Economics Bulletin

Volume 29, Issue 4

A proxy-variable search procedure

Jaqueson K. Galimberti
Federal University of Santa Catarina (UFSC)

Abstract

This paper proposes a proxy-variable search procedure, based on a sensitivity analysis framework, aiming to provide a useful tool for the applied researcher whenever he faces measurement or proxy-variable uncertainties. Extending from the sensitivity analysis literature it proposes two main methodological innovations. The first relates to the usage of a proxies grouping process to obtain averaged coefficient estimators for theoretical explanatory variables that have more than one possible measure. The second is a proposal of using the actual empirical distribution of the available data to base the inference over the confidence probabilities in choosing each possible measure as proxy for a theoretical variable. This is done using the widely known bootstrapped residuals technique. Besides the methodological main focus, an empirical application is presented in the context of cross-country growth regressions. This empirical application provided favorable evidence to the neoclassical view about the specification of the human capital effect on growth. The results also emphasized how neglecting educational quality differentials might lead to wrong conclusions about the robustness of the relationship between human capital accumulation and economic growth.

Financial support from REUNI program of Brazilian Education Ministry (MEC) is gratefully acknowledged. The author thanks for helpful suggestions from the referee, and for the clear conduction of the submission process by the editor. The program code and data to replicate the empirical work with the econometric software EViews 6.0 is available upon author's request, or from his webpage at: http://sites.google.com/site/jkgeconoeng/.

Citation: Jaqueson K. Galimberti, (2009) "A proxy-variable search procedure", *Economics Bulletin*, Vol. 29 no.4 pp. 2531-2541.
Submitted: Aug 29 2009. **Published:** October 07, 2009.

1. Introduction

Any researcher that has ever tried to econometrically match a theoretical model to its empirical counterpart might has suffered with any kind of estimation uncertainty. In turn, such uncertainties might be arising from diverse sources like theoretical ambiguities, or the existence of competing theories, methodological caveats, measurement errors, and non-direct observance of the theoretical variables. The usual method to circumvent these problems has been to run the so-called sensitivity analysis procedures which provide the researcher some measures of confidence, on Bayesian grounds, that could be put on the results first obtained. However, the literature surrounding such sensitivity procedures has mainly focused on the model uncertainty issues, leaving an unfilled gap regarding the arising uncertainties from the measures choice problem. This later source of uncertainty became known as the proxy-variable search problem since Leamer's (1978) work, and it is the main focus of this paper. Whilst theory usually provides clear pictures about the relationships expected to be found in the real world, it seldom specifies, between the available measures, which one is the one that best represents each of its theoretical variables.

On this context the main aim of this paper is to propose a proxy-variable search procedure, based on a sensitivity analysis framework, which is intended to be a useful tool for the applied researcher whenever he faces measurement or proxy-variable uncertainties. The paper is outlined in the following sections. In Section 2 the proposed procedure is contextualized on the previous literature, and then described. Section 3 presents an empirical application to cross-country growth regressions. Finally the paper ends with some concluding remarks.

2. Proxy-variable Search Procedure

Suppose that from a theoretical point of view a given variable $\boldsymbol{\gamma}$ might be explained by a pool of *N* explanatory variables **X**, according to the general specification in equation (1). If all of these variables are directly observable, and there is no uncertainty about it, an applied researcher would easily obtain the parameters from this equation using an appropriate estimation method.

$$\boldsymbol{\gamma} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} \tag{1}$$

Now, in a situation where there is more than one possible measure for one or more theoretical explanatory variables the researcher will have to deal with the proxy-variable uncertainty. The proxy-variable search procedure here proposed is an attempt to provide an objective method to deal with this problem. It consists of an adaptation of the sensitivity analysis proposed by Sala-i-Martin (1997), where the main difference relates to the presence of subsets of the explanatory variables that have more than one possible proxy. This kind of proxies grouping has already been used by Crain and Lee (1999) with an Extreme-Bounds Analysis (EBA) framework. However, as pointed out by Sala-i-Martin (1997) such EBAs, when viewed as tests, have the drawback of extremely labeling potential explanatory variables as "robust" or "nonrobust".

Adopting Sala-i-Martin's (1997) notation, consider the following general form regression:

$$\boldsymbol{\gamma} = \boldsymbol{\alpha}_{\mathbf{j}} + \boldsymbol{\beta}_{\mathbf{yj}}\mathbf{y} + \beta_{zj} z + \boldsymbol{\beta}_{\mathbf{xj}}\mathbf{x}_{\mathbf{j}} + \boldsymbol{\varepsilon} \tag{2}$$

where **y** is a vector of fixed variables, i.e., those theoretical variables that have only one available measure, and thus must appear in all regressions, z is one possible proxy from a subset **Z** of possible measures for one of the theoretical explanatory variables from **X**, and $\mathbf{x_j}$ is the j possible combination of measures for the remaining theoretical explanatory variables from **X**. If g is the number of theoretical explanatory variables with more than one possible measure (or the number of groups), and m_i is the number of possible measures for each of these groups, the total number of regressions, M, could be calculated from equation (3).

$$M_Z = \frac{\prod_{i=1}^{g} m_i}{m_z} \qquad M = \sum_{z=1}^{g} M_z = g\prod_{i=1}^{g} m_i \tag{3}$$

The basic idea is to draw the distribution of the slope coefficient estimators (β_{zj}) of each possible proxy for the theoretical variables, averaging these estimators over all the possible specifications while letting fixed the proxy and the others variables that have a unique directly observable measure. This idea could be synthesized in the following three rules: (i) the sensitivity analysis should be done on every variable for which there is data availability of more than one measure, or more than one possible functional form to enter into the model specification, and those variables that do not meet this condition should be considered as a fixed variable; (ii) one measure of each explanatory variable should be included in each regression; (iii) for each variable measure or functional form that the sensitivity analysis is applied, every possible combination with the other explanatory variables measures should be a regression.

As soon as one obtains all these estimates the next question is how to infer about their distribution. In Sala-i-Martin (1997), the confidence measure is defined as the fraction of the density function over the side[1] where the larger area of the distribution lies, regardless of whether this is in accordance with the theoretically expected sign for the variable. This is justifiable whenever the expected sign for the variable is ambiguous, or there is no theoretical prior about it. So, another adaptation in the method here refers to the way the confidence probabilities are calculated from the averaged slope coefficient estimators. In the present proposed procedure the researcher is allowed to define, based on his theoretical priors, from which side of the distribution the probabilities are calculated, while still allowing for the ambiguous case too.

To compute the estimator's cumulative distribution function (CDF), Sala-i-Martin (1997) proposed two distinct assumptions. First, the distribution of the estimates of β_z across models may be normal, where the mean and the standard deviation are obtained from equations (4) and (5), respectively. Notice that a weighting scheme based on the (integrated) likelihoods (L_{zj}) is used to give more weight to the regressions that are more likely to be the true model. While this weighting scheme is the main subject of the recent studies of Bayesian Model Averaging (BMA), this paper simply follows the same criteria used by Sala-i-Martin (1997), given by equation (6). For a literature review on BMA see Hoeting, Madigan, Raftery and Volinsky (1999).

$$\hat{\beta}_z = \sum_{j=1}^{M_z} \omega_{zj}\beta_{zj} \tag{4}$$

[1] Zero divides the area under the density in two.

$$\hat{\sigma}_z^2 = \sum_{j=1}^{M_z} \omega_{zj} \sigma_{zj}^2 \qquad (5)$$

$$\omega_{zj} = \frac{L_{zj}}{\sum_{i=1}^{M_z} L_{zj}} \qquad (6)$$

A second case would be to assume that the distribution of the estimates of β_z across models is not normal, obtaining the aggregate CDF as the weighted average of all the individual models CDF's (Φ_{zj}), as showed in equation (7).

$$\Phi_z(0) = \sum_{j=1}^{M_z} \omega_{zj} \Phi_{zj}\left(0/\hat{\beta}_{zj}, \hat{\sigma}_{zj}^2\right) \qquad (7)$$

Clearly, inherent to both of these cases is the assumption of normality of the distribution of the residuals (ε) from the estimated regressions of equation (2). As this assumption may be too strong, depending on the context of study, a third case would be to use the actual empirical distribution of the available data. This extension is here proposed using the plug-in principle with a bootstrapped residuals technique (Efron, 1979), which may be summarized in the following four steps. First, the parameters from (2) are estimated for all the M_z regressions for the proxy z under evaluation, also calculating its estimated residuals. Second, a residuals bootstrap sample is built drawing with replacement from the estimated residuals, and then 'plugged-in' with the parameters estimates to generate a bootstrapped sample of the dependent variable (y*). Third, this latter bootstrapped sample is used to estimate new values for the parameters of the model, specially our focused slope coefficient (β_{zj}). Fourth, the second and third steps are repeated many times in order to obtain a bootstrapped distribution of the proxy slope coefficient, from which it is straightforward to obtain the fraction of this empirical distribution lying on each side of zero.

Finally, after obtaining the confidence measures of all possible proxies for the theoretical variables, it remains only to compare then and choose the ones with the greater probability as the best empirical representations for the theoretical model under study.

3. Empirical Application: cross-country growth regressions

One of the most highlighted research fields in the recent economic growth empirics has been the model uncertainty issue. The seminal work on this field was the Levine and Renelt (1992) sensitivity analysis which used a variant of Leamer's (1983) extreme bounds analysis to test for the robustness of coefficient estimates to the inclusion of other relevant variables on growth equations. Their mainly finding was that very few macroeconomic variables are robustly correlated with cross-country growth rates. As already pointed, an alternative model averaging procedure was proposed by Sala-i-Martin (1997), which yielded less severe results based on a less restrictive concept of robustness. Nevertheless, there are still many variables that are theoretically expected to be important, but are found to be not significantly correlated with growth. According to Brock and Durlauf (2001), the inclusion or exclusion of most variables is typically arbitrary, a phenomenon labeled the "open-endedness" of growth theory.

In order to illustrate the proxy-variable search procedure from the previous section considers the general cross-country empirical specification of equation (8). As it can be seem this specification has as basis the main traditionally accepted factors of economic growth added of an export variable, which is intended to represent the so-called export-led led growth (ELG) hypothesis.

$$\gamma_i = \alpha + \beta_1 Y_i^0 + \beta_2 N_i + \beta_3 I_i + \beta_4 H_i + \beta_5 X_i + \varepsilon_i \quad (8)$$

where γ_i is the growth rate of output per worker, Y_i^0 is a measure of the initial level of output per worker, N_i is a measure for the growth rate of the labor force, I_i is a measure of the ratio of investment to GDP, H_i is a measure of human capital per worker, X_i is a measure of exports output, ε_i is the i.i.d. error term with mean zero and finite variance σ^2, and the subscript *i* refers to the country.

While there is some agreement about the relevance of the first four factors of growth over country's performances, there is a great theoretical controversy about their specification relative to growth rates, especially about the human capital factor. From a neoclassical point of view, which is derived from the human-capital-augmented neoclassical growth model of Mankiw, Romer and Weil (1992), differences in the levels of human capital stock are related to differences in output levels across countries and growth rates of human capital stock should be connected to growth rates of output. Now from the endogenous growth models standpoint, going at least as far back as to Nelson and Phelps (1966), differences in the levels of human capital stock are related to differences in output growth across countries.

Regarding the effects of exports on growth the controversy comes mainly from an empirical standpoint. Once one accepts that the hypothesis of export-led growth is theoretically plausible it remains the question about how an export measure must enter in the growth regressions specifications. Giles and Williams (2000) surveyed much of the empirical literature regarding the ELG hypothesis, where the most widely used measures were the ratio of exports to output, the growth rate of exports, and the product between these two.

This discussion gives some justification for the necessity of a proxy-variable search procedure in order to allow the applied researcher to circumvent the uncertainty problem associated with the estimation of growth equations. Before applying the proxy-variable search over the specification of equation (8) a description of the dataset is provided, to then turn to the results obtained.

The gross data comes mainly from the Penn World Tables v.6.2 (Heston et. al., 2006) which offers internationally comparable annual macroeconomic data for almost all of the world economies. The sample consists of 72 countries selected according to the criteria described at the Appendices section. All the selected data refers to the constant prices entries and covers the period from 1974 to 2003. The unique exceptions are the human capital data which are computed into a five-year basis and are based on the average years of schooling from the Barro and Lee (2000) Dataset, and the General Index of Qualitative Indicators of Human Capital (QIHC-G) recently built by Altinok and Murseli (2007).

Two distinct measures of human capital stock were constructed following Wössmann (2003), and are both based on the Mincerian human capital theory with decreasing returns to education. While the first specification (9) assumes identical quality of education, the second specification (10) accounts for quality differentials in education between the countries.

$$H_i^M = e^{\sum_a r_a s_{ai}} \quad (9)$$

$$H_i^Q = e^{\sum_a r_a Q_i s_{ai}} \quad (10)$$

where r_a is the rate of return to education at schooling level a^2, s_{ai} is the average years of schooling at level a for country i, and Q_i is the QIHC-G for country i.

As the gross data was obtained in a panel data format (countries/years observations) some averaging procedure is needed to convert it to a cross-country format. Such procedure may also lead to the same specification uncertainties discussed in the previous sections. Thus, using some of the several possibilities to convert the gross data made it possible to obtain the measures that will serve as the input for the proxy-variable search procedure. Details about the construction of these cross-country measures are provided at Table A.2 in the Appendices section. Notice that as more than one measure was obtained for each of the theoretical explanatory variables there was no fixed variable in this illustration.

The total number of regressions on this experiment was of 13,440 and its results are presented at Table 1, where the first probability measure refers to the case where the distribution of the slope parameter estimates is assumed to be normal, the second refers to the case where that distribution is assumed to be not normal, and the last measure refers to the bootstrapped confidence probabilities, obtained from 2,000 replications[3].

With these results at hand it remains only to compare then and choose the best proxy for each theoretical variable. This is done by finding those proxies that showed the greater confidence probability. Whether a draw is observed the next criteria could be to take the specification which yielded the greater (integrated) likelihood. Notice that the computation of the third probability based on the bootstrapped actual data might be also useful as a tiebreaker when the first two probabilities give ambiguous results.

The results obtained from this illustration are quite singular to choose the following specification: the intercept estimate from the logarithm GDP trend growth regression[4] as proxy for the initial level of output per worker; the population (trend) growth rate as proxy for the growth rate of the labor force; the product between the ratio of investment to GDP and its (trend) growth as proxy for the investment; the (trend) growth rate of the quality-adjusted measure of human capital stock as proxy for human capital; and the exports (trend) growth rate as proxy for exports output.

While this experiment serves just as an illustration to the proposed procedure, its results allow some interesting observations. First, between all the theoretical variables the considered possible proxies for the human capital are the ones that appeared to lead to the greater uncertainty. Even so, the result of choosing a human capital growth rate measure as the best proxy for the human capital effect on growth is favorable evidence to the neoclassical view about this relationship. Second, notice also that the chosen proxy for the human capital was the quality-adjusted measure of human capital, even though the regressions using this measure had 15 available observations less than the non-adjusted measures.

[2] The rates of return to education are considered to be the same for all countries, obtained from the estimates of the world-average social rates of return to education by Psacharopoulos (1994, Table 2) corresponding to 20.0% at the primary level, 13.5% at the secondary level, and 10.7% at the higher level.

[3] All of the presented measures were obtained with the likelihood-weighted scheme. Even though the developed procedure also computes the unweighted probabilities, these latter results did not differ qualitatively from the former ones. Thus, they were suppressed from the presentation.

[4] To clarify what it is meant by this estimate see Table A.2 and its explanatory text.

Table 1 – Results for the Proxy-variable Search Procedure.

Variables / Measures		Average Estimates		Confidence Probabilities			No Regs
		Coefs.	Std.Dev.	Normal	Not Norm.	Bootstrap	
Initial Level of Output per worker	Y-ant	-4.12x10^{-7}	1.82x10^{-7}	0.9881	0.9429	0.9990	448
	ln(Y-ant)	-0.008396	0.002079	0.9999	0.9983	1.0000	448
	Y-avg	-4.23x10^{-7}	1.88x10^{-7}	0.9876	0.9435	0.9990	448
	ln(Y-avg)	-0.008301	0.002070	0.9999	0.9982	1.0000	448
	Y-alpha	-5.03x10^{-7}	1.90x10^{-7}	0.9958	0.9620	1.0000	448
	ln(Y-alpha)	-0.008776	0.001993	**0.9999**	**0.9988**	**1.0000**	448
Labor Force	Lab-growth	-0.494662	0.139330	0.9998	0.9985	1.0000	1,344
	Pop-growth	-0.650145	0.173772	**0.9999**	**0.9991**	**1.0000**	1,344
Investment	I-share	0.112071	0.027521	0.9999	0.9973	1.0000	672
	I-growth	0.419328	0.053578	1.0000	0.9999	1.0000	672
	I-product	1.967426	0.220456	**1.0000**	**1.0000**	**1.0000**	672
	ln(I)	0.015178	0.004368	0.9997	0.9754	1.0000	672
Human Capital	Hm-ant	0.000189	0.000881	0.5849	0.5831	0.6265	192
	ln(Hm-ant)	-0.001554	0.004679	0.3699	0.4115	0.2560	192
	Hm-avg	0.000502	0.000788	0.7378	0.7094	0.8555	192
	ln(Hm-avg)	0.002965	0.004745	0.7340	0.6527	0.8405	192
	Hm-alpha	0.000296	0.000818	0.6414	0.6274	0.7060	192
	ln(Hm-alpha)	0.000727	0.004600	0.5627	0.5387	0.5550	192
	Hm-growth	0.196666	0.171482	0.8743	0.7559	0.9880	192
	Hq-ant	-0.000277	0.001025	0.3937	0.3928	0.3265	192
	ln(Hq-ant)	-0.004185	0.005306	0.2151	0.2684	0.1135	192
	Hq-avg	0.000173	0.000922	0.5746	0.5521	0.5840	192
	ln(Hq-avg)	0.001019	0.005108	0.5791	0.5386	0.5775	192
	Hq-alpha	-0.000131	0.000947	0.4449	0.4331	0.3825	192
	ln(Hq-alpha)	-0.001568	0.005099	0.3792	0.3919	0.2765	192
	Hq-growth	0.377988	0.227817	**0.9515**	**0.9187**	**0.9945**	192
Exports	X-share	0.013579	0.006549	0.9809	0.9490	1.0000	672
	ln(X-share)	0.003785	0.002348	0.9465	0.9199	0.9950	672
	X-growth	0.264957	0.051244	**1.0000**	0.9976	**1.0000**	672
	X-product	0.214365	0.060425	0.9998	**0.9980**	1.0000	672

Notes: the bold probabilities are the best choice for each theoretical variable.

Finally, the bootstrapped confidence probabilities computed were quite proximate from the other two probabilities computed on the basis of the assumptions of normality and non-normality of the distribution of the slope coefficients estimates. While this result might indicate that this computational costly procedure provided little new information, it is important to emphasize that its proposition had as its justification a context where the residual normality hypothesis is doubtful. In a short check, a Jarque-Bera normality test over the residuals from the chosen specification resulted in a non-rejection of the null hypothesis of normally distributed errors, under the most usual significance levels.

Concluding Remarks

This paper proposed a proxy-variable search procedure, based on a sensitivity analysis framework. The main aim in developing this procedure was to provide a useful tool for the applied researcher whenever he faces measurement or proxy-variable uncertainties. Extending from the sensitivity analysis literature, especially from the prominent contribution of Sala-i-Martin (1997), it proposes two main methodological innovations. The first relates to the usage of a proxies grouping process to obtain averaged coefficient estimators for theoretical explanatory variables that have more than one possible measure, leading to measurement uncertainties. The second is a proposal of using the actual empirical distribution of the available data to base the inference over the confidence probabilities in choosing each possible measure as proxy for a theoretical variable. This is done using the widely known bootstrapped residuals technique.

An illustration of an empirical application to the procedure developed is presented in the context of cross-country growth regressions. While this illustrative example is on a cross-section regressions context, it is important to remark that the procedure could be easily extended to a panel data context. Two main concerns in such an extension would arise. First the bootstrapping re-sampling procedure would have to be focused only on the cross-sectional units in order to keep their time series properties. Second, the usual question of whether it is appropriate to estimate the individual effects as random or fixed effects would arise leading to another source of estimation uncertainty that would have to be taken into account in the procedure.

Finally, besides the methodological main focus, the empirical application provided favorable evidence to the neoclassical view about the specification of the human capital effect on growth. The results also emphasized how neglecting educational quality differentials might lead to wrong conclusions about the robustness of the relationship between human capital accumulation and economic growth.

References

Altinok, N. and H. Murseli (2007) “International database on human capital quality” *Economics Letters* **96**, p. 237-244.

Barro, R. J. and J. Lee (2000) “International Data on Educational Attainment: Updates and Implications” *Cid Working Paper* **42**, p. 1-33.

Brock, W. and S. N. Durlauf (2001) “Growth empirics and reality” *World Bank Economic Review* **15**, p. 229-272.

Crain, W. M. and K. J. Lee (1999) “Economic Growth Regressions for the American States: A Sensitivity Analysis” *Economic Inquiry* **37**(2), p. 242-257.

Deaton, A. and A. Heston (2008) “Understanding PPPs and PPP-based national accounts” *NBER Working Papers* **14499**, p. 1-53.

Efron, B. (1979) “Bootstrap Methods: Another Look at the Jackknife” *The Annals of Statistics* **7**(1), p. 1-26.

Giles, J. A. and C. L. Williams (2000) “Export-led growth: a survey of the empirical literature and some non-causality results. Part 1” *Journal of International Trade & Economic Development* **9**(3), p. 261-337.

Heston, A., R. Summers, and B. Aten (2006) “Penn World Table Version 6.2” *Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania*.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999) “Bayesian Model Averaging: A Tutorial” *Statistical Science* **14**(4), p. 382-401.

Kakwani, N. (1997) “Growth Rates of Per-Capita Income and Welfare: An International Comparison” *The Review of Economics and Statistics* **79**(2), p. 201-211.

Leamer, E. E. (1978) *Specification searches: ad hoc inference with nonexperimental data*. New York: Wiley.

Leamer, E. E. (1993) “Let’s Take the Con out of Econometrics” *American Economic Review* **73**, p. 31-43.

Levine, R. and D. Renelt (1992) “A sensitivity analysis of cross-country growth regressions” *American Economic Review* **82**, p. 942-963.

Mankiw, N. G., D. Romer, and D. N. Weil (1992) “A Contribution to the Empirics of Economic Growth” *The Quarterly Journal of Economics* **107**(2), p. 407-437.

Nelson, R. R. and E. S. Phelps (1966) “Investment in Humans, Technological Diffusion, and Economic Growth” *American Economic Review* **56**(2), p. 69-75.

Psacharopoulos, G. (1994) “Returns to Investment in Education: A Global Update” *World Development* **22**, p. 1325-1343.

Sala-i-Martin, X. X. (1997) “I Just Ran Two Million Regressions” *American Economic Review* **87**(2), Papers and Proceedings of the Hundred and Fourth Annual Meeting of the American Economic Association, p. 178-183.

Wössmann, L. (2003) “Specifying Human Capital” *Journal of Economic Surveys* **17**(3), p. 239-270.

Appendices: Sample and Data Specifications

The sample consists of 72 countries selected according to the following criteria: (i) data availability for the period from 1974 to 2003; (ii) exclusion of countries for which oil production is the dominant industry[5]; (iii) exclusion of countries whose data receive a grade "D" from the Penn World Tables (Deaton and Heston, 2008); (iv) exclusion of countries whose populations in 1974 were less than one million. Table A.2 presents a list of the selected countries.

Table A.1 – Countries included in the sample.

Continent / Income Class	Low and Lower Middle Income (obs. = 34)	Upper Middle and High Income (obs. = 38)
Africa (obs. = 19)	Benin (BEN), Cameroon (CMR), Republic of Congo (COG)*, Egypt (EGY), Ghana (GHA), Jordan (JOR), Kenya (KEN), Malawi (MWI), Mali (MLI), Rwanda (RWA)*, Senegal (SEN), Sierra Leone (SLE)*, Syria (SYR)*, Tanzania (TZA), Tunisia (TUN), Zambia (ZMB), Zimbabwe (ZWE).	Israel (ISR), South Africa (ZAF).
America (obs. = 19)	Bolivia (BOL), Colombia (COL), Dominican Republic (DOM), El Salvador (SLV)*, Guatemala (GTM)*, Honduras (HND), Nicaragua (NIC)*, Paraguay (PRY), Peru (PER)*.	Argentina (ARG), Brazil (BRA), Canada (CAN), Chile (CHL), Costa Rica (CRI)*, Jamaica (JAM)*, Mexico (MEX), Panama (PAN)*, United States (USA), Uruguay (URY).
Asia/Oceania (obs. = 16)	China (CHN), India (IND)*, Indonesia (IDN), Nepal (NPL)*, Pakistan (PAK)*, Philippines (PHL), Sri Lanka (LKA)*, Thailand (THA).	Australia (AUS), Hong Kong (HKG), Japan (JPN), Republic of Korea (KOR), Malaysia (MYS), New Zealand (NZL), Singapore (SGP), Turkey (TUR).
Europe (obs. = 18)		Austria (AUT), Belgium (BEL), Denmark (DNK), Finland (FIN), France (FRA), Germany (GER), Greece (GRC), Hungary (HUN), Ireland (IRL), Italy (ITA), Netherlands (NLD), Norway (NOR), Poland (POL), Portugal (PRT), Spain (ESP), Sweden (SWE), Switzerland (CHE), United Kingdom (GBR).

Note: *Countries without data on the quality of education, which counts to 12 for the Low and Lower Middle Income Class and 3 for the Upper Middle and High Income Class.

Regarding the data, except for those of human capital that were constructed from the Barro & Lee Dataset (B&L) and the Altinok & Murseli tables (A&M) using equations (9) and (10), every others measures were obtained directly or indirectly[6] from the Penn World Tables v.6.2 (PWT). Table A.2 presents details about data sources and the measures construction.

[5] The countries excluded on this basis are the same of Mankiw, Romer and Weil (1992) in addition to the OPEC countries: Algeria, Bahrain, Ecuador, Gabon, Iran, Iraq, Nigeria, Oman, Qatar, Saudi Arabia, United Arab Emirates, and Venezuela. Lesotho is also excluded because the sum of private and government consumption far exceeds GDP in most part of the years, indicating that labor income from abroad constitutes an extremely large fraction of GNP.

[6] Two measures were indirectly obtained: (i) Labor force stock: *LAB=(RGDPCH*POP)/RGDPWOK*; and, (ii) Exports: *Exports/GDP=(OPENK+KNFB)/200*, where the net foreign balance as a percentage of the GDP (*KNFB*) can be obtained by the formula *100-KC-KI-KG=KNFB*, where *KC*, *KI*, and *KG* are the percentage shares of consumption, investment and government spending, respectively, in GDP.

Table A.2 – Data Sources and Construction.

Variables / Measures		Description	Source	
			Table	Entry
Output per worker (obs.=72)	γ	Log-Lin trend growth rate. (*)	PWT	RGDPWOK
	Y-ant	RGDP at 1973.	PWT	RGDPWOK
	ln(Y-ant)	Log. RGDP at 1973.	PWT	RGDPWOK
	Y-avg	Average RGDP (1970-73).	PWT	RGDPWOK
	ln(Y-avg)	Log. Avg. RGDP (1970-73).	PWT	RGDPWOK
	Y-alpha	Exp. Log-Lin intercept. (*)	PWT	RGDPWOK
	ln(Y-alpha)	Log-Lin intercept. (*)	PWT	RGDPWOK
Labor Force (obs.=72)	Lab-growth	Log-Lin trend growth rate. (*)	PWT	RGDPCH, POP, RGDPWOK
	Pop-growth	Log-Lin trend growth rate. (*)	PWT	POP
Investment (obs.=72)	I-share	Average share on GDP (1974-03).	PWT	KI
	I-growth	Log-Lin trend growth rate. (*)	PWT	KI
	I-product	Product between I-share and I-growth.	PWT	KI
	ln(I)	Log. Avg. share on GDP.	PWT	KI
Mincerian Human Capital (obs.=72)	Hm-ant	Hm at 1970.	B&L	TYR15
	ln(Hm-ant)	Log. Hm at 1970.	B&L	TYR15
	Hm-avg	Average Hm (1970-00).	B&L	TYR15
	ln(Hm-avg)	Log. Average Hm (1970-00).	B&L	TYR15
	Hm-alpha	Exp. Hm Log-Lin intercept. (*)	B&L	TYR15
	ln(Hm-alpha)	Hm Log-Lin intercept. (*)	B&L	TYR15
	Hm-growth	Hm Log-Lin trend growth rate. (*)	B&L	TYR15
Quality-adjusted Human Capital (obs.=57)	Hq-ant	Hq at 1970.	B&L, A&M	TYR15, QIHC-G
	ln(Hq-ant)	Log. Hq at 1970.	B&L, A&M	TYR15, QIHC-G
	Hq-avg	Average Hq (1970-00).	B&L, A&M	TYR15, QIHC-G
	ln(Hq-avg)	Log. Average Hq (1970-00).	B&L, A&M	TYR15, QIHC-G
	Hq-alpha	Exp. Hq Log-Lin intercept. (*)	B&L, A&M	TYR15, QIHC-G
	ln(Hq-alpha)	Hq Log-Lin intercept. (*)	B&L, A&M	TYR15, QIHC-G
	Hq-growth	Hq Log-Lin trend growth rate. (*)	B&L, A&M	TYR15, QIHC-G
Exports (obs.=72)	X-share	Average share on GDP (1974-03).	PWT	OPENK, KNFB
	ln(X-share)	Log. Avg. share on GDP.	PWT	OPENK, KNFB
	X-growth	Log-Lin trend growth rate. (*)	PWT	OPENK, KNFB
	X-product	Product between X-share and X-growth.	PWT	OPENK, KNFB

Notes: those measures marked with an "*" were obtained by least squares estimation of log-linear trend regressions of the form: $ln\ x_{it} = \alpha_t + \beta_t t + \varepsilon_{it}$, where *i* and *t* indexes for the country and the year, respectively. From these estimates one obtains possible proxies for the initial level (the intercept estimates α_i) and for the growth rate of the focused variable. Notice that many other procedures of averaging through time could be added to enrich the analysis here proposed. An interesting discussion about aggregation over time could be found at Kakwani (1997).