I Just Ran Two Million Regressions

By XAVIER X. SALA-I-MARTIN*

Following the seminal work of Robert Barro (1991), the recent empirical literature on economic growth has identified a substantial number of variables that are partially correlated with the rate of economic growth. The basic methodology consists of running cross-sectional regressions of the form

$$(1) \qquad \gamma = \alpha + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \cdots + \beta_n \mathbf{x}_n + \varepsilon$$

where γ is the vector of rates of economic growth, and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are vectors of explanatory variables, which vary across researchers and across papers. Each paper typically reports a (possibly nonrandom) sample of the regressions actually run by the researcher. Variables like the initial level of income, the investment rate, various measures of education, some policy indicators, and many other variables have been found to be significantly correlated with growth in regressions like (1). I have collected around 60 variables which have been found to be significant in at least one regression.

The problem faced by empirical growth economists is that growth theories are not explicit enough about what variables $\mathbf{x}_j$ belong in the "true" regression. That is, even if it is known that the "true" model looks like (1), one does not know exactly what particular variables $\mathbf{x}_j$ should be used. If one starts running regressions combining the various variables, variable $\mathbf{x}_1$ will soon be found to be significant when the regression includes variables $\mathbf{x}_2$ and $\mathbf{x}_3$, but it becomes nonsignificant when $\mathbf{x}_4$ is included. Since the "true" variables that should be included are not known, one is left with the question: what are the variables that are really correlated with growth?

An initial answer to this question was given by Ross Levine and David Renelt (1992).[1] They applied Edward Leamer's (1985) *extreme-bounds test* to identify "robust" empirical relations in the economic growth literature. In short, the extreme-bounds test works as follows. Imagine that there is a pool of N variables that previously have been identified to be related to growth and one is interested in knowing whether variable z is "robust." One would estimate regressions of the form

$$(2) \qquad \gamma = \alpha_j + \beta_{yj}\mathbf{y} + \beta_{zj} z + \beta_{xj}\mathbf{x}_j + \varepsilon$$

where $\mathbf{y}$ is a vector of variables that always appear in the regressions (in the Levine and Renelt paper, these variables are the initial level of income, the investment rate, the secondary school enrollment rate, and the rate of population growth), z is the variable of interest, and $\mathbf{x}_j \in \mathcal{X}$ is a vector of up to three variables taken from the pool $\mathcal{X}$ of N variables available. One needs to estimate this regression or model for all the possible M combinations of $\mathbf{x}_j \in \mathcal{X}$. For each model j, one finds an estimate, β_{zj}, and a standard deviation, σ_{zj}. The *lower extreme bound* is defined to be the lowest value of $\beta_{zj} - 2\sigma_{zj}$, and the *upper extreme bound* is defined to be the largest value of $\beta_{zj} + 2\sigma_{zj}$. The *extreme-bounds test* for variable z says that if the lower extreme bound for z is negative and the upper extreme bound is positive, then variable z is not robust. Note that this amounts to saying that if one finds a single regression for which the sign of the coefficient β_z changes or becomes insignificant, then the variable is not robust.

Not surprisingly, Levine and Renelt's conclusion is that very few (or no) vari-

* Department of Economics, Columbia University, 420 West 118th St., New York, NY 10027, and Universitat Pompeu Fabra, Barcelona, Spain.

[1] The data for this paper were taken from the World Bank Research Department's Web page.

ables are robust. One possible reason for finding few or no robust variables is, of course, that very few variables can be identified to be correlated systematically with growth. Hence, some researchers' reading of the Levine and Renelt paper concluded that nothing can be learned from this empirical growth literature because no variables are robustly correlated with growth. Another explanation, however, is that the test is too strong for any variable to pass it: if the distribution of the estimators of β_z has some positive and some negative support, then one is bound to find one regression for which the estimated coefficient changes signs if enough regressions are run. Thus, giving the label of nonrobust to all variables is all but guaranteed.

I. Moving Away from Extreme Tests

In this paper I want to move away from this "extreme test." In fact, I want to depart from the zero–one labeling of variables as "robust" vs. "nonrobust," and instead, I want to assign some level of confidence to each of the variables. One way to move away from the extreme-bounds test is to look at the entire distribution of the estimators of β_z. In particular, one might be interested in the fraction of the density function lying on each side of zero: if 95 percent of the density function for the estimates of β_1 lies to the right of zero and only 52 percent of the density function for β_2 lies to the right of zero, one will probably think of variable 1 as being more likely to be correlated with growth than variable 2.[2] The immediate problem is that, even though each individual estimate follows a Student-t distribution, the estimates themselves could be scattered around in a strange fashion. Hence, I will operate under two different assumptions.

[2] Zero divides the area under the density in two. For the rest of the paper, and in order to economize on space, the larger of the two areas will be called CDF(0), regardless of whether this is the area above zero or below zero [in other words, regardless of whether this is the CDF(0) or 1 − CDF(0)].

A. *Case 1: The Distribution of the Estimates of β_z across Models Is Normal*

In order to compute the cumulative distribution function [CDF(0)], one needs to know the mean and the standard deviation of this distribution. For each of the M models, compute the (integrated) likelihood, L_j, the point estimate β_{zj}, and the standard deviation σ_{zj}. With all these numbers one can construct the mean estimate of $\hat{\beta}_z$ as the weighted average each of the M point estimates, β_{zj}:

$$\hat{\beta}_z = \sum_{j=1}^{M} \omega_{zj}\beta_{zj} \tag{3}$$

where the weights, ω_{zj}, are proportional to the (integrated) likelihoods

$$\omega_{zj} = \frac{L_{zj}}{\sum_{i=1}^{M} L_{zi}}. \tag{4}$$

The reason for using this weighting scheme is to give more weight to the regressions or models that are more likely to be the true model. (Incidentally, this is another reason for using regressions with the same number of explanatory variables, since models with more variables will tend to have better fit. To the extent that the fit of model j is an indication of its probability of being the true model, a likelihood-weighted scheme like the one proposed here should be reasonable.)

I also compute the average variance as the weighted average of the M estimated variances, where the weights are given by (4):

$$\hat{\sigma}_z^2 = \sum_{j=1}^{M} \omega_{zj}\sigma_{zj}^2. \tag{5}$$

Once the mean and the variance of the normal distribution are known, I compute the CDF(0) using the standard normal-distribution.

B. *Case 2: The Distribution of the Estimates of β_z across Models Is Not Normal*

If the distribution is not normal, one can still compute its CDF(0) as follows. For each of the M regressions, compute the

individual CDF(0), denoted by $\Phi_{zj}(0/\hat{\beta}_{zj}, \hat{\sigma}^2_{zj})$. Then compute the aggregate CDF(0) of β_z as the weighted average of all the individual CDF(0)'s, where the weights are, again, the integrated likelihoods given by (4). In other words,

$$(6) \qquad \Phi_z(0) = \sum_{j=1}^{M} \omega_{jz}\Phi_{zj}(0/\hat{\beta}_{zj}, \hat{\sigma}^2_{zj}).$$

A potential problem with this method is that it is possible that the goodness of fit of model j may not be a good indicator of the probability that model j is the true model. This might happen, for example, when some explanatory variables in the data set are endogenous: Models with endogenous variables may have a (spurious) better fit. Thus, the weights corresponding to those given to these models will tend to be larger, and in fact, they may very well dominate the estimates. It may be found that only one or two of the models get all of the weight in the estimated weighted average, and these one or two models may suffer from endogeneity bias. It can be argued that, when this is a serious problem, the unweighted average of all the models may be superior to the weighted averages, so I also computed unweighted versions of (3), (5), and (6).

II. Specifications and Data

Even though I depart from Levine and Renelt when it comes to "testing" variables, I keep their specification in the sense that I am going to estimate models like (2). Model j combines some variables which appear in all regressions ($\mathbf{y}$), the variable of interest (z), with the trio $\mathbf{x}_j$ taken from the pool X of the remaining variables proposed in the literature. The reason for keeping some variables in all regressions and the reason for allowing the remaining variables to come only in trios is that the typical growth regression in the literature has (at least) seven right-hand-side variables. I found a total of 62 variables in the literature. If I tested one variable and allowed the remaining 61 to be combined in groups of 6, I would have to estimate 3.4 billion regressions, which would take me about four years to estimate, using my computer.[3] A possible alternative was to run regressions with only three or four explanatory variables. The problem then would be that a lot of the regressions would be clearly misspecified (missing important variables is more of a problem than introducing irrelevant variables). Given these problems, I decided to follow Levine and Renelt and allow all the models to include three fixed variables, so when I combine these three variables along with the tested variable and then with trios of the remaining 59 variables, I always have regressions with seven explanatory variables.

Of all the variables in the literature, I chose a total of 62. The selection was made keeping in mind that I want variables that are measured at the beginning of the period (which is 1960) or as close as possible to it to minimize endogeneity. This eliminated all those variables that were computed for the later years only.

The next thing I needed to do was to choose the three fixed variables (i.e., the variables that appear in all regressions). These variables need to be "good" a priori. By this I mean that they have to be widely used in the literature, they have to be variables evaluated in the beginning of the period (1960) to avoid endogeneity, and they have to be variables that are somewhat "robust" in the sense that they systematically seem to matter in all regressions run in the previous literature. One obvious variable here is the *level of income in 1960,* since most researchers include it in their analysis and find it to be significant (this is the conditional convergence effect). The other two variables chosen are the *life expectancy in 1960* and the *primary-school enrollment rate in 1960.* Both are reasonable and widely used measures of the initial stock of human capital.

In summary, I have a total of 62 variables. I will use three of them in all regressions, so for each variable tested I will combine the remaining 58 variables in sets of three. Hence, I will estimate 30,856 regressions per variable or a total of nearly 2 million regressions. I should mention that, even though I

[3] Some regressions are repeatedly estimated. Repetition could be reduced (and, hence, speed increased), but only at a high cost in terms of memory usage.

do not report these results, I performed the extreme-bounds test on the 59 tested variables and found that only one passes it.[4] However, when I look at the t ratios, I see that some variables are significant almost all of the time (or over 90 percent), while others are significant less than 10 percent or even 1 percent of the time.

III. Results

I will only report here the results for the variables that appear to be "significantly" correlated with growth. By this I mean those variables whose weighted CDF(0) is larger than 0.95. The full results are reported in Sala-i-Martin (1996).[5]

Column (i) of Table 1 reports the estimated weighted mean [described in (3)] of the estimated coefficients for each variable. Column (ii) reports the weighted standard error [described in (5)]. Column (iii) reports the level of significance under the assumption of nonnormality, as described by equation (6) (the levels of significance under normality can be computed by the reader using the average mean and standard deviations reported in columns (i) and (ii), respectively). The table shows that 22 out of the 59 variables appear to be "significant." These variables include the following:

1. Regional Variables: *Sub-Saharan Africa, Latin America* (negatively related to growth), and *Absolute Latitude* (far away from the equator is good for growth). These variables are from the Barro and Jong Wha Lee (1993) data set.[6]
2. Political Variables: *Rule of Law, Political Rights,* and *Civil Liberties* (good for growth); *Number of Revolutions and Military Coups* and *War* dummy (bad for growth). All of these are from the Barro and Lee (1993) data set.
3. Religious Variables: *Confucian, Buddhist,* and *Muslim* (positive); and *Protestant* and *Catholic* (negative). All of these variables are from Barro (1996).
4. Market Distortions and Market Performance: *Real Exchange Rate Distortions* and *Standard Deviation of the Black Market Premium* (both from Barro and Lee [1993] and both negative).
5. Types of Investment: *Equipment Investment*

TABLE 1—MAIN RESULTS OF REGRESSIONS (DEPENDENT VARIABLE = GROWTH)

Independent variable	(i) β	(ii) SD	(iii) CDF[a]
Equipment investment	0.2175	0.0408	1.000
Number of years open economy	0.0195	0.0042	1.000
Fraction Confucian	0.0676	0.0149	1.000
Rule of law	0.0190	0.0049	1.000
Fraction Muslim	0.0142	0.0035	1.000
Political rights	−0.0026	0.0009	0.998
Latin America dummy	−0.0115	0.0029	0.998
Sub-Saharan Africa dummy	−0.0121	0.0032	0.997
Civil liberties	−0.0029	0.0010	0.997
Revolutions and coups	−0.0118	0.0045	0.995
Fraction of GDP in mining	0.0353	0.0138	0.994
SD black-market premium	−0.0290	0.0118	0.993
Primary exports in 1970	−0.0140	0.0053	0.990
Degree of capitalism	0.0018	0.0008	0.987
War dummy	−0.0056	0.0023	0.984
Non-equipment investment	0.0562	0.0242	0.982
Absolute latitude	0.0002	0.0001	0.980
Exchange-rate distortions	−0.0590	0.0302	0.968
Fraction Protestant	−0.0129	0.0053	0.966
Fraction Buddhist	0.0148	0.0076	0.964
Fraction Catholic	−0.0089	0.0034	0.963
Spanish colony	−0.0065	0.0032	0.938

[a] Nonnormal.

[4] The detailed results can be found in Sala-i-Martin (1996).

[5] It turns out that the "levels of significance" found under the assumption of normal distribution and under the assumption of nonnormal distribution are virtually identical. This may indicate that the distribution is close to normal or that, for each variable, there is only one model that takes all the weight.

[6] The data for this paper were taken from the NBER Web page.

and *Non-Equipment Investment* (both positive, although the coefficient for nonequipment investment [$\beta = 0.0562$] is about one-fourth the coefficient for equipment investment [$\beta = 0.2175$]; see Bradford De Long and Lawrence Summers [1991]).[7]

6. Primary Sector Production: Jeffrey Sachs and Andrew Warner's (1995) *Fraction of Primary Products in Total Exports* (negative) and Robert Hall and Charles Jones's (1996) *Fraction of GDP in Mining* (positive).[8]
7. Openness: Sachs and Warner's (1996) *Number of Years an Economy Has Been Open Between 1950 and 1990* (positive).
8. Type of Economic Organization: Hall and Jones's (1996) *Degree of Capitalism* (positive).
9. *Former Spanish Colonies.*

It is interesting to note some of the variables that are not in the table (because they appear not to be important): no measure of government spending (including investment) appears to affect growth in a significant way. The various measures of financial sophistication, the inflation rate, and its variance do not appear to matter much. (In fairness to the authors who proposed these variables, I should say that they specifically say that they affect growth in nonlinear ways, and my analysis allowed these variables to enter in a linear fashion only.) Other variables that do not seem to matter include various measures of scale effects (measured by total area and total labor force), outward orientation, tariff restrictions, the black-market premium, and the recently publicized "ethno-linguistic fractionalization" (which is supposed to capture the degree to which there are internal fights among various ethnic groups).[9]

[7] The data for this paper were taken from the World Bank Research Department's Web page.

[8] The data for Sachs and Warner (1995) were provided by Andrew Warner; the data for Hall and Jones (1996) were taken from Charles Jones's Web page.

[9] See Sala-i-Martin (1996) for the complete list of variables, with their estimated coefficients and levels of significance.

As mentioned earlier, the likelihood-weights used up to now are valid only to the extent that all the models are true regression models. If there are models with spurious good fits, then a nonweighted scheme may be superior. In Sala-i-Martin (1996) I report the detailed results. Suffice to say that only four variables that are above the magic line of 0.95 according to the weighted CDF(0) drop below that mark when an unweighted average of the individual CDF(0)'s is used. These variables are *Civil Liberties, Revolutions and Coups, Fraction of GDP in Mining,* and the *War* dummy. On the other hand, only one variable with a CDF(0) above 0.95 gets a CDF(0) below 0.95: the *Ratio of Liquid Liabilities to GDP*, which is a measure of the degree of financial development.

IV. Conclusions

My claim in this paper is that, if one is interested in knowing the coefficient of a particular variable in a growth regression, the picture emerging from the empirical growth literature is not the pessimistic "nothing is robust" obtained with the extreme bound analysis. Instead, a substantial number of variables can be found to be strongly related to growth.

REFERENCES

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

_____. "Determinants of Democracy." Mimeo, Harvard University, July 1996.

Barro, Robert J. and Lee, Jong-Wha. "International Comparisons of Educational Attainment." *Journal of Monetary Economics*, December 1993, *32*(3), pp. 363–94.

De Long, J. Bradford and Summers, Lawrence. "Equipment Investment and Economic Growth." *Quarterly Journal of Economics*, May 1991, *106*(2), pp. 445–502.

Hall, Robert and Jones, Charles. "The Productivity of Nations." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 5812, November 1996.

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

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

Sachs, Jeffrey and Warner, Andrew. "Economic Reform and the Process of Economic Integration." *Brookings Papers of Economic Activity*, August 1995, (1), pp. 1–95.

_____. "Natural Resource Abundance and Economic Growth." Mimeo, Harvard Institute for International Development, 1996.

Sala-i-Martin, Xavier. "I Just Ran Four Million Regressions." Mimeo, Columbia University, December 1996.