December 2012

Gini's Mean Difference offers a response to Leamer's critique

By
Shlomo Yitzhaki

Abstract

Gini's mean difference has decomposition properties that nest the decomposition of the variance as a special case. By using it one may reveal the implicit assumptions imposed on the data by using the variance. I argue that some of those implicit assumptions can be traced to be the causes of Leamer's critique. By requiring the econometrician to report whether those assumptions are violated by the data, we may be able to offer a response to Leamer's critique. This will reduce the possibility of supplying "empirical proofs" which in turn may increase the trust in econometric research.

Key Words: Gini, variance, regression

Mailing address:
Shlomo Yitzhaki
Dept of Economics
Hebrew University
E-mail: shlomo.yitzhaki@huji.ac.il

Gini's Mean Difference offers a response to Leamer's critique

A popular method in reaching quantitative conclusions in economics is the method of regression. The most popular one is the Ordinary Least Squares (OLS), which is based on the properties of the variance. To simplify the presentation, I will restrict the arguments to the OLS. Comments on other methods will be offered in a separate section.

In an influential paper, Edward Leamer (1983) criticized the fragility of econometric estimates. He urged fellow econometricians to study the reasons causing the fragility of econometric research, .."lest we lose our customers in government, business, and on the boardwalk of Atlantic city." (p-3).

Almost thirty years later, the state of the art and the major points of controversy are best summarized in a recent paper by Angrist and Pischke (2010).

"Just over a quarter century ago, Edward Leamer (1983) reflected on the state of empirical work in economics. He urged empirical researchers to “take the con out of econometrics” and memorably observed (p. 37): “Hardly anyone takes data analysis seriously. Or perhaps more accurately, hardly anyone takes anyone else’s data analysis seriously.” Leamer was not alone … Perheps credible empirical work in economics is a pipe dream. " (p-1).

Angrist and Pischke description of Leamer's argument is:

"Leamer (1983) diagnosed his contemporaries’ empirical work as suffering from a distressing lack of robustness to changes in key assumptions — assumptions he called “whimsical” because one seemed as good as another. The remedy he proposed was sensitivity analysis, in which researchers show how their results vary with changes in specification or functional form.” (p-1).

Angrist and Pischke response to Leamer's critique is in listing the improvements in research design, better data collection, better definitions of the research question, and more. I do not deny the improvements pointed out.

However, as far as I can see, the methodology of estimation has not changed in a qualitative way. More computer power allows more complicated modeling and data mining. But some of the assumptions that may cause the results and it is not clear whether they are supported by the data, are still there.

To see whether Leamer’s criticism is still valid let me ask the following question: Is it possible that two investigators, using the same data and an identical model, can reach opposite conclusions concerning the effect of one variable on another?

Golan and Yitzhaki (2010) supply a positive answer to this question. They show that if one uses Gini regression (Gini's Mean Difference, or GMD) and the other OLS regression, then the signs of some regression coefficients may differ.

Clearly, if the researchers use different functional forms then they have a higher chance of reaching contradicting conclusions. Since, in a typical econometric research, the investigator may run thousands of regressions, and presents only a few regressions, one should view econometric research as a proof that the model presented can be supported by the data and not as representative of the evidence in the data.

The variance is cov(X, X), while one possible presentation of the GMD is cov(X, F(X)), where F(X) is the cumulative distribution function. The difference between the GMD and the variance is in the metric used to define variability. The former relies on the "city block" metric and the latter on the Euclidean metric. I argue that the "city block" metric imposes less whimsical assumptions on the data and has two advantages in econometric research: (a). The money metric that governs the budget constraint belongs to the city block metric. (b). The GMD imposes less whimsical assumptions on the data. The properties of the Gini regressions are presented in a book by Yitzhaki and Schechtman (2012).

The aim of this note is to present a non-technical discussion of the major characteristics of the Gini regression, and the way it offers an adjustment of the econometric technique to meet Leamer's critique.

The structure of the note is the following: Section 2 interprets the term “whimsical assumptions” and presents a list of those assumptions in a typical OLS regression. Section 3 comments on other regression methods, while section 4 concludes.

2. The definition of a "whimsical" assumption

An econometric model can be viewed as imposing a structure on the data. In this process one has to decide which properties of the data to stress and which to ignore. Ignoring a possibility can be justified if it does not affect the conclusions in a drastic way. I define "drastic" to be the change of the sign of a regression coefficient. The reason for this definition is that a change in a sign reverses the conclusion reached.

I define a "whimsical assumption" as an assumption, imposed on the data, and has the potential of changing the sign of the regression coefficient. A whimsical assumption can either change the sign of a regression coefficient alone, or in a combination with other whimsical assumptions.

I restrict the discussion to whimsical assumptions that are used in an OLS regression and can be avoided by the use of the Gini regression. In what follows, I first describe those assumptions and then discuss their implications. Among those assumptions are:

(a). The reliance on a symmetric correlation measure.

The regression is an asymmetric relationship between variables. To attest that, note the distinction between the dependent and independent variables. Imposing a symmetric correlation on the data, may in some cases affect the sign of the correlation.

(b). The reliance on a symmetric variability measure.

The variance is symmetric around the mean. Moreover, it is sensitive to extreme observations. On the other hand, one of the basic assumptions in economics is the declining marginal utility of income. To accommodate economic theory, one needs a measure of variability that does not contradict economic theory.

(c). The assumption of the linearity of the model.

Most econometric models are based on a linear model in the parameters. This assumption need not be respected by the data, leading in some cases to a change in the sign of a regression coefficient.

(d). The (almost) free use of transformations

A transformation applied to a variable in a regression is an accepted procedure that does not need a justification. Transformations change the distribution of the random variable, and in some cases, may change the sign of a regression coefficient.

It should be stressed that in some cases the change in sign can occur because of a combination of two or more whimsical assumptions. Therefore, in what follows I discuss the effect of combinations of assumptions.

2(a): The reliance on a symmetric correlation coefficient.

The OLS regression is based on the properties of the Pearson correlation coefficient. This coefficient is based on the properties of the covariance:

$$\rho(X,Y) = \rho(Y,X) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Conventional wisdom says that the correlation is bounded between minus one and one. But by applying monotonic transformations on the variables one can change the bounds on the correlation.

For example, if (X, Y) have a bivariate standard lognormal distribution (that is, their natural logs have a standard bivariate normal distribution), then the range of Pearson's ρ is [-0.368, 1] (De Veaux, 1976). Denuit and Dhaene (2003) present an example with Pearson's correlation coefficient converging to zero, while the variables are connected by a monotonic transformation. Denuit and Dhaene (2003) conclude that: "it is possible to have a random couple where the correlation is almost zero even though the components exhibit the strongest kind of dependence possible for this pair of marginals" (p-3).

The effect of these properties on econometric analysis should not be underestimated. Note that econometricians tend to almost freely apply monotonic transformations to the variables. By doing so they may affect the correlation between the variables, which in turn may affect the relative explanatory power of an independent variable, or alternatively, may change the sign of the correlation. The conclusion is that one should be careful when applying seemingly innocent monotonic transformations to variables when using Pearson's correlation coefficient.

The GMD has two correlation coefficients defined between two variables and the range of each one of them is between [-1,1]. The definitions are:

$$\Gamma_{XY} = \frac{\mathrm{cov}(X, F(Y))}{\mathrm{cov}(X, F(X))} \text{ and } \Gamma_{YX} = \frac{\mathrm{cov}(Y, F(X))}{\mathrm{cov}(Y, F(Y))} .$$

In general, the two correlation coefficients may have different signs. The two correlation coefficients are equal if the distributions are symmetric. Only in this case, the decomposition of the variability of a linear combination of the random variables is identical in structure to the decomposition of the variance. On top of that the Gini methodology enables the researcher to check whether the correlation between two variables is monotonic, i. e., whether the sign of the correlation does not change along the distribution of each variable.

The existence of two correlation coefficients between two variables offers a built in specification test. To see this, note that an optimization procedure results in an orthogonality condition. Otherwise, the procedure does not offer an "optimal" solution, because one can improve upon the target of the optimization. In a regression, together with the requirement that the regression line passes through the mean, this implies that a covariance between each independent variable and the residuals is set to zero. Because there are two correlation coefficients, one is imposed on the data by the regression method while the other can be used to test whether the other covariance is also equal to zero. If it is, it is safe to use extrapolations along this variable. Otherwise, the regression should not be used for extrapolation.

2(b): A symmetric measure of variability

A fundamental assumption in economic theory is the law of positive and declining marginal utility. To handle issues that involve decisions under risk or social welfare one has to have a modification of the concept of mean preserving spread.

The mean preserving spread is composed of at least two simultaneous changes in the distribution: one involves an increase and the other involves a decrease in the relevant variate (i.e., income of an agent in welfare economics or return in a state of nature in finance). Let us decompose the mean preserving spread into an increase and decrease in different outcomes. Also, let us concentrate on the mean and variability measure used. An increase in the income of one agent (or in a state of nature in finance) increases the mean but it can either increase or decrease the variability measure. To be compatible with expected utility theory, it must be that the increase (decrease) in the mean is greater than the increase (decrease) in the measure of variability. This property holds in the GMD but may be violated in the case of the variance. This property also holds in a class of asymmetric measures based on the GMD, which is the extended Gini family. Yitzhaki and Lambert (2012b) use this weakness of the variance in order to construct an example demonstrating that a risk-averse expected utility maximizer may prefer a higher variance in the rate of return on his portfolio.

2 (c): The linearity assumption.

The basic assumption in the OLS regression is that the regression curve is linear in the parameters. Economic theory does not require this assumption, and on top of that this assumption does not necessarily hold in the data. The estimated regression coefficient can be described as a weighted average of slopes, defined between adjacent observations of the independent variable. The weights are determined by the regression method used (Yitzhaki, 1996). This means that the linear regression is actually a linear approximation to the (unknown) regression curve. The use of the Gini regression enables to check:

(a) Whether the regression curve is monotonic. That is, whether the sign of the regression curve is the same along the sections of the independent variable.

(b) Whether a monotonic non-decreasing transformation can change the sign of the regression coefficient.

Note that (a) is weaker than (b). That is, a necessary condition for (b) to hold is that (a) holds.

To see the significance of this assumption, note that the regression coefficient is a weighted average of slopes defined between adjacent observations, and that economic theory does not require imposing linearity on the regression curve. Hence, in welfare economics, optimal taxation, and portfolio analysis, the use of the GMD (or the extended Gini) enables imposing economic theory (i.e., the declining marginal utility of income) on the econometric model. Otherwise, as is pointed in Yitzhaki (1996), the economist and the econometrician analyzing the data, who may be the same person wearing a different hat, may contradict each other.

2 (d). The (almost) free use of transformations.

Econometricians tend to apply transformations to variables almost freely. Transformations change the distribution of the variable concerned, and they also change the Pearson correlation coefficients with other variables. On top of that, transformations may affect properties of the data. For example, in demand analysis, the data may reflect the property that the expenditure of each agent is equal to his income. This implies that the sum of the marginal propensities to spend equals one. Once the log transformation is applied to one of the variables, the estimated model does not have this property. If one wants to preserve this property, one has to impose it on the model. In a Gini regression there are restrictions on imposing transformations on the variables. A linear model without imposing transformations on the variables has the advantage that all properties that are held in the data also hold in the estimated model. It is especially important not to allow truncation of variables that are used in the regression, like turning a continuous variable to a binary one, such as the frequent use of participating/non participating distinction. In an Instrumental Variable regression truncation may affect the sign of the regression coefficient.

3. Comments on other regression methods:

There are several regression methods that compete with the OLS, among them, Mean Absolute Deviation (MAD) regression, Least Absolute Deviation (LAD) regression, Quantile regression (the absolute deviation from a quantile of the residuals) and Maximum Likelihood (ML) regression.

The first three regressions, i.e., MAD, LAD and Quantile regressions are based on variability measures that belong, like the GMD, to the L_1 (city bloc) metric and as such, we should expect them to have properties that are identical to those of the GMD. However, it is shown by Yitzhaki and Lambert (2012a) that they can be viewed as between-group GMD component in the decomposition of the GMD in a way that resembles ANOVA, and is referred to as ANOGI. Hence, using them to achieve robustness is similar to using the between-group variance instead of the variance of the residuals as a target of the OLS regression. It means that all intra-group variability is dismissed as irrelevant. I believe that this is a heavy price to pay to achieve robustness.

Another regression method is the maximum likelihood regression. Maximum Likelihood aims at the mode of the distribution. The mode is a reasonable target whenever there is no weight, in terms of the target function, attached to different outcomes, because it implies to be right in most cases. In economics, and especially in the areas of finance and income distribution, to be right most of the time and to avoid attaching high values to extremely important cases does not seem a reasonable strategy.

4. Concluding Comments:

The argument of this note is that there are too many tools in the arsenal of the researcher to influence the results of the regression. In some sense the target of the overzealous researcher is to prove his point, which in some cases translates into finding the model that can get the results the researcher believes in. My aim is to get better revelations of the hidden and redundant assumptions that may be responsible for the results. For this purpose, we need a methodology that "reveals more" (the term was coined by Lambert and Decoster, 2005).

I believe that Leamer's critique can be partially answered by developing a better technique that incorporates the structure of the OLS as a special case. Moreover, it can be used together with the OLS, to see how robust are the conclusions derived by the regression.

The suggested methodology is based on the properties of the GMD (and the extended Gini family) that by now have many known properties that are similar to the variance, but reveal more.

The "reveal more" includes the following:

The basic assumption in a regression is that there exists a linear model connecting the variables. The OLS and Gini methods result in regression coefficients which are weighted average of slopes between adjacent points of the independent variable. The method of regression determines the weighting scheme. Hence, linearity is a “whimsical” assumption. If the underlying model is not linear then changing the method of regression may result in an estimate with a different sign.

To see the effect of this assumption consider the following: Some variables are not related to each other in a monotonic way. As an example, let us consider age. The association of many variables with age is a U-shape (or an inverse U-shape) relationship. In such a case, the composition of the sample, and the way the age variable is introduced in the regression, together with monotonic transformations and the regression method may determine the sign of the regression coefficient with respect to age. Also, these factors may determine the sign of the regression coefficients that are included as independent variables together with age. The use of the Gini methodology enables us to identify those variables. If such a variable is involved in the regression, and affects the sign of the regression coefficient of a key variable, then one has to inform the reader and to use a more complicated modeling structure.

The same problem exists if the relationship is monotonic but not linear. The reason: the regression method or a monotonic transformation can change the magnitude of the regression coefficient of the variable and the magnitude of its correlation with other variables. This in turn can change the sign of a regression coefficient of another variable in the regression.

The Gini method enables the researcher to test for linearity. If linearity is rejected, then the model should be viewed as a linear approximation, not useful for prediction. Also, at least in two areas in social science, economic theory calls for asymmetric treatment of the data. This is caused by the assumption of declining marginal utility of income that is assumed in the areas of decision making under risk and income distribution. To avoid contradiction between statistical and economic theories we have to impose economic theory on the statistical methodology. The extended Gini

allows the researcher to impose (and reveal to the reader) her social (or risk) attitude, and to impose the statistical measure of variability on the analysis. This way one reveals her social attitude first and then imposes it on the analysis. Alternatively, one is required to perform sensitivity analysis using alternative social preference.
Monotonic transformations, which are a legitimate tool to use in modeling the relationships between variables, can change the sign of the relationship in the case of a non-monotonic relationship. Monotonic transformations include using a different functional form, restricting the sample from above and below, and binning (making a continuous variable a discrete one). There should be a ban on using transformations of variables because they change the properties of the data. Using the Extended Gini regression keeps the properties of the data intact, while applying a transformation on the weighting scheme.

The implication of using the Gini methodology will be to increase the number of sensitivity tests one has to perform which will result in reducing of the quantity of results "proven" by regressions. Researchers will have to admit that sometimes we do not have the answer. However, it is clear that using the GMD does not exhaust all the possibilities. I have restricted the discussion to those assumptions that can be revealed by using the Gini.

In some sense, to be able to respond to Leamer's critique requires a way of "licensing" the way research in social science should be done. Similar to the requirements of the Food and Drug Administration we should impose (looser) regulations on how research in the social science should be performed and what properly conducted research should report.

Acknowledgement: I am grateful to Peter Lambert and Ingram Olkin for helpful comments on a draft of this note.

References:

Angrist, J. and J.S. Pischke (2010). The credibility revolution in empirical economics: how better research design is taking the con out of econometrics, Working Paper No. 15794, NBER. http://www.nber.org./papers/w15794.

De Veaux D. R., (1976). Tight upper and lower bounds for correlation of bivariate distributions arising in air pollution modeling, Technical Report No. 5 (1976), Department of Statistics, Stanford University.

Denuit, M and J. Dhaene (2003). Simple characterizations of comonotonicity and countermonotonicity by extremal correlations. *Belgian Actuarial Bulletin* 3, 22-27.

Golan, Y. and S. Yitzhaki (2010). Who does not respond in the social survey: an Exercise in OLS and Gini regressions. Draft. Presented at the 31st IARIW, http://www.iariw.org/c2010.php.

Lambert, P. J. and A. Decoster (2005). The Gini Coefficient Reveals More, *Metron*, LXIII, 3, 373-400.

Leamer, E. (1983). Let's Take the Con Out of Econometrics, *American Economic Review* 73, 1, 31-43.

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

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

Yitzhaki, S. and P. Lambert (2012a). Is Higher Variance necessarily bad for Investment. Working paper. http://ssrn.com.

Yitzhaki, S. and P. Lambert (2012b). The Relationship between Gini's Mean Difference and the Absolute Deviation from a Quantile. Working paper. http://ssrn.com.

Yitzhaki, S. and E. Schechtman (2012). *The Gini Methodology: A primer on a statistical methodology*. Springer:N.Y.