This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

Sensitivity Analysis through mixed Gini and OLS regressions
Paper submitted to the NSF

By

Shlomo Yitzhaki

The Government Statistician, Israel and
Professor emeritus, The Hebrew University
Shlomo.yitzhaki@huji.ac.il

Abstract

About thirty years ago Edward Leamer criticized the credibility of empirical research in economics. Since then there were huge improvements in research design, data collection and econometric methodology. On the other hand, the huge increase in computing power has increased the number of instruments available for the use of the over-zealous researcher who wants to prove his point. I suggest developing the mixed Gini and Ordinary Least Squares regression. It enables unraveling, tracing and testing the role of several whimsical assumptions imposed on the data in regression analysis. Among those assumptions are the linearity assumption, the use of monotonic increasing transformations, and the symmetry between distributions that is imposed by the Pearson correlation coefficient. My conjecture is that the new technique will reduce drastically the number of results that are claimed to be supported by empirical "proofs".

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

Sensitivity Analysis through mixed Gini and OLS regressions

1. What is the fundamental question?

A popular method in reaching quantitative conclusions in research is the method of regression. The most popular one is the Ordinary Least Squares, which is based on the properties of the variance. To simplify the presentation, I will restrict my arguments to the OLS, although the arguments apply, with some modifications, to other methods. It is clear that this area of research suffers from a lack of credibility. This was pointed out by Edward Leamer who states that "Hardly anyone takes data analysis seriously."[1] Leamer traced the lack of credibility to lack of robustness because of sensitivity to key assumptions he called as "whimsical". My interpretation of whimsical assumptions is those that are imposed on the data, affect the coefficients in a drastic way but are not supported by the data.

In a recent paper Angrist and Pischke responded by pointing out the huge improvements in research design, better data collection, better definitions of the research question, and more.[2] 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 are not supported by the data, and that may drive the results, 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 partial effect of one variable on another?

Golan and Yitzhaki supply a positive answer to this question. They show that if one investigator uses Gini regression and the other OLS regression, then the sign of some regression coefficients differ.[3] Yet both regression methods rely on plausible properties of regression model and can be described as innocent applications of the methodology. .

[1] E. Leamer, Let's Take the Con Out of Econometrics, *American Economic Review* 73 no. 1 (1983): 37,

[2] J. Angrist and J.S. Pischke, 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

[3] Y. Golan and S. Yitzhaki, 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

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

My argument 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 searching for the model that can get the desired results. My aim is to provide a method that exposes the hidden and redundant assumptions that are responsible for the results. For this purpose, we need a methodology that "reveals more" (the term was coined by Lambert and Aronson).

I believe that Leamer's critique can be answered by developing a better technique that incorporates the properties of 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 Gini's Mean Difference (and the Extended Gini family), which has many properties that are similar to those of the variance (and which nests variance), but reveals more about the critical underlying statistical assumptions.

The "reveal more" includes the following:

The basic assumption in a regression is that there exists a linear model connecting the variables. Yitzhaki showed that both the OLS and Gini methods result in a regression coefficient which is a weighted average of slopes between adjacent points of the independent variable.[4] The method of regression determines the weighting scheme. Hence, linearity is a “whimsical” assumption. If the underlying model is not linear then the method of regression would result in different estimate.

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 increasing transformations and the regression method, may determine the sign of the regression coefficient with respect to age. These factors may also determine the sign of the regression coefficients that are included as independent variables together with age.

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

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

The Gini methodology enables the researcher and the reader to know whether non-monotonic relationships among the variables exist.[5]

The same problem exists if the relationship is monotonic but not linear. The reason is that the regression method or a monotonic transformation can change the magnitude of the regression coefficient 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 also enables the researcher to test for linearity. If linearity is rejected, then the model should be viewed as a linear approximation that is not useful for prediction.

In at least two areas of social science, economic theory calls for asymmetric treatment of the data. It arises because of the assumption of declining marginal utility of income that is made in the areas of decision making under risk and income distribution. It is shown in Yitzhaki that statistical theory may contradict economic theory if some "whimsical" assumptions--like linearity of the model that includes income as an independent variable is imposed but not supported by the data.[6] To illustrate, a researcher that uses a linear expenditure system using OLS in order to design poverty reducing subsidies estimates the income elasticities of the commodities by the properties of the top deciles, totally ignoring the poor. 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 that reflects the social attitude on the analysis. This way one first reveals one's social attitude and then imposes it on the analysis.

Monotonic transformations, which are a legitimate tool to use in modeling the relationship 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 or/and below, and binning (making a continuous variable a discrete one). Of course such transformations of variables change the properties of the data, and therefore should be used sparingly and, when used, should be carefully documented and justified. The Extended Gini regression

[5]E.S. Schechtman, S. Yitzhaki and T. Pudalov, Gini's Multiple Regressions: Two Approaches and Their Interaction (2010). http://ssrn.com

[6] Yitzhaki (1996) Ibid.

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

keeps the properties of the data intact, while applying a transformation on the weighting scheme.

Another instrument in the hand of an over-zealous or sophisticated researcher is the Pearson correlation coefficient. Its "official" range is between minus one and one. But, if the underlying distributions of the variables are different then its range can be limited. As an example, consider two lognormally distributed variables, for which the Pearson correlation coefficient is bounded from below by -0.36.[7] That is, by applying the transformation e^x to two normally distributed variables, we are able to change the correlation coefficient from minus one to minus -0.36. The problem is especially relevant and severe in the field of finance, where additive and multiplicative relationships are mixed. (A transfer of a dollar from one asset to the other is additive, while interest over time is multiplicative, because compound interest is used). Levy and Swartz have shown that the Pearson correlation between the returns of two assets will converge to zero provided that one estimates them for a long enough period, no matter what is the periodical correlation coefficient.[8]

The implications of advancing the domain:

To answer these questions, we need to explain the difference between the OLS/variance and Gini-based methodologies. Both are based on averaging the difference between all pairs of observations. The difference is in the metric used to define the distance between observations: the variance is based on Euclidean metric, the Gini on "city block". (That is, one can move east/west or north/south as in Manhattan). It is not obvious a priori which metric is more appropriate for particular social science applications.

As Yitzhaki develops, the Gini is the only measure of variability that can be decomposed in a way that resembles the decomposition of the variance, but it reveals more.[9] In this context "reveals more" means that the structure of the decomposition of the Gini of a linear combination of random variables can be identical to the structure

[7] E. Schechtman and S. Yitzhaki, On the Proper Bounds of the Gini Correlation, *Economics Letters* 63, 2 (May 1999), 133-138.

[8] H. Levy and G. Schwarz, Correlation and the Time Interval over which the Variables Are Measured, *Journal of Econometrics* 76 (1997): 341

[9] S. Yitzhaki, Gini's Mean Difference: A Superior Measure of Variability for Non-Normal Distributions, *Metron* LXI, 2 (2003): 285-316.

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

of the decomposition of the variance, provided that certain properties of the underlying distributions hold. Therefore, by using the decomposition of the Gini we can learn about the implicit assumptions behind the use of the variance method. Among the implicit assumptions is the imposition of a symmetric relationship in correlations and linearity. Under the Gini methodology there are two correlation coefficients between two variables, and if they are equal then one gets the same decomposition as is done under the variance. If they are not equal, then a symmetric relationship is imposed on the asymmetric relationship, leading to some of the results mentioned above.

Thus using the Gini methodology expands the number of sensitivity tests one has to perform to demonstrate a robust relationship, which will result in reducing of the quantity of results “proven” by regressions. The range of cases for which researchers will have to admit that no answer can be confidently claimed will likewise expand.

To see how this can be done, note that the Gini methodology allows for mixed regressions in the following sense: one can run an OLS regression as well as a Gini regression. If the sign of all regression coefficients are identical and the values do not differ very much, then we may conclude that the regression results are robust. On the other hand, if the signs are not equal, then the researcher should run a mixed regression where the investigator can choose which independent variable to treat according to OLS and which according to the Gini. This enables the researcher to move from one regression to the other in a step-wise way so that one can identify the variable(s) that are responsible for the change in sign that can happen in another independent variable. Should one detect such a change, then one can investigate the simple regression coefficients for a non-monotonic or non-linear relationship.[10]

Adopting the methodology will reduce the number of new "findings" in the social science, which in turn will increase the trust in empirical results, further supporting the “credibility revolution” championed by Angrist and Pischke.[11]

At this stage, one can replicate every text-book in econometrics. However, the Gini regression is much more complicated than the OLS, friendly Software has to be developed and it is not clear that there are low hanging fruits in every area.

[10] Golan and Yitzhaki (2010), Ibid.

[11] Angrist and Pischke (2010), Ibid.

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

For Further Reading:

Schechtman, E; S. Yitzhaki and T. Pudalov (2010). Gini's multiple regressions: two approaches and their interaction. http://ssrn.com. Forthcoming, *Metron*.

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.