Working Paper

I Didn't Run a Single Regression

Author(s):
Müller, Christian

Publication Date:
2006-02

Permanent Link:

https://doi.org/10.3929/ethz-a-005118441 →

KOF Konjunkturforschungsstelle
Swiss Institute for
Business Cycle Research

Arbeitspapiere/ Working Papers

Christian Müller

I didn't run a single regression

No. 128, February 2006

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

I didn't run a single regression

Christian Müller

Swiss Institute for Business Cycle Research (KOF) at the
Swiss Federal Institute of Technology Zürich (ETHZ)
CH-8092 Zürich, Switzerland
Tel.: +41.44.632 46 24
Fax: +41.44.632 12 18
Email: cmueller@kof.ethz.ch

First version: January 27, 2006

This version: January 30, 2006

Abstract

Growth regression economics are haunted by the fact that results are easily overthrown by regressing alternative model specifications. Recent research therefore aims at obtaining robust regression results by systematically running multiple models and picking surviving variables. This note shows that a very popular of these approaches, the robust regression due to Sala-i-Martin (1997) very likely leads to inconsistent conclusions but may be remedied by refining the 'testimation' algorithm. To that aim I do not need to run a single regression.

JEL classification: C50

Keywords: robust estimation, growth regression

The question what are the determinants of economic growth and hence welfare has always been one of the key issues in economics. Therefore, many theoretical and econometric studies have tried to shed light on this subject. Especially in the aftermath of Barro's (1991) survey on potential growth factors, the growth regression literature itself has experienced enormous growth. However, particularly to non-experts the evidence seems to be very confusing to the effect that hardly anything appears to be a robust finding as to what really causes economies to grow faster than others.

Consequently, several attempts have been made to generate 'robust' empirical results. For example, Levine and Renelt (1992) applied Leamer's (1985) extreme bound analysis concluding that any variable that changes sign or becomes insignificant in any single regression model variant should be labelled non-robust. Granger and Uhlig (1990) modify (and simplify) Leamer's (1985) approach by letting the researcher choose how 'extreme' the selection should actually be. Against that Sala-i-Martin (1997) (henceforth SIM) proposed an alternative that is based on a systematic re-sampling of the potential regression models. He derives a test statistic that measures the dispersion of the coefficient of interest across models. If the probability mass of the corresponding empirical cumulative distribution function is far away from zero, then the corresponding variable appears robust. Both these approaches are now well established in the literature, a recent application is due to Sturm and de Hahn (2005), for example.

This note fills a gap in SIM's argument that arises due to the omission of an explicit statement of the null and alternative hypotheses. It is pointed out that particular assumptions are needed to derive the

robustness statistic. Unfortunately, it turns out that the SIM proposal easily runs into inconsistency which severely limits its use in applied research.

The structure of the paper is as follows. The next section describes the SIM approach in more detail. Then, the arguments are reviewed, the tacit assumptions highlighted, and the limits of the method are discussed. Finally, a potential remedy is presented and conclusions are drawn. As this is rather a technical note, not a single regression is run.

I Sala-i-Martin revisited

The starting point is a regression model of the following type

$$(1) \quad \begin{aligned} y &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \\ u &\sim i.i.d.(0, \sigma_u^2) \end{aligned}$$

with y being the independent variable, usually the economies' growth of income, and $x_i, i = 1, 2, 3$ are the (potential) explanatory variables for growth while u represents orthogonal white noise.

Let me call $x_1 \in X_1$ the robust variables, x_2 the variable(s) in question and $x_3 \in X_3$ noise variables. The researcher wants to know whether $x_2 \in X_1$ or $x_2 \in X_3$, in other words, is x_2 signal or is it noise?

As SIM observes, standard t-tests appear not fully suitable to answer this question since their results are not robust with respect to the choice of x_3. Therefore, he suggests to build a statistic on a sample of estimated β_2 coefficients (and their standard deviations) where the values are drawn from the set of models that is given by the M possible combinations of x_3. In SIM's example the set X_3 comprises $K_3 = 58$

variables while the vector x_3 is made up of $k_3 = 3$ variables. Therefore, a total of $M = \frac{K_3!}{(K_3-k_3)!k_3!} = 30,856$ models can be estimated for every choice of x_2.

Calling $\hat{\beta}_{2,j}$ and $\hat{\sigma}_{\beta_2,j}$ the jth draw of the estimated coefficient and its standard deviation respectively, SIM calculates

$$(2) \qquad \tilde{\beta}_2 = \sum_{j=1}^{M} \omega_j \hat{\beta}_{2,j}$$

$$(3) \qquad \tilde{\sigma}^2_{\beta_2} = \sum_{j=1}^{M} \omega_j \hat{\sigma}^2_{\beta_2,j}$$

with $\omega_j, \sum_{j=1}^{M} \omega_j = 1$ as weights reflecting the model fit in terms of the relative likelihood value. Using $\tilde{\beta}_2$ and $\tilde{\sigma}^2_{\beta_2}$ as mean and standard deviation respectively, SIM constructs the cumulated normal density function (CDF) of the across-equation mean of β_2. It is then easy to see whether or not the probability mass is far enough away from zero to call x_2 robust.

II Tacit assumptions and their implications

In this section a possible set of assumption is discussed that would be able to justify SIM's statistic. Notice that there may be alternative sets, however, as SIM does not provide a set himself, the current discussion appears warrantable. The key argument of the SIM method is the derived CDF. Therefore, its quality is crucial for the conclusions that may be drawn from it. The CDF needs two parameters to be identified. Thus instead of looking at the CDF it is sufficient to scrutinise $\tilde{\beta}_2$ and $\tilde{\sigma}^2_{\beta_2}$. For both of them to be useful, it first of all need to be assumed that their estimates given in (2) and (3) converge to their true values,

i.e. they need to be reliable estimates of their population means.

The literature knows several measures of convergence (see *e.g.* Hamilton 1994, chapter 7). Since SIM does not define the measure he has got in mind, I suppose a very weak one. Thus, we first have to assume that all $\hat{\beta}_{2,j}$ are drawn from the same population. This population shall be the population of consistent estimates for β_2. We may hence write

$$\hat{\beta}_{2,j} \sim \left(\beta_2, \sigma^2_{\beta_{2,j}}\right) \tag{4}$$

which indicates that all population members have the same expected value. In order to simplify matters we may confine our analysis to β_2. The arguments regarding $\sigma^2_{\beta_2}$ would be virtually identical and will therefore not be discussed. Likewise, the choice of ω_j is not going to be discussed since the only interesting one would be zeros for certain j (see the discussion below). As this is extremely unlikely for likelihood ratios, it is not worth considering.

The question thus is under what circumstances will (4) hold? Remember that $\hat{\beta}_{2,j}$ is the estimation result of a regression defined in (1). Therefore, every single regression must be such that (4) applies. It is again possible to use standard results. One central (and mild) assumption of standard regression techniques is

$$\lim_{R\to\infty} \frac{1}{N} \sum x_2^i u^i = 0, \tag{5}$$

where $i = 1, 2, \ldots, N$ is the number of observations (*i.e.* the number of countries in the sample). A stronger assumption would for example require independence of x_2 and residuals. Less technically speaking, (5) demands that the explanatory variable of interest is asymptotically uncorrelated with the innovations. This condition is important to ensure consistency of the estimate in every jth draw. Disregarding (5)

would almost surely introduce a bias in $\hat{\beta}_{2,j}$ and since SIM shows no indication of supposing malpractice on parts of the growth researchers it appears save to put (5) on the list of tacit assumptions. The problem now becomes how to make sure that (5) holds for all j. Remember that the choice of x_3 differs for different values of j, that is for different model variants. Therefore, if a set $x_{3,j}$ is related to y by a coefficient $\beta_{3,j} \neq 0$ then consistency of $\hat{\beta}_{2,j+i}, i \neq 0$ is only available if

$$(6) \quad Corr\left(x_{3,j}, x_{2,j+i}\right) = 0.$$

In other words, the variables from the noise set must not be correlated with the variable of interest, that is with x_2. This, however, seems to be a very unlikely situation. Putting it the other way round, (6) imposes a sharp selection criterion for the SIM approach to work since many theoretically attractive variables will fail to pass (6). All those variables that have to be deselected can of course not be subjected to the SIM test and hence the SIM robustness check appears severely limited. It may be interesting to note that SIM found ten dummy variables such as Sub-Saharan Africa, number of revolutions and military coups and religious orientation (Confucian, Buddhist, *etc.*) to be robust out of 22 robust variables in total. This rather high share may thus simply reflect the fact that these dummy variables are very likely to be independent of the remaining potential explanatory variables. Therefore, they probably comply with (6) making reliable inference feasible. In contrast, other robust variable may not have been detected because the related coefficient estimates are inconsistently estimated. Accordingly, the 'robust variables' may not be robust at all, it just cannot be told.

A seemingly simple way to circumvent (6) is to assume

(7) $\beta_{3,j} = 0 \; \forall j.$

Doing so is equivalent to avoiding an omitted variable bias at a later stage: Suppose that the first variable of interest that is checked for robustness, say $x_2(1)$, turns out non-robust according to SIM. Then another variable $x_2(2)$ would be chosen out of $X_3(1)$ and the first candidate be moved to $X_3(2)$. If it happens that this new $x_2(2)$ appears robust it would automatically imply that the previous inference was wrong since (7) was violated. Thus, every robust variable that is found reduces the reliability of the testing procedure.

Would it pay to to shift $x_{2,j+1}$ to X_1 instead and restart the whole analysis? Probably not very much so, since any further detection of a robust variable would invalidate the (previous) decision to shift. In fact, the only feasible such robustness check is able to check exactly one variable x_2. Otherwise, only non-robust variables can be identified, provided of course that the null hypothesis is not rejected for any x_2 possible.

In short, finding a robust variable appears not very desirable although it is the overriding objective of the whole approach. However, how bad is the effect of a robust variable actually? To answer this question I define s as the number of robust variables in X_3, and by $M^* \leq M$ I denote the number of consistent estimations. We may now calculate the share of admissible, that is consistent, regressions of the total number of regressions. For example, for $k_3 = 3$ as in SIM, and a hypothetical $s = 1$ we have to calculate the number of pairs of variables out of the set X_3 that can be complemented with the robust variable in order to obtain consistent estimators. Only those triplets where

the robust variable is included generate consistent estimates because there will be no omitted variable bias. Therefore, M^* can be given as $M^* = \left(\begin{array}{c} K_3 - 1 \\ k_3 - 1 \end{array} \right) = \left(\begin{array}{c} K_3 - 1 \\ 2 \end{array} \right)$ using Euler's binomial coefficient notation. After some algebra it turns out that M^* is proportional to K_3^2 whereas M increases proportional to K_3^3. Thus, M^*/M is proportional to K_3^{-1} which implies that the share of consistent $\beta_{2,j}$ estimates entering (2) approaches zero as K_3 increases. In general,

$$M^* = \begin{cases} \left(\begin{array}{c} K_3 - s \\ k_3 - s \end{array} \right), \ \forall s \leq k_3 \\ 0, \text{else}, \end{cases}$$

and hence M^*/M is proportional to K_3^{-s}. There is no nonzero s for which (2) provides a consistent estimate unless (6) holds. In particular, if $s > k_3$ and x_2 is non robust then not a single regression will yield consistent estimates. For $s = k_3$ there will be exactly one valid regression, no matter how large K_3 is. Relating this result to SIM, one might note that if the 12 robust (non-dummy) variables SIM claims to have found were really non robust, then just less than half of the nearly 2 million regressions delivered consistent estimates for β_2. As remarked before, these consistent estimates may have been obtained for the dummy variables.

Summarising this section gives the following picture. We may either drop (7) and find us put back where we started from, namely to the position where we have to choose x_1, x_2, and x_3 and play around with various such choices. The chances of learning about the true determinants of growth would remain as thin as before. Or, we could assume that (7) holds and trust that research on growth is fruitless, *i.e.* newly suggested explanatory variables are non-robust. As neither of these two perspectives is very attractive, the next section suggests a slight

manipulation of the SIM method that solves some of the problems.

III An alternative route

The previous section demonstrated that it is possible to consistently accept non-robustness of all variables in question. Suppose now, that instead of identifying robust variables, we only find non-robust variables by the method of SIM. Then, if there are any robust variables available, they must be in X_1. Thus a promising alternative is to start with a rather large set X_1 and reduce it as far as possible by throwing out those elements that in fact belong to X_3. The following algorithm can be applied. Start with some choice of X_1, x_2, and x_3 where now, X_1 should contain many variables, in particular those which are in focus. Apply the SIM method. If x_2 turns out non-robust, move it to X_3 and chose another x_2 from the (now smaller) set X_1. If however, x_2 turns out robust, put it (back) to X_1 and select another x_2 from X_1. Repeat these steps until all variables in X_1 have been tested and only robust variables remain in X_1. Notice that this procedure is consistent because (7) can always be maintained. A serious caveat however, arises if one considers the possibility of wrongly accepting the null. If that happens and it will almost surely happen when X_1 has many elements, then (7) is also almost surely violated. A way out here would be not to shift the non-robust x_2 variables to X_3 but remove them from the exercise altogether because (7) could still be maintained. Unfortunately, the only choice that cannot be checked is the definition of the initial set X_3. That regrettably mimics again the current situation where it is exactly the initial choice of explanatory variables in growth regressions which

leads to these contradictory results. Nevertheless, due to its consistency the alternative approach seems to be preferable to the suggestion by SIM.

IV Summary and conclusions

Sala-i-Martin (1997) has suggested an intuitively appealing way to check the robustness of explanatory variables in economic growth regressions. Complementing the intuition with explicit assumptions about the properties of the proposed statistic reveals a severe drawback, however. In fact, even under rather mild assumptions the possibilities to really find robustness appear very limited. The proposed method will in general not be applicable to the whole set of potentially robust explanatory variables and the result of the robustness check cannot be regarded robust itself.

An alternative algorithm has been suggested. Although it will not solve the entire problem, it provides an internally consistent way to address robustness.

References

Barro, Robert J., "Economic Growth in a Cross Section of Countries," *The Quarterly Journal of Economics*, 1991, *106* (2), 407 – 43.

Granger, Clive W. J. and Harald Uhlig, "Reasonable extreme–bounds analysis," *Journal of Econometrics*, 1990, *44*, 159 – 70.

Hamilton, James D., *Time Series Analysis*, 1st ed., Princeton, New Jersey USA: Princeton University Press, 1994.

Leamer, Edward E., "Sensitivity Analyses Would Help," *American Economic Review*, 1985, *57* (3), 308 – 13.

Levine, Ross and David Renelt, "A Sensitivity Analysis of Cross-Country Growth Regressions," *American Economic Review*, 1992, *82* (4), 942 – 63.

Sala-i-Martin, Xavier X., "I Just Run Two Million Regressions," *American Economic Review*, 1997, *87* (2), 178 – 83.

Sturm, Jan-Egbert and Jakob de Hahn, "Determinants of long-term growth: New results applying robust estimation and extreme bound analysis," *Empirical Economics*, 2005, *30* (3), 597 – 617.