for the failure. An increasing/decreasing/flat curve means that the conditional mean (given the plotted explanatory variable) of the residuals is positive/negative/zero. Hence, one can see at which range(s) of the explanatory variable the fitted regression has missed.

Properties (d) and (e) are useful in order to explore the sensitivity of the regression coefficients with respect to monotonic transformations (Grether (1974); Yitzhaki (1990); Yitzhaki and Schechtman (2004)). These properties will be illustrated in the next section.

6. AN ILLUSTRATION: THE TWO EXPLANATORY VARIABLES CASE.

We start by expressing the multiple regression coefficients of the two explanatory variables case as explicit functions of the simple regression coefficients so that it will be clear how each simple regression coefficient affects the multiple regression coefficients. Next we discuss the logical inconsistency between measurement and policy instruments used in order to deal with the effect of family size on economic well-being, while the third section presents the empirical analysis intended to shed some light on this internal inconsistency.

6.1. The two explanatory variables case.

To be able to illustrate the properties of the Gini multiple regression in details and to investigate the roles of the different components, it is convenient to restrict ourselves to the two explanatory variables case. The first target of this section is to illustrate the similarity between the semi-parametric GMD regression and OLS regression. Generally, each variance is substituted by the appropriate GMD and each covariance is substituted by an appropriate co-Gini. As a result we can present the Gini and OLS multiple regression coefficients as solutions of the corresponding sets of linear equations, with the simple regression coefficients serving as constants. To complete the equivalence, it is shown in Yitzhaki (1996) that the simple regression coefficients under both methods are weighted averages of slopes defined between adjacent observations of the explanatory variables, so that the only difference between the two approaches is in the weighting schemes used to derive the simple regression coefficients. As a result, other terms in the GMD regression such as the partial correlation can be easily defined and explicitly expressed imitating the appropriate concepts in OLS.

Restricting (2.1) to two explanatory variables, the matrix a = r'x (which is the equivalent of x'x in OLS) is equal

$$a = \begin{pmatrix} \mathrm{cov}(x_1, r_1) & \mathrm{cov}(x_2, r_1) \\ \mathrm{cov}(x_1, r_2) & \mathrm{cov}(x_2, r_2) \end{pmatrix} \tag{6.1}$$

where r_i is the vector of ranks of the explanatory variable X_i. Note that the matrix **a** is not necessarily symmetric. Dividing each row by the GMD of the diagonal variable and solving the

linear equations, we get an explicit presentation which is identical to the OLS presentation. In order to show the similarity of b_N to OLS regression, let us rewrite (2.2) as follows:

$$\begin{pmatrix} b_{N01.2} \\ b_{N02.1} \end{pmatrix} = \frac{1}{1 - \Gamma_{12}\Gamma_{21}} \begin{pmatrix} 1 & -b_{N21} \\ -b_{N12} & 1 \end{pmatrix} \begin{pmatrix} b_{N01} \\ b_{N02} \end{pmatrix} \tag{6.2}$$

where $\Gamma_{12}\Gamma_{21}$ is the symmetric version of the Gini correlation,[11] b_{Nij} (i=0,1,2; j=1,2) indicates the simple Gini regression coefficient of variable i on j, with 0 denoting the dependent variable. The regression coefficients in the multiple regression are $b_{N0i.j}$.

To make the analysis easier, note that

$$b_{N12}b_{N21} = \frac{\mathrm{cov}(x_1, r_2)}{\mathrm{cov}(x_2, r_2)} \frac{\mathrm{cov}(x_2, r_1)}{\mathrm{cov}(x_1, r_1)} = \frac{\mathrm{cov}(x_1, r_2)}{\mathrm{cov}(x_1, r_1)} \frac{\mathrm{cov}(x_2, r_1)}{\mathrm{cov}(x_2, r_2)} = \Gamma_{12}\Gamma_{21} .$$

Using the above equation we rewrite the coefficients in the multiple regression as functions of the simple regression coefficients as follows:

$$\begin{pmatrix} b_{N01.2} \\ b_{N02.1} \end{pmatrix} = \frac{1}{1 - b_{N12}\, b_{N21}} \begin{pmatrix} b_{N01} - b_{N02}\, b_{N21} \\ b_{N02} - b_{N01}\, b_{N12} \end{pmatrix} . \tag{6.3}$$

Note that the denominator in (6.3) is always non-negative because:

$$b_{N12}b_{N21} = \Gamma_{12}\Gamma_{21} \le 1 .$$

Therefore, the sign of $b_{N0i.j}$ is determined by the sign of $(b_{N0i} - b_{N0j}\, b_{Nji})$.

6.2. The problem to be solved.

In this section we present an example to illustrate the properties of the Gini regressions, the specification tests, and the visual inspection of whether the association between random variables is monotonic. In this example there is an incompatibility between the measurement of performance of a policy intended to improve the income distribution and the policy instruments that the government uses. An empirical examination can help to decide which way is the "right" way. The issue is the following: when the problem of interest is to measure inequality in economic well-being, differences in family sizes are commonly taken into account by looking at income per capita or some kind of an equivalence scale. Almost all equivalence scales in use are

[11] See Schechtman and Yitzhaki (1987, 1999) for the properties of the Gini correlation.

based on dividing the income by a number which is a function of the family size. Viewing consumption per adult equivalent as representing economic well being we should expect consumption expenditures to be related to income and family size in a multiplicative relationship. On the other hand, most of the policy instruments in use in the income tax and benefits systems are based on adjustment of the tax to family size by giving a tax relief which is based on decreasing the tax (or increasing the benefits) by an amount which is only a function of the family size. We argue that those instruments represent an additive relationship between consumption and family size. It should be emphasized that we are not dealing with the normative issue of how the tax system should treat families of different sizes. The normative issue needs further research. All we deal with is the issue - are the way in which performance of tax systems in the area of reducing inequality is measured and the policy instruments used compatible? In the rest of this section we present the distinction between a multiplicative relationship and an additive one in a formal way.

The first step in the measurement of inequality in economic well being is to rank households according to economic well-being. One way of doing that is to use a **multiplicative** scale so that the ability to consume is defined as

$$E(Y,N) = E\left(\frac{Y}{a(N)}\right), \qquad (6.4)$$

where Y is the net income of the household, N is the family size, a(N) is the adult equivalent scale, and E(Y,N) is the equalized income that represents economic well-being. For example, the European Union is using the scale of $a(N) = N^{0.5}$ as its official scale. Feldstein (1976) principle of horizontal equity is that *"If two individuals would be equally well off (have the same utility) in the absence of taxation), they should also be equally well off if there is a tax"* (Italic at source, 1976, p-83). Our interpretation to the principle is that the ranking of families according to before tax economic well-being should be identical to the ranking of families according to after-tax economic well being. Would we want the tax and benefit function to keep Feldstein's principle of horizontal equity, the structure of the tax and benefit function should have been

$$T(N,Y) = a(N)\, t\left(\frac{Y}{a(N)}\right) \quad , \qquad (6.5)$$

where T(N,Y) is the total tax minus the benefits that the family receives and t( ) is the tax function defined over adult equivalent income. [12],[13] On the other hand, when looking at tax and benefit systems, it turns out that most countries rely on an additive scale. This is the case whenever child allowances or exemptions are used to handle family size. In the case of exemptions, the structure of the tax function is:

$$T(N,Y) = T(Y-E(N)). \qquad (6.6)$$

That is, to adjust the tax to family size, a constant amount of E(N) is deducted from before tax income, as it is for example in the U. S.. In the case of tax-exempt allowances, the tax function is:

$$T(N,Y) = T(Y) - AL(N), \qquad (6.7)$$

where AL is the allowance, as is the case in Israel and Britain. In both cases (6.6) and (6.7) it is as if the state recognizes a given amount that should be added to the income to keep horizontal equity intact.

We argue that (6.4) on one hand and (6.6) and (6.7) on the other hand are incompatible because they violate Feldstein's principle of horizontal equity. To shed some empirical light on this issue it is worth to check whether the Engel curve, which relates consumption to income and family size, is additive or multiplicative. This means that we view consumption as representing economic well being and household's size as representing needs. Assume that the appropriate specification of the Engel curve is linear. That is,

$$\hat{C} = \alpha + \beta Y + \gamma N. \qquad (6.8)$$

Then it is reasonable to argue that the effect of an additional family member on consumption is to increase consumption by a constant and then the appropriate tax adjustment should be of the exemption or allowance type. If, on the other hand, the appropriate specification of the Engel curve is multiplicative, that is of the type:

$$\ln(\hat{C}) = \alpha + \beta \ln(Y) + \gamma \ln(N) \qquad (6.9)$$

then the effect of an additional family member on consumption is to increase consumption by a given percentage, which can be viewed as supporting a tax function of the multiplicative type as

[12] In some applications the model used is $T(N,Y) = N\, t\left(\frac{Y}{a(N)}\right)$, so that each member of the household is counted as one. (See Ebert (2005) and Yoram Ben-Porath's comment in Bruno and Habib (1976)).

[13] The French tax system resembles this structure.

in (6.5). Our empirical research question is to choose between the two alternative specifications – (6.8) or (6.9) as representing the appropriate specification of the Engel curve.

6.3. Empirical findings

To illustrate the properties of the Gini regression and the tools it provides to the investigator, we use Israeli Survey of households' Expenditures, 2008. (For a description of the sample, see Central Bureau of Statistics (2009), S.P. 1363). The data consists of 5971 observations. Each observation includes a weight which represents its weight in the population. Consumption includes the depreciation and value of forgone interest on capital invested in housing and vehicles. Income is after-tax overall income, which includes money income *plus* in-kind income *minus* income tax and social security taxes.

The structure of the empirical illustration is the following: we first present the simple regression coefficients that are the basic components of the multiple regression coefficients and only later we present the multiple regression coefficients. In our presentation, our main interest is to find out whether the relationships between the variables are monotonic. The final stage is to see which model fits the data better: the additive or the multiplicative? We start with the relationship between the dependent variable and the explanatory variables.

Table 6.1 presents the simple regression coefficients between consumption expenditures and after tax (net) income using linear and multiplicative specifications. For comparison we also present the OLS estimates.

Table 6.1: Simple OLS and Gini regression coefficients –consumption as a function of income*

	OLS			Gini					
Model	a	b	R^2	a (mean)	a (median)	b	$R(y, \hat{y})$	$R(\hat{y}, y)$	GR
Linear	4756	0.533 (0.000)	0.514	3505	2641	0.621 (0.011)	0.792	0.803	0.346
Multiplic.	4.216	0.538 (0.000)	0.489	2.735	2.707	0.698 (0.014)	0.811	0.791	0.372

* Standard errors in parentheses. In Gini regression the standard errors were calculated using Jackknife fast method.[14] Standard errors of the OLS are rounded to three decimal points.

[14] In using the jackknife method in a regression context there are two options: when dropping

The left-hand part of Table 6.1 presents the OLS simple regression coefficients of the additive and multiplicative specifications (Equations (6.8) and (6.9)). The right-hand side presents the semi-parametric Gini regression equivalents. Because we are only interested in the components of the multiple regression, the standard errors for the constant term are not estimated. As can be seen from Table 6.1, the marginal propensity to spend, i.e., the simple regression coefficient, is smaller for the OLS than for the Gini in both specifications. This is a result of a combination of two factors: the Gini regression tends to give lower weights to extreme observations, and the marginal propensity to spend tends to decline with income. The constant terms of the Gini regressions are estimated in two ways depending on whether the regression passes through the median of the dependent variable or through the mean. The median constant is lower than the mean constant, especially for the linear regression. This indicates that the residuals tend to be larger for high income groups. It should be mentioned that the standard errors of the Gini and OLS regression coefficients are not comparable because for the Gini it is assumed that the explanatory variable is a random variable, while under regular OLS software it is assumed that only the error term is random. Also note that the measures of goodness of fit are not automatically comparable. In order to make them comparable one should look at R for OLS, rather than at R^2. Figure 6.1 presents the LMA of consumption vs income (Line of Independence Minus Absolute Concentration Curve).The curve is concave and smooth, and does not intersect the horizontal axis implying that the relationship between consumption and income is monotonically increasing so that there is no monotonic transformation of consumption that can change the sign of the regression coefficient.

---

an observation from the sample, should one re-estimate the whole model again or is it permissible to drop an observation and to evaluate the effect on the regression coefficient. The former approach seems to be the appropriate one but it is time consuming and the difference between the two methods seems negligible. By fast method it is meant that the model was not re-estimated.

Figure 6.1: LMA curve of consumption as a function of Income

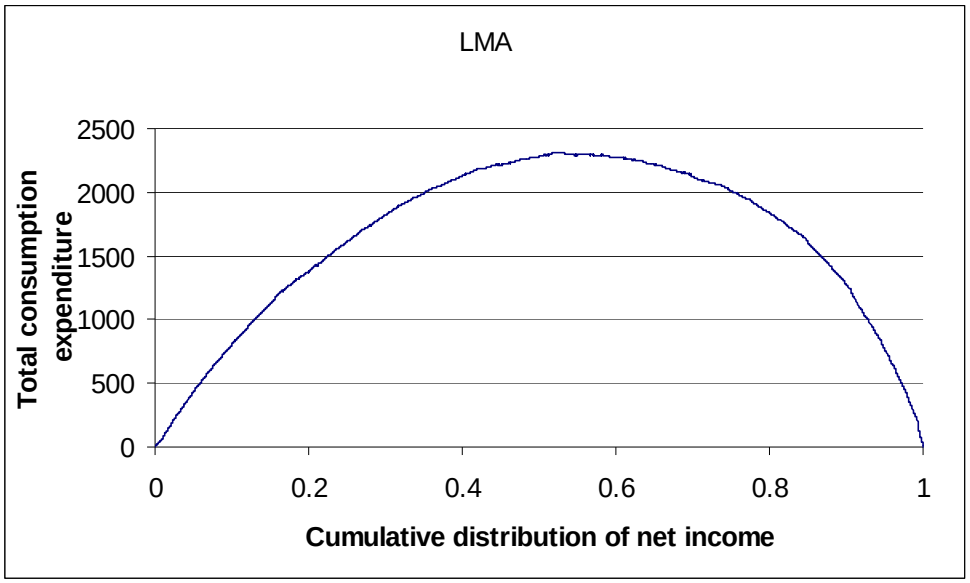


Table 6.2 is similar to Table 6.1, except that this time the regression is with respect to family size. Here, the results are a bit different than in Table 6.1. Under the linear specification, having an additional person increases consumption by 1552 according to the Gini regression while under OLS the amount is smaller and is equal to 1251. On the other hand, under the multiplicative specification, the OLS estimates of the percentage increase in consumption due to an increase in the household size is 0.49 which is lower than the Gini (0.47).

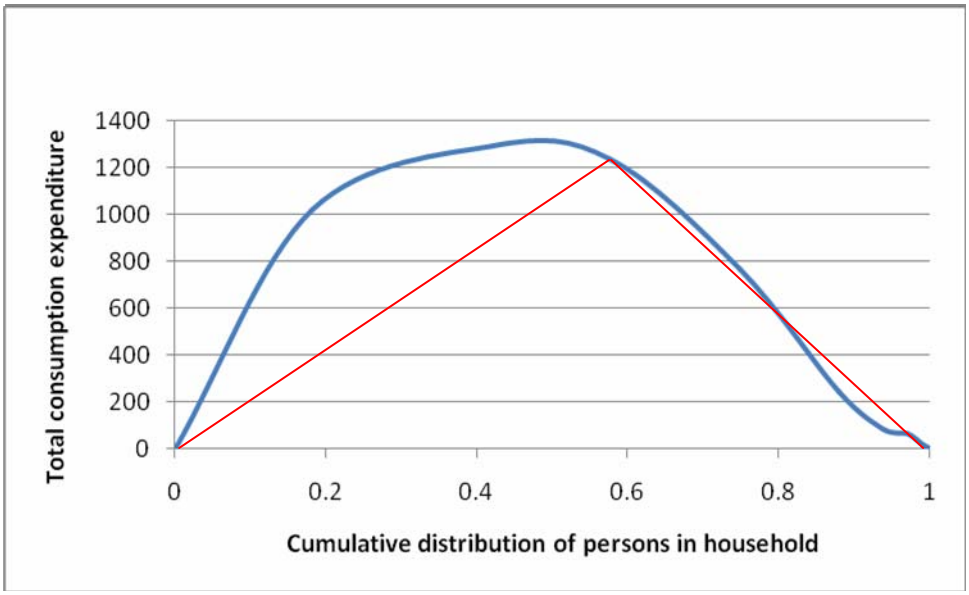
Table 6.2: Simple OLS and Gini regressions- consumption as a function of Household's Size*

	OLS			Gini					
Model	**a**	**b**	**R^2**	**a (mean)**	**a (median)**	**b**	**$R(y, \hat{y})$**	**$R(\hat{y}, y)$**	**GR**
Linear	8191	1251 (2.686)	0.093	7195	5674	1551.6 (59.47)	0.399	0.399	0.056
Multiplicative	8.73	0.486 (0.001)	0.217	8.742	8.766	0.474 (0.015)	0.459	0.463	0.110

* Standard errors in parentheses. In Gini regression standard errors were calculated using Jackknife fast method.

Figure 6.2 offers an insight to the results. Unlike the curve in Figure 6.1 which is concave and smooth, Figure 6.2 changes from a concave to a convex curve. This means that while the overall regression coefficient is positive (as seen in Table 6.2), it is positive among small households, but it is negative among the largest households as is illustrated next: using the Gini linear specification and estimating the regression coefficient for households of size 4 and above, which include 42 % of the observations, we get β= -173.5, while restricting the regression to households of size 5 and larger, which amount to about 25 % of the observations, we get an even smaller slope, β= - 827.3.

Figure 6.2: LMA curve of consumption as a function of Household's size



Note, however, that the curve does not cross the horizontal axis, which implies that no monotonic transformation of household's size can change the sign of the regression coefficient. The explanation to this result lies in an additional decomposition that one can perform, which decomposes the regression coefficient into two components: a within-component (intra) and a between-component (inter). (See Yitzhaki and Schechtman, 2009). Assume that one divides the range of the explanatory variable into two sections, one section is composed of the sixty percent of the smallest households, while the other section is composed of the remaining 40 percent of the largest households. Then the overall regression coefficient can be expressed as a weighted sum of intra and inter section regression coefficient weighted by the appropriate measure of variability used (variance of X for OLS; GMD of X for GMD regression). The between-group

component is reflected by the triangle which starts at the origin, reaches the curve at the end of the first section and ends up on the horizontal axis at one. The intra-group components are reflected by the areas enclosed between the curve and the edges of the triangle. In our example (Figure 6.2) the between-group component contributes to the overall regression coefficient more than the intra-group components. Therefore it determines its sign. However, it should be clear that no linear model can explain such a pattern.

We now turn to describe the regression coefficients between the explanatory variables. Table 6.3 presents that last elements needed for the multiple regression coefficients.

Table 6.3: The simple regression coefficients between the explanatory variables

Household's size as a function of income

Model	OLS			Gini					
	a	b	R^2	a (mean)	a (median)	b	R(y, $\hat{y}$)	R($\hat{y}$,y)	GR
Linear	2.713	0.000042 (0.000)	0.055	2.393	1.96	0.000065 (0.000)	0.321	0.343	0.031
Multip.	-1.583	0.28 (0.000)	0.144	-2.113	-2.087	0.3366 (0.0117)	0.404	0.411	0.062

Income as a function of household's size

Model	OLS			Gini					
	a	b	R^2	a (mean)	a (median)	b	R(y, $\hat{y}$)	R($\hat{y}$,y)	GR
Linear	9961.86	1285.38 (3.687)	0.055	8584	6406	1709 (82.4)	0.343	0.321	0.026
Multip.	8.781	0.514 (0.001)	0.144	8.802	8.863	0.4938 (0.024)	0.411	0.404	0.080

* Standard errors in parentheses. In Gini regression standard errors were calculated using Jackknife fast method. See footnote 14 and Yitzhaki (1991).

It is worth mentioning that in the OLS the ratio between a regression coefficient and the regression coefficient in a reversed regression is equal to the ratio of the variances. (i.e., in OLS b_{yx} /b_{xy} = Var(y)/Var(x)) (Goldberger, 1984). In the Gini regression no such relationship has to hold. And at least in theory, they can even have different signs! Figures 6.3 and 6.4 present the LMA curves for the additive model. Figure 6.3 which portrays net income as a function of the

size of the household indicates that while for small households the conditional expected value of income is increasing, this sign of the regression coefficient changes to a negative one among 45 percent of the largest households. Moreover, there is a small range (between 90-th to 95-th percentiles of household's size) in which the curve is below the horizontal axis, which means that a monotonic transformation that shrinks all the rest of the range of household's size will yield a negative regression coefficient in the OLS. Also, there is a transformation of income that can change both OLS and Gini regression coefficients. Obviously this will not be an acceptable treatment of the data. Figure 6.4, on the other hand, which presents household size as a function of income, is a concave and smooth curve.

Figure 6.3: LMA curve of Income as a function of Household's size

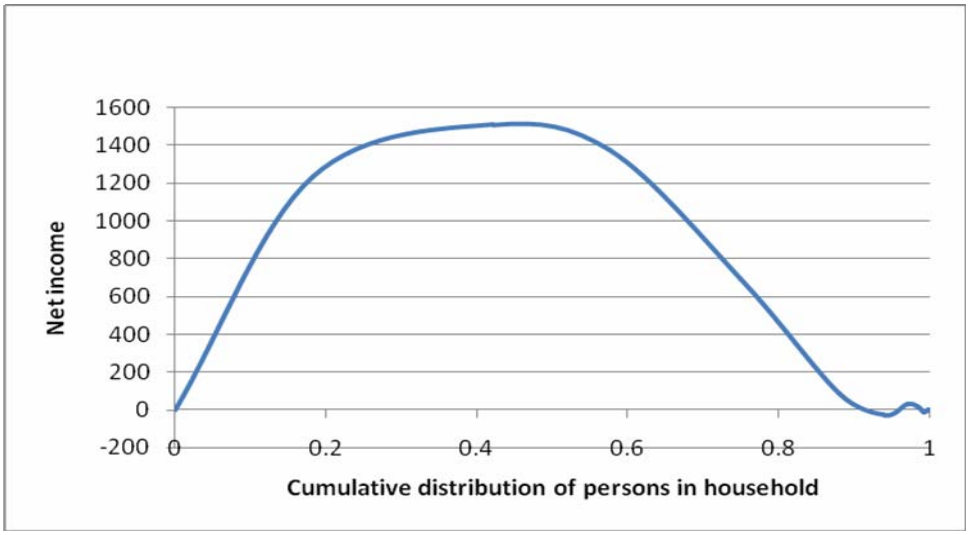
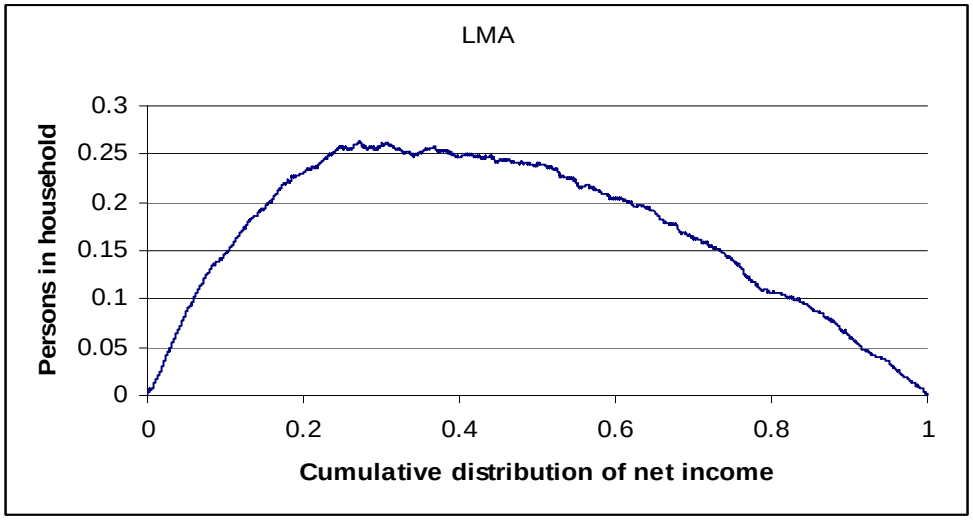


Figure 6.4: LMA curve of Household's size as a function of Income



As far as we can see, no simple model can explain such results. One possible explanation is that there are two models of behavior: the fertility in one group follows the regular pattern of bringing children to the world subject to having the ability to support them, while the fertility in the other group is not related to income. Our conclusion is that no simple model can explain such a curve. Further research is needed to explain those results.

Figure 6.5: LMA curve of Ln Income as a function of Ln Household's size

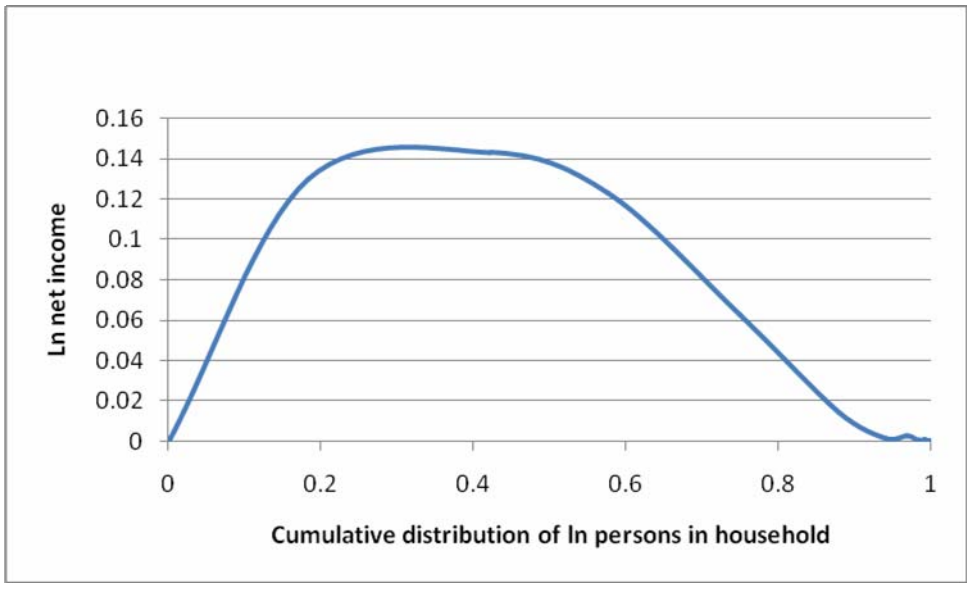


Figure 6.5 is added in order to explain the difference between the additive and the multiplicative models: since the horizontal axis portrays the cumulative distribution, it is the same as the horizontal axis of Figure 6.3. Hence the only difference between the additive and the

multiplicative model is that the latter shrinks the inconvenient deviations that existed in the additive model. As a result, in the latter model there is no monotonic transformation that can change the sign of the regression coefficient. However, the negative relationship in income as a function of the household's size for large households continues to hold. This is an indication that we should expect a better fit of the multiplicative model than the additive one.

Having described the components of the multiple regression coefficients, we now move to present the results of the multiple regression. Table 6.4 presents the results.

Table 6.4: The multiple regression: consumption as a function of income and household's size

Model	OLS				Gini						
	a	b (income)	b (size)	R^2	a (mean)	a (med.)	b (income)	b (size)	$R(y, \hat{y})$	$R(\hat{y}, y)$	GR
Linear	3132	0.508 (0.000)	598.6 (1.981)	0.534	2173	1346	0.585 (0.011)	556.4 (41.01)	0.801	0.813	0.363
Multip.	4.603	0.47 (0.000)	0.245 (0.001)	0.536	3.064	3.037	0.645 (0.013)	0.156 (0.012)	0.823	0.805	0.394

* Standard errors in parentheses. In Gini regression standard errors were calculated using Jackknife slow method.[15]

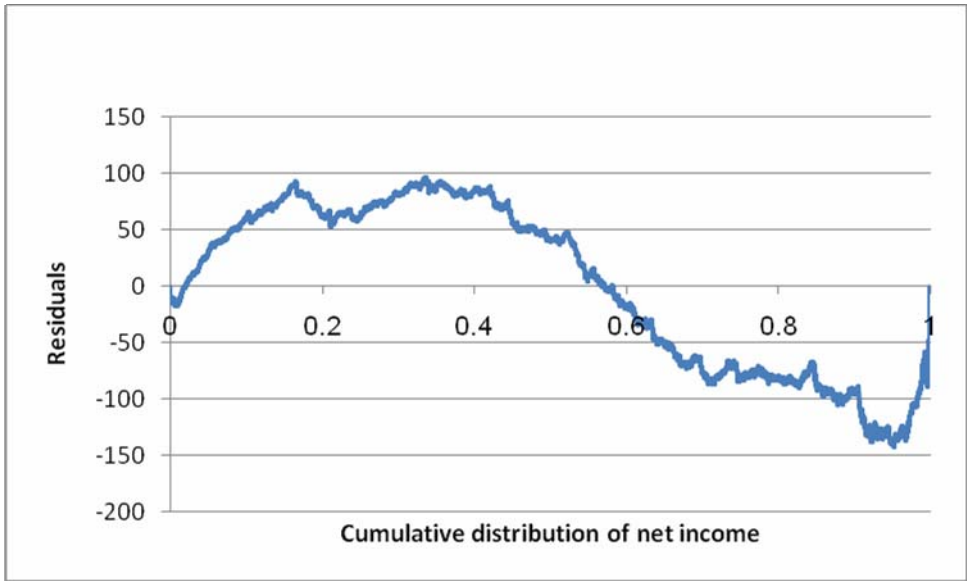
Comparison of the regression coefficients between the OLS and the Gini in the additive and multiplicative models indicates that the marginal propensity to spend is smaller under the Gini than under OLS regression, with the multiplicative model showing larger differences. On the other hand, both specifications indicate that the effect of an additional member in the household is larger under OLS than under the Gini method. The difference between the estimates of the simple regression coefficient and the parallel estimates in the multiple regression case indicates the effect of the association between the explanatory variables. Would the explanatory variables be statistically independent, then there should have been no difference between the two. However, when the explanatory variables are correlated, and the relationship is not linear then the effect of the correlation may be different under OLS and Gini regressions. The fact that

[15] Under the slow method, each time an observation was dropped from the sample, the model was re-estimated.

the two methods yield regression coefficients that are different calls for a further inspection of the way that the models fit the data.

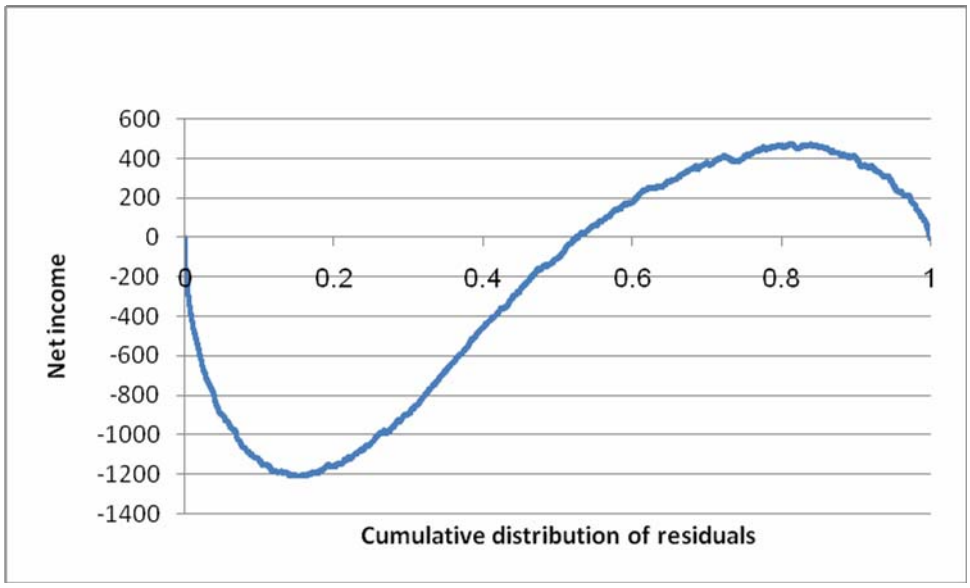
Figure 6.6 presents the LMA curve of the residuals as a function of net income. By construction, the area between the curve and the horizontal axis is equal to zero (recall that $cov(e_N , F(X))=0$). A perfect fit of the model to the data will result in a curve which oscillates randomly around the horizontal axis. As can be seen, this is unlikely the case. For the lower 55 percent of observations of income (approximately) there is a positive (Gini and Pearson) correlation between the residuals and income, while the highest 45 percent of the observations on income reveal a negative correlation.

Figure 6.6: LMA of residuals as a function of income: Linear Specification



Note that by construction, the area enclosed between the curve and the horizontal axis equals to zero.

Figure 6.7: LMA of Income as a function of the residuals



To check the quality of the specification Figure 6.7 presents the LMA curve of income as a function of the residuals. Results show that Figure 6.7 is almost a mirror image of Figure 6.6. For small values of residuals the conditional correlation is negative, while for large values of residuals the correlation is positive. Overall, $cov(x,r(e_N))$= -505.44. To test whether the specification of the model is correct we estimated the (simple) Gini regression coefficient of income on the residuals and found that the regression coefficient $b_{x,r(e)}$ =-0.194, $\hat{\sigma}_b$ = 0.042, and since the estimator of the regression coefficient is approximately normally distributed (Schechtman *et al.* , 2008b), it turns out that the value of the test statistics is Z=-4.609 and the linearity of the model with respect to income is rejected.[16]

We now inspect the quality of the specification of the model with respect to household's size. Figure 6.8 presents the LMA curve of the residuals as a function of household's size. Again, we remind the reader that the area enclosed between the curve and the horizontal axis is zero by construction. Similar to the case of income, there is a positive correlation between the residuals at low levels of household's size and negative correlation for large household's size. However, for about 25 percent of the observations of middle size households the curve is horizontal and close to the horizontal axis, indicating a good fit of the model to the observations.

Figure 6.8: LMA curve - linear specification: household's size

[16] The properties of the estimators are presented in Schechtman and Yitzhaki (1987) and Schechtman *et al*. (2008).

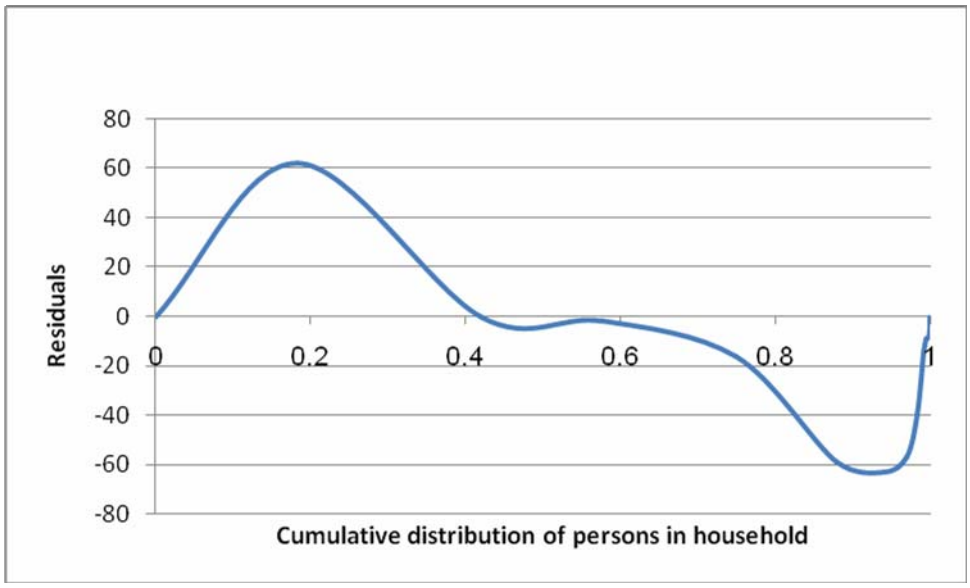


Figure 6.9: LMA of household size as a function of residuals: linear specification

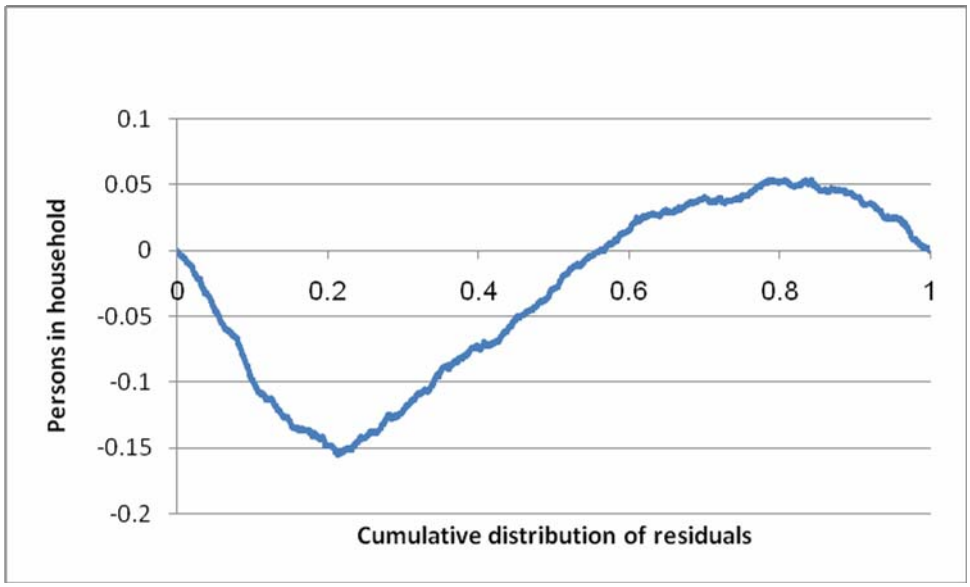


Figure 6.9 presents the LMA curve of household's size as a function of the residuals. Again we get a mirror image of Figure 6.8 although the quality of the "mirror" is worse than the one we got when we dealt with income. For small values of residuals we got a negative correlation between household's size and residuals, while for large values of residuals we got a positive correlation. To test for the quality of the linear specification, we ran a (Gini simple)

regression of household's size on the residuals. The estimated Gini regression coefficient is -0.0000247, its standard error is estimated to be 4.5197E-11, so that the test statistics indicate that the linear specification is rejected. [17]

We now turn to inspect the multiplicative specification.

Figure 6.10: Multiplicative specification: residuals as a function of Ln(net income)

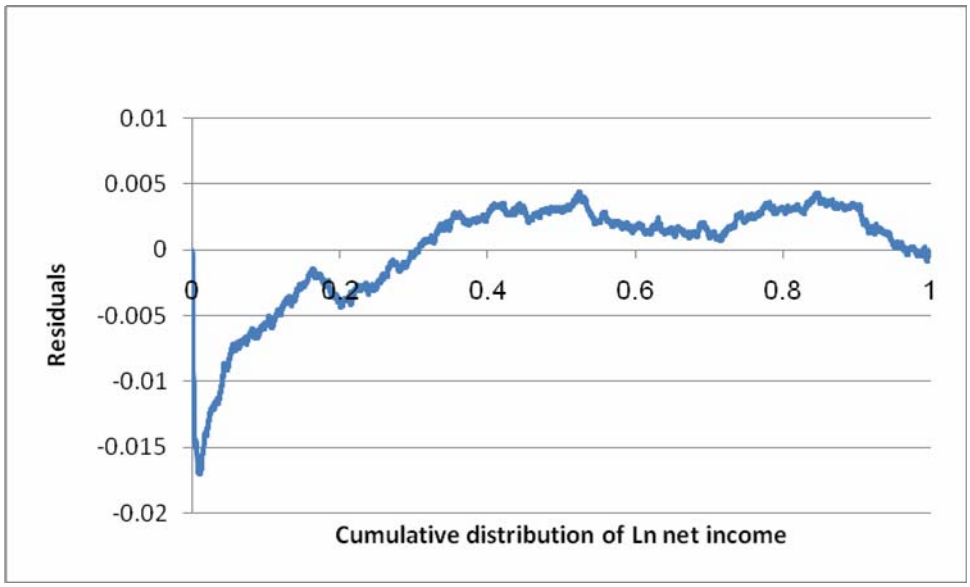


Figure 6.11: LMA curve of ln(net income) as a function of the residuals

[17] There are two problems with these results. The first problem is that household's size is a discrete variable. In this case there is a mismatch between the LMA curve and the definition of cumulative distribution, because the empirical cumulative distribution is defined as a step function while in an LMA (and Lorenz) curves one connects different points of the curve by a straight line, which implies continuity. (See Schechtman and Yitzhaki, 2008). The other problem is the issue of rounding errors because of small numbers involved. Therefore one should be careful in interpreting this result. Further research is required to resolve this issue.

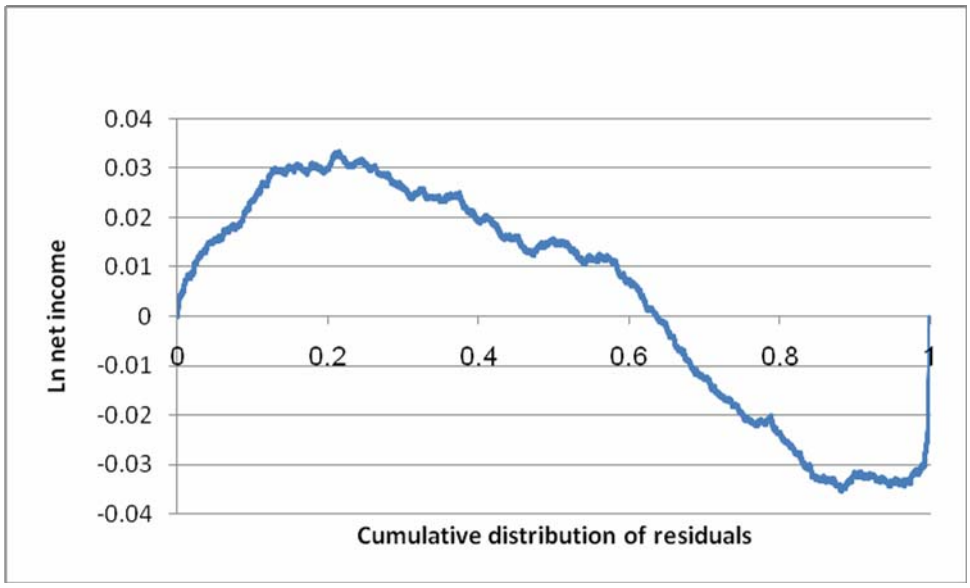


Figure 6.10 presents the LMA curve of the residuals as a function of ln(net income). Note that because the horizontal axis portrays the cumulative distribution of Ln(net income) and the cumulative distribution of net income is not affected by monotonic increasing transformation, the horizontal axes in Figure 6.10 and Figure 6.6 are identical. The difference is only in the vertical axes. Comparisons of the two figures reveal that while in Figure 6.6 low levels of income are associated with negative correlation with the residuals, the multiplicative specification has changed the sign and the order of correlations. Except for the lowest 20 percent of observations of income the curve seems flat indicating a low level of correlation between the residuals and income. Figure 6.11 presents the LMA curve of Ln(net income) with respect to residuals. For low level of residuals the correlation is positive while for high level of residuals the correlation is negative. The estimated cov(x,r(e))= -0.00803, the regression coefficient of the simple (Gini) regression of ln(income) as a function of the residuals is $b_{x,r(e)}$ =-0.0365, the estimated standard error is $\hat{\sigma}_b$ = 0.047, the test statistics is Z=-0.772, so that we fail to reject the hypothesis that the model is linear with respect to income. [18]

We turn now to the specification of the multiplicative model with respect to household's size. Figure 6.12 presents the LMA curve of the residuals as a function of the household's size. As can be seen, for low levels of household's size the correlation is negative, for middle sized households it is positive and for large households it is again negative. Note that by construction, the overall area between the curve and the horizontal axis is equal to zero.

[18] Standard errors were calculated using Jackknife fast method.

Figure 6.12: multiplicative specification: household's size

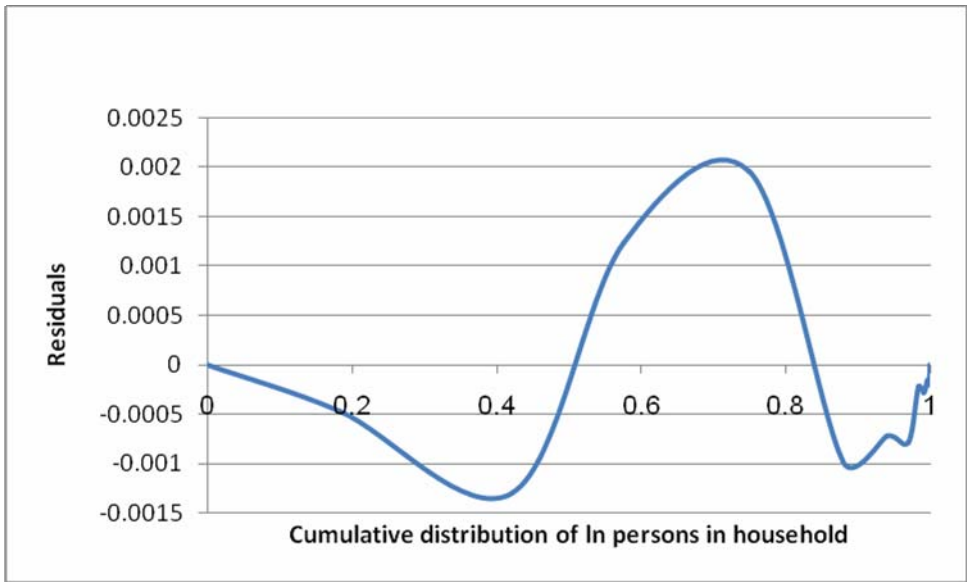
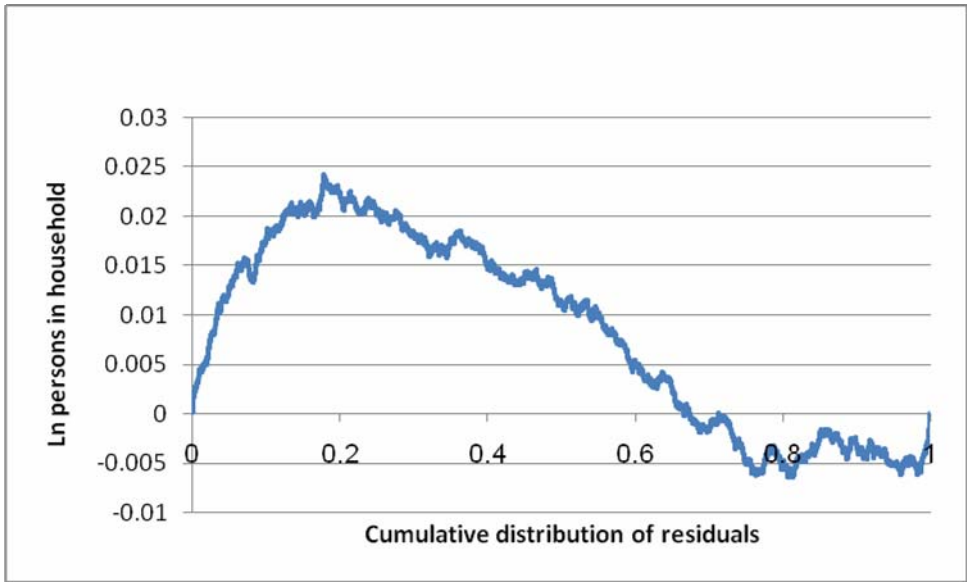


Figure 6.13: LMA curve of household's size with respect to residuals



One can observe that the (Gini) correlation between household's size and residuals is positive for small households and slightly negative for large ones. The Gini covariance is

cov(x,r(e))= 0.01644, the (Gini simple) regression coefficient is $b_{x,r(e)}$ =0.0748, its standard error is $\hat{\sigma}_b$ = 0.026, so that the test statistics is Z=2.907. This means that although we have rejected the specification with respect to family size in the multiplicative model too, the multiplicative model fits the data better than the additive one.

To conclude the empirical examination, we have found the multiplicative model to fit the data better than the additive one. If one has to choose between Equations (6.8) and (6.9) then (6.9) is supported by the data better than (6.8).

7. CONCLUDING REMARKS

In this paper we have extended the simple regression based on Gini's mean difference into a multiple regression framework. Similar to the simple regression case, the multiple Gini regression offers two kinds of regressions. The first is a semi parametric regression, which is an imitation of the OLS regression, and like it, the estimator can be explicitly written. The advantages are that no model has to be specified, and it is less sensitive to outliers than the OLS regression (since it is based on ranks). The other regression is based on minimization of the Gini of the residuals.

The combination of the two methods offers a built-in specification test. It is based on the fact that the Gini method has two covariances between each pair of random variables. In estimating a regression model one covariance between the residuals and each explanatory variable is set to zero, so that the other covariance can be used as a test for the specification of the model. One can start by estimating a linear approximation to the regression curve, and if one wants to use it for prediction, then the prediction will be restricted to the variables in which the model is linear.

The connection between the Gini parameters and concentration curves enables one to verify monotonicity of the regression curve. The importance of verifying the monotonicity is stressed by Heckman, Urzua, and Vytlacil (2004) in the context of an instrumental variable. In a Gini regression framework, monotonicity is not an assumption imposed on the data. It is a property that says that the regression coefficients on each section of the data have identical signs. It is up to the judgment of the reader to decide whether this requirement is violated, and to what extent. Schechtman *et al*. (2008a) suggest the forms of statistical tests that will enable the user to test for the intersection of absolute concentration curves.

The basic OLS regression has many refinements. The similarity between the OLS and the semi-parametric Gini regression gives the hope that many of the refinements of the OLS can be developed for the Gini regression as well. A first step in this direction is offered in Yitzhaki and Schechtman (2004), where the instrumental variable approach is developed in a GMD framework. An additional tool is the decomposition of the Gini (ANOGI) (Frick *et al.*, 2006) which enables the user to imitate the decomposition of the variance (ANOVA) and to get an additional property – the stratification in a distribution. ANOGI is the tool that enables the decomposition of the Gini regression coefficient into the contributions of different sections of the explanatory variable, as discussed in Yitzhaki and Schechtman (2009). Dividing the residuals into positive and negative groups and applying ANOGI enables the user to get further insights about the distribution of the negative and positive residuals along the range of the explanatory variable. Further research is needed to fully utilize the properties of the Gini in regression analysis.

References

Banks, J., Blundell, R., and Lewbel, A. (1997). Quadratic Engel curves and consumer demand, *Review of Economics and Statistics*, 79, 4, (November), 527-539.

Bassett, G., Jr. and Koenker R. (1978). Asymptotic theory of least absolute error regression, *Journal of the American Statistical Association,* 73, 618-622.

Bowie, C. D. and Bradfield, D. J. (1998). Robust estimation of beta coefficients: Evidence from small stock market, *Journal of Business Finance & Accounting*, 25(3)&(4), April/May, 439-454.

Bruno, M. and J. Habib (1976). Taxes, family grants and redistribution. *Journal of Public Economics,* vol. 5, pp. 57-79.

Central Bureau of Statistics (2009). Family expenditure survey 2008 Special Series No. 1363, Jerusalem, Israel.

D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples, *Biometrika*, 58, 341-48.

D'Agostino, R. B. (1972). Small sample probability points for the D test of normality, *Biometrika*, 59, pp 219-221.

Ebert, U. (2005). Optimal anti poverty programmes: horizontal equity and the paradox of targeting, *Economica*, 72, 453-468.

Eubank, R. L. and Spiegelman, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques, *Journal of the American Statistical Association*, 85, 387-392.

Frick, R. J., Goebel, J., Schechtman, E., Wagner, G., and Yitzhaki, S. (2006). Using analysis of Gini (ANOGI) for detecting whether two sub-samples represent the same universe: the German Socio-Economic Panel Study (SOEP) experience, *Sociological Methods and Research*, 34, 4, (May), 427-468.

Feldstein, M. (1976). On the theory of tax reforms, *Journal of Public Economics*, 6, 77-104.

Goldberger, A. S. (1984). Reverse regression and salary discrimination, *The Journal of Human Resources*, 19, 293-319.

Grether, D. M. (1974). Correlation with ordinal data, *Journal of Econometrics*, 2, 241-246.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivative, *Journal of the American Statistical Association*, 84, 408, Theory and Methods, 986-995.

Heckman, J. J. (2001). Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture, *Journal of Political Economy*, 109 (4), 673-748.

Heckman, J. J., Urzua, S. and Vytlacil, E. (2004). Understanding instrumental variables in models with essential heterogeneity, mimeo. http://jenni.ucicago.edu/underiv/.

Hettmansperger, T. P. (1984). *Statistical Inference Based on Ranks*, New York: John Wiley & Sons.

Jurečková , J. (1969). Asymptotic linearity of a rank statistic in regression parameter, *Annals of Mathematical Statistics*, 40, 1889-1900.

Jurečková ,J. (1971). Nonparametric estimates of regression coefficients, *Annals of Mathematical Statistics*, 42, 1328-1338.

Kakwani, N. C. (1980). *Income Inequality and Poverty*, Oxford: Oxford University Press.

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

Kozek, A. S. (1990). A nonparametric test of fit of a linear model, *Communications in Statistics - Theory and Methods*, 19(1), 169-179.

Lewbel, A. (1995). Consistent nonparametric hypothesis tests with an application to Slutsky symmetry, *Journal of Econometrics*, 67, 379-401.

McKean, J. W. and Hettmansperger, T. P. (1978). A robust analysis of the general linear model based on one step R-estimates, *Biometrika*, 65, 571-579.

Neill, J. W. and Johnson, D. E. (1984). Testing for lack of fit in regression - a review, *Communications in Statistics - Theory and Methods*, 13(4), 485-511.

Olkin I. and Yitzhaki, S. (1992). Gini regression analysis, *International Statistical Review*, 60, 2, (August), 185-196.

Rilstone, P. (1991). Nonparametric hypothesis testing with parametric rates of convergence, *International Economic Review*, 32, 1 (February), 209-227.

Schechtman, E. and Yitzhaki, S. (1987). A measure of association based on Gini's mean difference, *Communications in Statistics - Theory and Methods*, A16, 1, 207-231.

Schechtman, E. and Yitzhaki, S. (1999). On the proper bounds of the Gini correlation, *Economics Letters*, 63, 133-138.

Schechtman, E. and Yitzhaki, S. (2008). Calculating the extended Gini coefficient from grouped data: a covariance presentation approach. *Bulletin of Statistics & Economics*, 2, S08, (Spring), 64-69.

Schechtman, E., Shelef, A., Yitzhaki, S. and Zitikis, R. (2008a). Testing hypotheses about absolute concentration curves and marginal conditional stochastic dominance. *Econometric Theory*, 24, 1044-1062.

Schechtman, E., Yitzhaki, S. and Artzev, Y. (2008b) Who does not respond in the household expenditure survey: An exercise in extended Gini regressions. *Journal of Business & Economics Statistics*, 26, Number 3, pp. 329-344.

Yitzhaki, S. (1990). On the sensitivity of a regression coefficient to monotonic transformations, *Econometric Theory*, 6, No. 2, 165-169.

Yitzhaki, S. (1991). Calculating jackknife variance estimators for parameters of the Gini method, *Journal of Business & Economic Statistics*, 9, No. 2, (April), 235-9.

Yitzhaki, S. (1996). On using linear regressions in welfare economics, *Journal of Business & Economic Statistics*, 14, 4, 478-486.

Yitzhaki, S. (1998). More than a dozen alternative ways of spelling Gini, *Research on Economic Inequality*, 8, 13-30.

Yitzhaki, S. (2003). Gini's mean difference: A superior measure of variability for non-normal distributions, *Metron*, LXI, 2, 285-316.

Yitzhaki, S. and Olkin, I. (1991). Concentration curves and concentration indices, in Karl Mosler and Marco Scarsini (eds.) *Stochastic Orders and Decisions under Risk*, Institute of Mathematical Statistics: Lecture-Notes Monograph Series Vol 19, 380- 392.

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

Yitzhaki, S. and Schechtman, E. (2009). Identifying monotonic and non-monotonic relationships. Paper presented in ISI 57th Meeting Proceedings: *Development of Progress in Applied Statistical Modelling for the Economic Sciences Using Non-paramtric Regression Methods*, Statistics South Africa.
www.statssa.gov.za/isi2009/ScientificProgramme/IPMS/1665.pdf

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

WORKING PAPER SERIES

No.53

Gini's multiple regressions: two approaches and their interaction

Shlomo Yitzhaki[*], Edna Schechtman[**], Taina Pudalov[***]

September, 2010

*Central Bureau of Statistics and Hebrew University.
** Ben Gurion University.
*** Central Bureau of Statistics.