Columbia University

Department of Economics
Discussion Paper Series

AN ANALYSIS OF SAMPLE SELECTION BIAS IN CROSS-COUNTRY GROWTH REGRESSIONS

Jayant Ray
Francisco L. Rivera-Batiz

Discussion Paper #:0102-10

Department of Economics
Columbia University
New York, NY 10027

February 2002

Columbia University
Department of Economics Discussion Paper No. 0102-10
An Analysis of Sample Selection Bias in Cross-Country Growth Regressions[*]
Jayant Ray and Francisco L. Rivera-Batiz
February 2002

Abstract:

Sample sizes in cross-country growth regressions vary greatly, depending on data availability. But if the selected samples are not representative of the underlying population of nations in the world, ordinary least squares coefficients (OLS) may be biased. This paper re-examines the determinants of economic growth in cross-sectional samples of countries utilizing econometric techniques that take into account the selective nature of the samples. The regression results of three major contributions to the empirical growth literature by Mankiw-Romer-Weil (1992), Barro (1991) and Mauro (1995), are considered and re-estimated using a bivariate selectivity model. Our analysis suggests that sample selection bias could significantly change the results of empirical growth analysis, depending on the specific sample utilized. In the case of the Mankiw-Romer-Weil paper, the value and statistical significance of some of the estimated coefficients change drastically when adjusted for sample selectivity. But the results obtained by Barro and Mauro are robust to sample selection bias.

[*] The authors are grateful to Francisco Ciocchini, Mitali Das, Phoebus Dhrymes, Marc Henry, Rohini Pande, David Weinstein and the participants of a seminar at Columbia University for helpful comments and suggestions.

I. INTRODUCTION

Numerous empirical studies have emerged in the last decade analyzing the determinants of economic growth in the world [see, for example, Barro (1991), Mankiw, Romer and Weil (1992), Mauro (1995), Sachs and Warner (1997), Easterly and Levine (1997), Knack and Keefer (1997), Rodrik (1998), and Hanushek and Kimko (2000), among many others]. No doubt this growing literature has been stimulated by the emergence and proliferation of cross-country growth databases [see Barro and Lee (1994), and World Bank (2000a)]. But despite the greater availability of information in recent years, the databases utilized in growth regressions have substantial gaps. A cursory glance at recent papers in the area shows that the selected samples can vary from a few dozen countries to over one hundred.[1]

Although sample selection is sometimes intentional (as when the determinants of growth are examined within a specific region), countries are generally excluded from analysis due to data gaps in key variables under study. The selected samples may therefore differ sharply from the underlying population of countries in the world. As a result, ordinary least squares and other methods utilized to estimate the cross-country growth equations could yield biased coefficients. This is particularly significant since many poor and so-called transition economies are excluded from cross-country samples, yet the results of the empirical growth regressions are meant to apply especially to them, as nations facing the early stages of the development process.

This paper re-examines the determinants of economic growth in a cross-section of countries, utilizing econometric techniques that take into account the selected nature of the sample.[2] More specifically, we reproduce and re-analyze the empirical models estimated in three papers published in the *Quarterly Journal of Economics* during the 1990s: Barro (1991),

[1] For instance, the sample size in the output growth analysis of Barro (1997) varies between 87 and 90, in Levine and Zervos (1998) the countries range from 32 to 45, in Knack and Keefer (1997) it is 29 countries, and in Hanushek and Kimko (2000) the number of observations is between 78 and 80.

[2] This paper will focus exclusively on sample selection bias issues. Others have examined a variety of additional econometric problems in the empirical growth literature, including causality, functional specification, sensitivity to inclusion and exclusion of particular variables in the equation, outliers, etc. See, for example, Levine and Renelt (1992), Krueger (1998), Temple (1999), Durlauf (2000), and Bils and Klenow (2000).

Mankiw, Romer and Weil (1992) and Mauro (1995). These papers were chosen not only because they have been and continue to be widely cited in the literature, but also because they differ widely in the variables they include as determinants of growth. The sample sizes in the papers also diverge, ranging from a low of 58 in the case of Mauro (1995) to a high of 98 in Barro (1991).

It will be shown that sample selection bias is present in some of the regressions examined. Furthermore, econometric methods used to re-estimate these equations taking into account sample selectivity can yield significant differences in the estimated impact of the variables explaining economic growth. At the same time, no evidence of sample selection bias is found in some of the equations analyzed in this paper. Indeed, some OLS regression coefficients display remarkable robustness to sample selectivity. This robustness applies to growth regression studies with even comparatively small sample sizes.

The next section of the paper presents the econometric framework and methodology used to examine the sample selection issue. Section III examines the variables determining the selection of countries in cross-sectional growth samples. Section IV presents a re-analysis of the Mankiw, Romer and Weil paper using an econometric model that incorporates a sample selection mechanism in addition to the growth regression model. Sections V and VI report the result of similar analyses of the Barro (1991) and Mauro (1995) papers. Section VII summarizes our main conclusions.

II. SAMPLE SELECTION BIAS IN EMPIRICAL CROSS-COUNTRY GROWTH ANALYSIS

The econometric framework underlying most cross-country growth analysis is:

$$y_i = \alpha' X_i + \varepsilon_i \qquad (1)$$

$$\varepsilon_i \sim N [0, \sigma^2_\varepsilon] \qquad (2)$$

where the dependent variable y_i is some measure of per-capita income growth or per-capita income for country i, X_i is a vector of explanatory variables influencing the dependent variable, and ε_i is a random error term, assumed to be distributed normally with mean zero and variance σ^2_ε. Under these, classical, conditions, ordinary least squares (OLS) is the best linear unbiased estimator of α'.

The problem with this analysis rests on the assumption that ε_i in equation (2) represents an error term that is randomly distributed among the population. In the cross-country growth literature, the sample of countries selected to estimate equation (1) is chosen on the basis of data availability; a selection criterion that, as we shall show below, does not necessarily constitute a random sample of the world population of countries. Thus there is a potential sample selection bias problem. In particular, if we take expectations in equation (1), then:

$$E\ (y_i\ *\ X_i \text{ and } y_i \text{ are observed}) = \alpha'\ X_i\ +\ E(\varepsilon_i\ *\ X_i \text{ and } y_i \text{ are observed}), \qquad (3)$$

where X_i is, as before, the vector of country characteristics observed in the sample. The expectation of the error term in the right-hand side of equation (3) is taken conditional on data being available for X_i and y_i and may differ from zero, thereby making the least squares estimates of equation (1) subject to specification error and therefore biased.

The critical issue, both theoretically and empirically, is what determines the inclusion of countries in the selected sample used in the regression equation in (1). If cross-country data were available randomly, the conditional expectation on the right-hand side of equation (3) would be zero and selectivity bias would not be a problem. If, on the other hand, the availability of required data were not randomly determined, then selection bias could be a significant issue

Most cross-country data are aggregate in nature and are usually produced by government statistical agencies. Whether a government invests in collecting, processing and publishing reliable information on a timely basis depends on the relative costs and benefits involved in carrying out these activities. What are the costs and benefits of supplying data?

Investing in the physical infrastructure required to collect reliable data is a costly venture that a country may not be able or willing to fund. The government also needs to employ the skilled human resources required to design reliable sampling procedures and to analyze the collected data on a regular basis. This human capital may be scarce in poor countries and its expense prohibitive. However, the cost of the materials, equipment and human resources involved may be miniscule compared to the non-economic (political) costs of making such data available. Published socioeconomic data may prove damaging to a government and its policies. It is well known, for example, that GDP (and other) data supplied by former socialist countries like Russia and the former Soviet Republics were utterly inaccurate and were often doctored to provide the rest of the world with a positive view of socialism. In developing countries, corrupt governments can severely limit the supply of reliable socioeconomic data as they manipulate information to provide a rosy picture of the performance of their regimes.

The key benefit of having reliable public data on key economic variables available is the resultant flow of information that this allows, to policymakers, the business community and the public in general. Spending and investment decisions made by the public and private sectors are crucially dependent on the use of socioeconomic data. By bounding the rationality of public and private economic decisions, the absence of accurate data on income, employment, interest rates, investment rates, inflation, educational attainment, etc. can have disastrous consequences. In addition, the availability of transparent economic data may be necessary for a country to become a member of international organizations such as the OECD and the International Monetary Fund. Such memberships can provide direct economic benefits to governments but also serve as signals to global market participants, allowing improved access to international trade and capital markets

Let C_i and B_i respectively represent the costs and benefits of collecting, processing and publishing information in country i, where any non-economic costs and benefits are attached a monetary equivalent. Then, in judging whether to gather reliable public data, the government will compare C_i and B_i, and make a decision on the basis of whether the profit function, Z_i , is positive or negative:

$$Z_i = C_i - B_i = \beta' V_i + U_i \qquad (4)$$

where V_i is a vector of observable variables influencing the relative costs and benefits of data gathering activities, β' is a vector of coefficients, and U_i is a stochastic disturbance to be specified below.

Equation (4) determines whether a country is included in the sample used to estimate the cross-country growth equation. If Z_i is positive, country i's government provides reliable, public data, and the country makes it into the sample. If, on the other hand, Z_i is zero or negative, then no data are supplied and the country does not make it into the sample.

The selectivity equation (4) can be incorporated into an econometric analysis that adjusts for sample selection bias in OLS estimates of the growth regression equation [see Greene (2000, chapter 20)]. One simple presentation of the problem is to see the regression equation (1) as part of a two-equation, bivariate classical regression system that includes the original growth equation (1) plus an additional equation specifying the selectivity criteria used to determine the sample in the growth equation, which corresponds to equation (4). Symbolically, the equation system can be described by:

$$y_i = \alpha' X_i + \varepsilon_i \qquad (1)$$

$$Z_i = \beta' V_i + U_i \qquad (4)$$

$$\text{and } \varepsilon_i \text{ , } U_i \sim N [0, 0, \sigma^2_\varepsilon, \sigma^2_u, \rho] \quad , \qquad (2')$$

where σ^2_ε is the variance of the growth regression error term, σ^2_u is the variance of the error term for the equation for Z_i, and ρ is the correlation coefficient between the two error terms, ε_i and $U_{i\,,}$ which is equal to $\rho = \sigma_{\varepsilon\,u} / \sigma_\varepsilon\, \sigma_{u,}$ with $\sigma_{\varepsilon\,u}$ the covariance between the error terms and σ_ε and σ_u the corresponding standard deviations.

The two equations, (1) and (4), could be estimated as a classical bivariate regression model. In reality, the variable Z_i is not generally observed since information on the costs and benefits of information processing in various countries is not readily accessible. However, one can still estimate the system in (1) and (4) by slightly modifying the selectivity equation to make it consistent with the data available. What is clearly known is whether a country has been able to collect the data, which determines whether a country is included in the growth regression sample. We can therefore substitute equation (4) with an alternative selectivity mechanism based on the observed dichotomous variable, S_i, where:

$$\begin{aligned} & S_i = 1 \qquad && \text{if } Z_i > 0, \qquad && (4') \\ \text{and} \quad & S_i = 0 \qquad && \text{otherwise.} && \end{aligned}$$

The probability of being included in the cross-country growth data sample is then:

$$\begin{aligned} P_i &= \Pr [S_i = 1] = \Pr [Z_i > 0] = \Pr [U_i > -\beta' V_i] \\ &= 1 - \Pr [U_i \leq -\beta' V_i] \\ &= 1 - F(-\beta' V_i) \\ &= F(\beta' V_i), \qquad (4'') \end{aligned}$$

where F is a symmetric, cumulative distribution function for U_i .

If U_i is assumed to have a normal distribution, then the probability of country i being in the growth cross-sectional data sample is given by:

$$\Pr [S_i = 1] = \int_{-4}^{\beta' V_i} \frac{1}{(2\pi)^{1/2}} \exp (t^2/2)dt \qquad (4'')$$

This is a probit model whose dependent variable is equal to one if a country is included in the cross-country growth sample and zero otherwise. It can be estimated to determine the impact of various explanatory (selection) variables on the likelihood of inclusion in the cross-country growth regression sample.

Since the probit equation (4") replaces the unobserved equation (4), the modified growth regression model taking into account the sample selection mechanism now consists of the original growth regression equation (1) and the probit selectivity equation (4"). These two equations can be estimated jointly by maximum likelihood.

An alternative specification, following Heckman (1979) is the so-called Heckit procedure. It involves the OLS estimation of equation (3) after including an estimate of E(U_i $*$ X_i and y_i are observed) as an additional variable in the equation. This estimate is obtained by first identifying and estimating the binary probit model in equation (4"), with the dependent variable equal to one if the country is in the sample and zero otherwise. The results are used to compute inverse Mills' ratios that are then introduced into the original regression equation (3)–in place of E(U_i $*$ X_i and y_i are observed)–to take sample selection bias in account. In the second stage of the Heckit estimation procedure, the growth regression equation would have an additional variable, $MILLS_i$, to identify the inverse Mills' ratio computed for each sample country. With $MILLS_i$ included as an independent variable, the cross-country growth equation can be estimated using ordinary least squares, providing consistent estimates [see Heckman (1979) and Maddala (1983, chapter 8)]. The results using the two-equation MLE model are both consistent and efficient, and the Heckit approach generally produces results that are very close to those of the MLE model, so the results of the two-equation MLE model are reported here, leaving the Heckit results for Appendix II.

The key issue at hand is to specify the vector V_i of explanatory variables in the probit equation (4"), establishing the likelihood of being in the growth regression sample. The probit model can then be estimated using a full data set that contains the sample of countries traditionally included in cross-country growth analysis plus an additional sample of countries not used in those studies due to the unavailable data. The next section specifies and examines the selectivity variables and how they influence the probability that a government will carry out the appropriate data gathering activities that would qualify the country to be included in cross-country growth regression analysis.

III. THE DETERMINANTS OF SAMPLE SELECTION IN EMPIRICAL GROWTH ANALYSIS

In this section, we specify the variables that influence whether or nor a country is included in the sample of countries for which cross-country growth regressions are carried out.

Since data gathering activities involve significant set-up costs, governments of poor countries are less likely to make those investments than those in rich countries. Using per-capita GDP available for a wide cross-section of countries in 1990, a set of three dummy variables was constructed: $POOR_i$, that is equal to one if the country has a per-capita GDP (expressed in 1985 international dollars, adjusted for differences in purchasing power) of less than $1,600 and zero otherwise, $MIDDLE_i$, equal to one if the country has a per-capita GDP between $1,600 and $8,600 and zero otherwise, and $RICH_i$, which is equal to one if the country's per-capita GDP was over $8,600 and zero otherwise. These categories follow those of the World Bank's categorization into low-income, middle-income and high-income.

TheWorld Bank (2000a) database yields data on GDP per-capita for 147 countries. This is significantly larger than the sample sizes used in most growth regression analyses, which have samples of less than 100 countries. This is not necessarily due to the unavailability of GDP per-capita (or other data) for the 1990s but because of the lack of reliable data from earlier years. For instance, in many growth regressions, the dependent variable is growth in GDP per-capita between 1960 and 1985 or growth between 1960 and 1990. Although a comprehensive sample can be constructed with GDP per-capita for 1990, such data is not so easily available for 1960.

One suspects that the likelihood that a country has reliable data not only on current but also past values of the relevant variables used in growth regression analysis is related to its level of income. Countries with low per-capita GDP in 1990 may be able to supply data for 1990, but unable to report it for earlier decades.

It is also to be expected that the greater the educational attainment in a country, the more likely its government will be able to collect reliable data. If most of a country's population has not attained a college education, it will lack the skilled manpower required to develop and maintain the statistical databases needed to survey the economy's fundamental economic indicators over time. This includes but is not limited to measuring inflation rates, maintaining a consistent system of national income accounts and balance of payments statistics, surveying of individual households to determine educational attainment, and developing a comprehensive historical set of economic statistics. On the other hand, the higher the proportion of college-educated workers in the population, the more likely the country will have the skilled labor force that is required to supply a reliable system of national statistics. As a proxy for the presence or absence of human expertise on data collection and analysis in a country, we will utilize the variable HIGHERED_i, which is equal to the proportion of persons 25 years of age or older in country i who had attained a tertiary educational level in 1990. As with most educational attainment indicators, this is a variable for which 1990 data have become widely available for developing countries but is exceedingly difficult to obtain for earlier decades.[3]

An important factor influencing the likelihood that a nation will have reliable, public information available is whether the country belongs to the International Monetary Fund (IMF). According to the IMF website: "a long-standing objective [of the IMF] has been the improvement of data and statistics practices among [its]membership." In the analysis below we use a dummy variable, NONIMF_i, equal to one if country i was not a member of the IMF in 1980

[3] This measure includes persons who have completed their tertiary education as well as persons who attended higher education institutions for a certain period of time but did not complete their degrees. The data are obtained by joining UNESCO, Barro-Lee (1994, 2000) and the detailed World Bank Higher Education Task Force data set, as presented in World Bank (2000b).

and zero otherwise. Significantly, most former Soviet Republics and East European economies were not IMF members in 1980. We expect $NONIMF_i$ to be negatively associated with the probability of being included in cross-country growth samples.

Urbanization facilitates the process of data collection. Data gathering activities are significantly more expensive for countries where the population is widely distributed in scattered, isolated rural communities. Indeed, in some countries, census and household surveys are conducted only in urban areas. We expect that a greater rate of urbanization should make it more likely that comprehensive socioeconomic data are available. The variable $URBAN_i$ is defined as the proportion of the 1980 population of country i residing in areas defined as urban.[4] It is anticipated that this variable will be positively related to the probability of a nation being included in growth regression samples.

Political institutions constitute yet another variable influencing the likelihood of data collection. In order for the public to make informed electoral decisions, effective democracies require that the government produce reliable, publicly accessible socioeconomic data that can be used to monitor the performance of the administration in power. Authoritarian states on the other hand, do not face such pressures and, all else being equal, are less likely to supply reliable public data. To measure this factor, we utilize the Freedom House measure of political rights. This measure, which we refer to as the variable $AUTHORITA_i$, is based on Freedom House's 1980 classification of countries on a scale of 1 to 7, with a higher value indicating fewer political rights (greater authoritarianism). [5]

The discussion so far suggests that the probability of a country i being included in the sample of countries used in empirical growth analysis, P_i, is equal to Pr $[\beta^{'} V_i + U_i > 0]$, with:

$$\beta' V_i + U_i = \beta_o + \beta_1 \, POOR_i, + \beta_2 \, MIDDLE_i + \beta_3 \, HIGHERED_i + \beta_4 \, NONIMF_i$$

[4] The source for these data are: World Bank (2000a) and World Bank (1981).

[5] These data is the same as that used in Barro (1991), supplemented with additional data for 1986 from Freedom House (1998), and with values of 7.0 assigned to countries which were under socialism and under the Soviet sphere of influence in 1980.

$$+ \beta_5 \text{ URBAN}_i + \beta_6 \text{ AUTHORITA}_i, + U_i \quad , \tag{5}$$

where the β_m are coefficients to be estimated.

To adjust for selectivity bias, the selection probit equation in (5) will be estimated jointly with the growth regression equation (1) by maximum likelihood (the Heckit model was also estimated and the results reported in Appendix II). The statistical significance of the correlation coefficient between the disturbance terms of the two equations--symbolized by ρ in equation (2')-- reflects whether sample selection bias is a potential problem in the estimation of the growth regression equation.

The implications of our analysis are bounded by the set of variables that have been used to explain sample selection. Any bias found in estimated regression coefficients is related to the sample censoring associated with those variables. Although we believe that the selection variables we have included (income, educational attainment, etc.) are the most relevant, it is possible that we have missed some crucial forces influencing sample selection. In our preliminary analysis, we considered additional variables such as ethnic fractionalization, political instability, size of the country, etc. as possible determinants of sample censoring. However, these variables were either closely correlated with those already considered above or were not available for a large number of countries. Future research may identify additional sample-censoring variables not examined in this paper.

IV. RE-ANALYZING MANKIW-ROMER-WEIL'S "A CONTRIBUTION TO THE EMPIRICS OF ECONOMIC GROWTH"

This section presents an assessment of the sample selection bias issue in the central model estimated by Mankiw-Romer-Weil in their influential paper on the empirics of economic growth. These authors derive theoretically and estimate empirically an augmented Solow model that incorporates the accumulation of human as well as physical capital. The empirical model is given by:

$$y_i = \alpha_o + \alpha_1 \ln(I/GDP)_i + \alpha_2 \ln(n_i + g_i + \delta_i) + \alpha_3 \ln(SCHOOL)_i + \varepsilon_i \quad , \qquad (6)$$

where y_i is the log of income per-capita, measured by real GDP in 1985 divided by the working-age population in that year; I/GDP_i is equal to the average share of real investment in real GDP during the sample period (1960 to 1985); n_i is the rate of growth of the working-age population between 1960 and 1985 (people aged 15 to 64); $(n_i + g_i + \delta_i)$ is the rate of growth of population, the rate of technical change plus the rate of depreciation, constrained by M-R-W to equal 0.05; and $SCHOOL_i$ is a measure of the rate of human capital accumulation, equal to the percentage of the working-age population enrolled in secondary education (average for the period 1965 to 1980).

Note that the M-R-W analysis is considered to be "growth analysis" although the dependent variable in equation (6) is the level of income per-capita, not growth. The reason is that equation (6) is derived from the implications of the Solow growth model, augmented by including human capital. We will thus informally refer to it as a "growth equation."

Mankiw-Romer-Weil used two main samples in their empirical analysis. One sample includes all 98 non-oil countries for which they had available data.[6] A second sample was created by excluding an array of "small" countries as well as countries that Summers and Heston catalogued as having low quality data available. This sample consists of 75 countries.[7]

In our selectivity analysis we will supplement the two samples used by M-R-W to include those censored countries for which information required by the growth regressions is not

[6] Oil countries are excluded on the basis that "the bulk of recorded GDP for these countries represents the extraction of existing resources, not value added [and] one should not expect standard growth models to account for measured GDP in these countries" [Mankiw,Romer and Weil (1992, p. 413)].

[7] M-R-W also carried out their analysis on a sample that included only OECD countries, with the explicit goal of examining the values of the estimated coefficients for this particular group of nations.

available. This augmented sample of countries is used in the selectivity probit equations, where the dependent variable is equal to one if a country i is in the growth regression sample and zero if it is not. The explanatory variables in this selectivity equation were specified in the last section and are summarized in equation (5). They are: POOR, MIDDLE, HIGHERED, NONIMF, URBAN, and AUTHORITA (Appendix I lists variable definitions and sample means). The augmented sample includes 147 countries for which we have data pertaining to the six selectivity variables.[8]

Tables I and II reproduce the results of the Mankiw-Romer-Weil analysis using a sample of 75 countries as well as our re-estimation after taking into account sample selectivity. Table I begins by presenting the results of the selectivity probit equation, which supplements the sample of 75 countries in the growth equation with 72 non-oil countries that were censored by M-R-W.

Note, first of all, that the correlation coefficient between the error terms of the selectivity and growth regression equations, Rho(1,2), is equal to -0.9 and is statistically significant at a 99 per cent level of confidence. This suggests that sample selection bias is a significant issue for the Mankiw-Romer-Weil analysis. What are the most relevant selectivity variables involved in the censoring of countries associated with the M-R-W sample? Table I shows that the level of income of a country is a key force increasing the likelihood that it will be included in the M-R-W's 75-country growth regression sample. The coefficients on the two dummy variables reflecting income level, POOR and MIDDLE, are both statistically significant at a level of confidence of 99%. The estimated coefficients are both negative, which suggests that being poor or middle income makes it less likely for a country to be included in the growth equation sample.

[8] There are still some non-oil countries that are excluded from our own, augmented, sample due to the unavailability of the data required for the selectivity analysis. However, these countries are almost all small countries --whose production structure, as M-R-W observe, may be idiosyncratically determined by non-Solow forces-- or countries whose production structure has been severely distorted by long-term conflict. Therefore, even if data were available, these countries would be exempted from the analysis because the structure of the model being tested is not intended to apply to them. The excluded countries include: Dominica, Afghanistan, Aruba, Antigua, Barbados, Bermuda, Bhutan, Bosnia/Herzegovina, Croatia, Brunei, Cambodia, Cuba, Djibouti, Eritrea, Equatorial Guinea, Grenada, Guadaloupe, Guam, Gaza/West Bank, Lebanon, Macedonia, Maldives, Martinique, Netherlands Antilles, North Korea, Qatar, Reunion, Samoa, Slovenia, St. Kitts/ Nevis, St.Lucia, Suriname, Sao Tome/Principe, Solomon Islands, Tonga, and Vanuatu.

Another variable whose coefficient is statistically significant is NONIMF, which assumes a value equal to one if the country is not a member of the IMF and zero otherwise. As Table I shows, if the country was not a member country of the IMF in 1980, it is less likely to be part of the sample used in M-R-W's growth regressions. This suggests that, in this sample, IMF membership is positively associated with the reliable production of data used in growth regressions.

The coefficients of the other three variables in the selectivity probit equation, HIGHERED, URBAN and AUTHORITY are not statistically significant at conventional levels of confidence.

Having presented the estimated coefficients of the selectivity probit equation, our re-estimation of the Mankiw-Romer-Weil growth regression equation is presented next. The first column in Table II presents the OLS-version of equation (6), as presented by Mankiw-Romer-Weil on page 420 of their paper (and which we were able to reproduce using their data). Column 2 shows the value obtained for a two-tailed test of the hypothesis that the estimated coefficients in the first column are equal to zero (for the OLS regression, this probability is based on the "t" distribution). Column 3 of Table II depicts the estimated coefficients of the "growth equation" examined by Mankiw-Romer-Weil but now estimated jointly with the selectivity probit equation as part of a bivariate regression model using maximum likelihood estimation. Column 4 presents probability values for significance tests on the estimated coefficients.

The results presented in Table II show the impact of adjusting for sample selection bias on the estimated coefficients of the M-R-W model. The coefficients on all explanatory variables change significantly when the growth equation is estimated jointly with the selectivity equation. The coefficient on the human capital accumulation variable ln(SCHOOL), drops from 0.73 in the OLS equation to 0.60 in the selectivity-adjusted equation. The coefficient on the variable $\ln(n+g+\delta)$ also changes sharply, from -1.50 to -0.73, and loses its statistical significance at any conventional level of confidence. The coefficient on the investment rate term is the least affected but it also shrinks significantly from 0.70 to 0.60.

The discussion so far has examined the results of the M-R-W's analysis that uses a sample of 75 countries. They also considered a larger, 98-country sample. We proceed to discuss next the results of our re-estimation of their 98-country growth (per-capita GDP) regression equation taking sample selectivity into account.

Table III presents the estimated coefficients of the selectivity probit equation, that included the 98 countries in the M-R-W sample plus 49 censored countries (Appendix I presents variable definitions and sample means). Note that the correlation coefficient between the error terms of the growth regression and the selectivity equation, Rho(1,2), is again statistically significant at a 99% level of confidence. However, the variables that are associated with selectivity are drastically different from those presented in Table I for M-R-W's sample of 75 countries. The addition of what are 23 mostly low and middle-income countries to the sample eliminates the role of income in sample selection: the coefficients of both the POOR and MIDDLE dummy variables, representing low and middle-income countries in the equation, lose their statistical significance. Of all the variables explaining sample selectivity, only NONIMF is now statistically significant.

Table IV shows the results of our re-estimation of the regression analysis carried out by M-R-W, adjusting for sample selection bias (as before, appendix I presents variable definitions and sample means). The first two columns of Table IV depict the OLS regression estimates as obtained by M-R-W. Columns three and four show the selectivity-adjusted coefficients and associated probabilities. As can be seen, there are differences in the two sets of coefficients, but these are not as substantial as those in the sample of 75 countries. The coefficient on ln(I/GDP) rises from 0.69 to 0.78, the coefficient on $\ln(n+g+\delta)$ drops slightly from -1.74 to -1.79, and the ln(SCHOOL) coefficient again declines, but this time from 0.65 to 0.57.

The estimates presented in this section illustrate how sample selection bias can lead to significant changes in the results of empirical growth analysis. But the differences obtained using the 75 and 98 country samples also suggest that the bias is dependent on the precise sample utilized. In the analysis carried out by M-R-W, using a 75-country sample leads to the exclusion of a number of low and middle-income countries that leads to a substantial sample selection bias. On the other hand, if the 98-country sample is used, this bias diminishes substantially.

V. RE-ANALYZING BARRO'S "ECONOMIC GROWTH IN A CROSS SECTION OF COUNTRIES"

The paper by Barro (1991) focuses on two questions. The first is whether there is convergence in levels of per-capita income across countries, that is, what is the impact of the initial level of income on subsequent economic growth. The second issue is the effect of the initial level of human capital on subsequent economic growth. The empirical analysis in Barro (1991) involved a variety of models including a wide array of variables. We re-estimated all of these models and found similar results across-the-board. For expository purposes, we will focus on the following model:

$$\begin{aligned} GR6085_i = {} & \alpha_o + \alpha_1\, GDP60_i + \alpha_2\, SEC60_i + \alpha_3\, PRI60_i + \alpha_4\, GCY_i \\ & + \alpha_5\, REV_i + \alpha_6\, ASASS_i + \alpha_7\, PPI60DEV_i + \varepsilon_i \end{aligned} \tag{7}$$

where $GR6085_i$ is the average annual growth rate between the years 1960 and 1985 of real GDP per-capita of country i, $GDP60_i$ is the value of real per-capita GDP for country i in 1960 (1980 base year), $SEC60_i$ is the secondary-school enrollment rate in 1960, $PRI60_i$ is the primary-school enrollment rate in 1960, GCY_i is the average ratio of real government consumption (exclusive of defense and education) to real GDP for the period of 1970 to 1985, REV_i is the number of revolutions and coups per year (1960-85 or sub-sample), $ASASS_i$ is the number of assassinations per million population per-year (1960-85 or sub-sample), and $PPI60DEV_i$ is the magnitude of the deviation of the 1960 PPP value for the investment deflator from its sample mean. The rationale for including these variables can be found in Barro (1991). Note however, that the sign and value of the coefficient on $GDP60_i$ is related to the issue of convergence and those on $SEC60_i$ and $PRI60_i$ reflect the impact of initial levels of human capital on subsequent economic growth.

The main sample used by Barro (1991) consists of 98 countries. We utilized Barro's data set to reproduce his results and then re-estimated Barro's regression equation in a model that takes into account sample selection bias, as described in earlier sections. The selectivity probit equation in this model involves a dependent variable that is equal to one if a country i is in the Barro (1991) growth regression sample and zero if it is not. The explanatory variables were specified in Section II and also summarized in equation (5); they are: POOR, MIDDLE, HIGHERED, NONIMF, URBAN, and AUTHORITA (appendix I presents variable definitions and sample means). The augmented sample includes 147 countries.

Tables V and VI reproduce the results of the Barro (1991) growth analysis as well as our re-estimation taking into account sample selectivity. Table V presents the results of the selectivity probit equation, using the sample of 98 countries in the growth regression equation supplemented by 51 countries censored from the analysis by Barro (1991).

Note first of all that the correlation coefficient between the error terms of the selectivity and growth regression equations –Rho(1,2)-- is equal to 0.7 and is statistically significant at a 99 per cent level of confidence. In addition, Table I shows that there are two variables that are statistically significant in determining whether a country was included in the Barro (1991) sample or not: NONIMF, which is a dummy variable equal to one if the country was not a member of the IMF, and AUTHORITA, which is an index of political rights, where greater values are attached to more authoritarian regimes. Both these variables reduce the likelihood of being included in the Barro (1991) sample.

In contrast to the Mankiw-Romer-Weil sample, the Barro sample does not appear to be censored on the basis of income. The coefficients on the POOR and MIDDLE dummy variables that measure whether a country was a low or middle-income country, are not statistically significant. Note that the results for the Barro (1991) sample differ sharply from those obtained in the 75-country M-R-W sample, for which the coefficients on the two dummy variables reflecting income level, POOR and MIDDLE, were both negative and statistically significant at a level of confidence of 99%. These results are closer to the 98-country sample used by M-R-W in which both income dummy variables were statistically insignificant in determining selectivity.

Having presented the results of the estimated probit equation, Table VI displays our re-estimation of the Barro (1991) growth equation (appendix I presents variable definitions and sample means for this equation). The first column shows the OLS-version of equation (7), represented in Barro (1991) as equation (1) in Table 1 on pages 410-12. Column 2 shows the probability that the estimated coefficients in the first column equal zero.[9]

Column 3 of Table VI depicts the estimated coefficients of the growth equation in Barro (1991) but estimated jointly with the selectivity probit equation as part of a bivariate regression model using maximum likelihood estimation. Column 4 presents the probability that each re-estimated coefficient is equal to zero. Table VI does not show any significant differences between the coefficients of the growth equations when sample selection is taken into account

Our re-estimation of the Barro (1991) growth regression model taking into account sample selectivity suggests that the Barro sample is not censored in a way that substantially biases the statistical results of the growth regression equations reported in that paper. The results are robust in this regard.

VI. SAMPLE SELECTION BIAS IN PAULO MAURO'S "CORRUPTION AND GROWTH"

The paper by Paulo Mauro (1995) on "Corruption and Growth," empirically analyzes the links between corruption and other institutional factors on economic growth. The sample of countries utilized is the smallest so far in our discussion: 58 countries. Since one may suspect that the results of ordinary least squares regressions using such a relatively small sample of

[9] We successfully reproduced all of the equations in Barro (1991) and re-estimated them using our regression model with selectivity. The results in Table VI are typical of our overall results and, for brevity, the analysis of these other growth regression equations are not presented here. They are available from the authors by request.

countries are subject to sample selection bias, this section focuses on re-examining the cross-country empirical growth regressions carried out in that paper.

Mauro (1995) estimated a wide range of empirical models, which included or excluded various explanatory variables. We focus on the following empirical growth equation:

$$GR6085_i = \alpha_o + \alpha_1 GDP60_i + \alpha_2 SEC60_i + \alpha_3 POP6085_i + \alpha_4 BUREAU_i + \varepsilon_i \qquad (8)$$

where $GR6085_i$ is the average annual growth rate of real GDP per-capita of country i between 1960 and 1985, $GDP60_i$ is the value of real per-capita GDP for country i in 1960, $SEC60_i$ is the secondary-school enrollment rate in 1960, $POP6085_i$ is the annual rate of growth of population between 1960 and 1985 in country i, and $BUREAU_i$ is the value of a bureaucratic efficiency index, which is negatively associated with corruption. The last variable is based on data constructed from assessments made by analysts from Business International (now in the Economist Intelligence Unit) of conditions prevailing in the country in question. Mauro (1995) combines three types of assessments made by analysts regarding: (1) “the degree to which business transactions involve corruption or questionable payments,” (2) “the efficiency and integrity of the legal environment as it affects business,” and (3) “the regulatory environment foreign firms must face when seeking approvals and permits, affecting the degree to which it represents an obstacle to business.” The resulting index of bureaucratic efficiency ranges from a minimum of 1.89 (lowest bureaucratic efficiency) to a maximum of 10 (highest bureaucratic efficiency).

The main sample used by Mauro (1995) consists of 58 countries. The growth regression equation (8) was re-estimated using a model that takes into account sample selection bias. The selectivity probit equation involves a dependent variable equal to one if a country i is in the Mauro (1995) growth regression sample and zero if it is not. The explanatory variables in this probit equation are: POOR, MIDDLE, HIGHERED, URBAN, and AUTHORITA (Appendix I presents variable definitions and sample means). The augmented sample includes 147 countries for which we found information on these variables.[10]

Tables VII and VIII reproduce Mauro's results and our re-estimation taking into account sample selectivity. Table VII presents first the results of the selectivity probit equation, using the sample of 58 countries in the growth regression equation supplemented by 89 countries censored from the analysis. The results shown in Table VII suggest that POOR and MIDDLE are statistically significant variables reducing the likelihood of inclusion of a country in the sample. Another variable explaining the pattern of selection into the sample is AUTHORITA, implying that countries with lower values for the political rights index (more authoritarian) also have a lower likelihood of inclusion in the sample.

The value of Rho (1,2), the correlation coefficient between the error terms of the selection probit equation and the growth regression equation is -0.73 and is statistically significant at a 98% level of confidence.

Table VIII presents one of the main equations estimated by Mauro (Table VII, pp. 702-3, equation 5).[11] The first column shows the OLS-version of equation (8). The second column presents the probabilities obtained from a two-tailed test of the hypothesis that the estimated coefficients in the first column each equal zero. Column 3 depicts the estimated coefficients of the growth equation estimated jointly with the selectivity probit equation as part of a bivariate regression model using maximum likelihood estimation. Column 4 presents the probability that each coefficient is equal to zero. Table VIII shows that the coefficients obtained with the selectivity adjustment differ from the OLS coefficients, but the changes are not substantial. In particular, the coefficient on the BUREAU variable, representing an index of bureaucratic efficiency, drops from 0.006 to 0.005, not a major adjustment.

[10] The NONIMF dummy variable had to be dropped because of the small number of "1"s in the Mauro (1995) sample. As Greene (1998, pp. 444-445) notes, the probit estimator tends to break down when an explanatory dummy variable is extremely unbalanced in terms of "1"s or "0"s, particularly in small samples.

[11] Using the same data as Mauro (1995), we closely reproduced his results and re-estimated them using our bivariate selection model. The model was applied not only to the equation reported in Table VIII but also to the other major growth equations reported in that paper. Since these results do not differ much from those in Table VIII, we do not report them here but will supply them upon request.

Our re-estimation of the Mauro (1995) growth regression model taking into account sample selectivity suggests that its sample, though relatively small, is not censored in a way that significantly affects its results.

VII. CONCLUSIONS

In the presence of sample selection bias, the ordinary least squares coefficients (OLS) obtained from cross-country growth regression equations may be biased. An alternative econometric model consists of a two-equation system with the growth regression equation estimated jointly with a selectivity equation that specifies the influence of a set of variables on the probability of selection into the growth regression sample. This model can be estimated by maximum likelihood and provides consistent and efficient coefficients in cross-country growth regression equations. This paper has adopted this bivariate selectivity methodology and presents a re-examination of three major contributions to the empirical cross-country growth literature published in the *Quarterly Journal of Economics*.

The paper first identified the variables that influence whether reliable cross-country growth data are available or not. Since most cross-country data are aggregate in nature, they are usually produced by government statistical agencies. Therefore, the likelihood of having reliable data for any particular country is related to: (1) level of income, with governments in poor countries less likely to make the necessary investments in data collection and processing , (2) educational attainment, with countries having more educated populations more likely to produce reliable data, (3) IMF membership, with countries belonging to the IMF more likely to produce transparent data on a regular basis, (4) urbanization, with more urbanized countries having a greater likelihood of providing country-wide data, and (5) political institutions, with democracies more likely to produce reliable data than authoritarian governments.

With this sample selectivity model in hand, we first reproduced the OLS results of the three papers examined –Mankiw-Romer-Weil (1992), Barro (1991)and Mauro (1995) – and then re-estimated the cross-country growth equations using the bivariate selectivity model. Our analysis suggests that sample selection bias could lead to significant changes in the results of empirical growth analysis depending on the specific sample utilized. In the Mankiw-Romer-Weil (1997) paper, we found that using their 75-country sample leads to the exclusion of a number of low-income and middle-income countries that results in a substantial sample selection bias. The value and statistical significance of the estimated growth equation coefficients reported by Mankiw-Romer-Weil for this sample of countries change drastically when adjusted for sample selectivity. But in re-examining these results using Mankiw-Romer-Weil's 98-country sample, we found much smaller differences in estimated coefficients. The impact of sample selection bias on the Mankiw-Romer-Weil results is thus dependent on the choice of sample.

We also found that the OLS cross-country growth equation coefficients in Barro (1991) are almost identical to those obtained using the bivariate selectivity equation. For this paper, countries with low levels of income do not appear to be over-represented in the group of censored countries. In fact, most of the variables included to explain sample selection are not statistically significant.

The comparatively small sample used by Mauro (1995) –58 countries– could be anticipated by some to signal that sample selection bias is an issue in that paper. However, our results show that sample selection bias is not necessarily associated with small sample size. For Mauro's paper, we found the selectivity adjustments changed the growth regression coefficients, but not to any significant extent. This confirms the danger of rejecting analyses that employ small samples just based on the fear that the analysis may be subject to sample selection bias. Our analysis finds such fears are justified for the 75-country sample used by Mankiw-Romer-Weil (1992) but unjustified for Mauro(1995)'s 58-country sample.

The diversity of our results leads us to conclude that future researchers in this field should consider incorporating the methodology used in this paper to examine the presence of sample selection bias in their analysis. Although this methodology has its limitations, it can provide substantial value added as a tool to explore the robustness of growth regression estimates.

TABLE 1

SELECTION PROBIT EQUATION, MANKIW-ROMER-WEIL MODEL

SAMPLE OF 75 SAMPLE COUNTRIES PLUS 72 CENSORED COUNTRIES

Dependent variable: $S_{i,}$ equal to one if the country is in the growth regression sample and zero otherwise.

Estimation technique	Bivariate regression model with selection
Growth equation sample	Mankiw-Romer-Weil Non-oil countries (75 countries)
Selection equation sample	147 countries

Selection equation Explanatory variable	Estimated coefficient (s.e.)	Prob. ∗Z∗ ∃0 (Z)
CONSTANT	0.9768 (0.9152)	0.3 (1.1)
POOR	-1.6029 (0.7678)	0.0 (-2.1)
MIDDLE	-1.1506 (0.5188)	0.0 (-2.2)
HIGHERED	-0.6087 (4.0170)	0.9 (-0.1)
NONIMF	-1.4583 (0.5906)	0.0 (-2.5)
URBAN	0.0103 (0.0096)	0.3 (1.1)
AUTHORITA	-0.0124 (0.1002)	0.9 (-0.1)
Rho (1,2)	-0.9 (0.1)	0.0 (-7.5)
Log-Likelihood	-109.6	

Note: Standard errors for the estimated coefficients are in parentheses, below the reported value of the coefficients; the second column of each equation presents the value of the probability obtained for a two tailed test of the hypothesis that the estimated coefficient equals zero, based on the standard normal distribution.

TABLE II

RE-ESTIMATION OF MANKIW-ROMER-WEIL'S CROSS-COUNTRY GROWTH EQUATION

SAMPLE OF 75 COUNTRIES

Dependent variable: log GDP per working-age person in 1985

Estimation technique	Ordinary least squares		Bivariate regression model with selection	
Growth equation sample	Mankiw-Romer-Weil Non-oil countries (75 countries)		Mankiw-Romer-Weil Non-oil countries (75 countries)	
Growth Equation Explanatory variable	Estimated coefficient (s.e.)	Prob. ∗t∗∃x (t)	Estimated coefficient (s.e.)	Prob ∗Z̦∗∃z (Z)
CONSTANT	7.81 (1.19)	0.0 (6.5)	9.50 (1.86)	0.0 (5.2)
ln (I/GDP)	0.70 (0.13)	0.0 (4.7)	0.60 (0.15)	0.0 (3.9)
$\ln (n + g + \delta)$	-1.50 (0.42)	0.0 (-3.7)	-0.73 (0.63)	0.2 (-1.2)
ln(SCHOOL)	0.73 (0.07)	0.0 (7.7)	0.60 (0.08)	0.0 (5.1)
$\bar{R}^2$	0.78		–	
Log-Likelihood	--		-109.6	

Note 1: Standard errors are in parentheses, below the estimated coefficients; the value of the t statistic is in parentheses, below the value obtained for a two tailed test of the hypothesis that the estimated coefficient equals zero (for the OLS regression, the probability is based on the "t" distribution, while for the selection-adjusted equations, the probability is based on the standard normal distribution).

Note 2: The OLS equation reports the R^2 but this cannot be computed for the selectivity-adjusted equation, which is estimated jointly by Maximum Likelihood. Instead, for the latter, we report the likelihood ratio statistic for the joint probit-regression model.

TABLE III

SELECTION PROBIT EQUATION, MANKIW-ROMER-WEIL MODEL

SAMPLE OF 98 SAMPLE COUNTRIES PLUS 49 CENSORED COUNTRIES

Dependent variable: $S_{i,}$ equal to one if the country is in the growth regression sample and zero otherwise.

Estimation technique	Bivariate regression model with selection
Growth equation sample	Mankiw-Romer-Weil Non-oil countries (98 countries)
Selection equation sample	147 countries

Selection equation Explanatory variable	Estimated coefficient (s.e.)	Prob. *Z* ∃0 (Z)
CONSTANT	0.5078 (1.0910)	0.6 (0.5)
POOR	1.3226 (0.9623)	0.2 (1.3)
MIDDLE	-0.0062 (0.6548)	0.9 (-0.1)
HIGHERED	3.9029 (5.6790)	0.5 (0.7)
NONIMF	-2.5247 (0.4309)	0.0 (-5.7)
URBAN	-0.0016 (0.0119)	0.9 (-0.1)
AUTHORITA	-0.0162 (0.0811)	0.8 (-0.2)
Rho (1,2)	0.875 (0.100)	0.0 (8.7)
Log-Likelihood	-119.5	

Note: Standard errors are in parentheses, below the estimated coefficients; the second column of each equation presents the value of the probability obtained for a two tailed test of the hypothesis that the estimated coefficient equals zero, based on the standard normal distribution.

TABLE IV

RE-ESTIMATION OF MANKIW-ROMER-WEIL CROSS-COUNTRY GROWTH EQUATION

SAMPLE OF 98 COUNTRIES

Dependent variable: log GDP per working-age person in 1985

Estimation technique	Ordinary least squares		Bivariate regression model with selection	
Growth equation sample	Mankiw-Romer-Weil Non-oil countries (98 countries)		Mankiw-Romer-Weil Non-oil countries (98 countries)	
Growth Equation Explanatory variable	Estimated coefficient (s.e.)	Prob. *t*∃x (t)	Estimated coefficient (s.e.)	Prob *Ẓ*∃z (Z)
CONSTANT	6.85 (1.18)	0.0 (5.8)	6.44 (1.45)	0.0 (4.4)
ln (I/GDP)	0.69 (0.13)	0.0 (5.2)	0.78 (0.12)	0.0 (6.6)
ln (n + g + δ̣)	-1.74 (0.42)	0.0 (-4.2)	-1.79 (0.51)	0.0 (-3.5)
ln(SCHOOL)	0.65 (0.07)	0.0 (9.0)	0.57 (0.08)	0.0 (6.8)
$\bar{R}^2$	0.77		–	
Log-Likelihood		--		-119.5

Note 1: Standard errors are in parentheses, below the estimated coefficients; the value of the t statistic is in parentheses, below the value obtained for a two tailed test of the hypothesis that the estimated coefficient equals zero (for the OLS regression, the probability is based on the "t" distribution, while for the selection-adjusted equations, the probability is based on the standard normal distribution).

Note 2: The OLS equation reports the R^2 but this cannot be computed for the selectivity-adjusted equation, which is estimated jointly by Maximum Likelihood. Instead, for the latter, we report the likelihood ratio statistic for the joint probit-regression model.

TABLE V

SELECTION PROBIT EQUATION, BARRO MODEL
SAMPLE OF 98 SAMPLE COUNTRIES PLUS 49 CENSORED COUNTRIES

Dependent variable: $S_{i,}$ equal to one if the country is in the growth regression sample and zero otherwise.

Estimation technique	Bivariate regression model with selection	
Growth equation sample	Barro’s 98 countries	
Selection equation sample	147 countries	
Selection equation Explanatory variable	Estimated coefficient (s.e.)	Prob. ∗Z∗ ∃0 (Z)
CONSTANT	1.70	0.2
	(1.34)	(1.3)
POOR	0.92	0.4
	(1.21)	(0.8)
MIDDLE	0.53	0.5
	(0.80)	(0.7)
HIGHERED	-2.04	0.7
	(5.96)	(-0.3)
NONIMF	-2.03	0.0
	(0.51)	(-3.9)
URBAN	0.01	0.5
	(0.01)	(0.7)
AUTHORITA	-0.36	0.0
	(0.10)	(-3.6)
Rho (1,2)	0.7	0.0
	(0.3)	(2.7)

Note: Standard errors are in parentheses, below the estimated coefficients; the second column of each equation presents the value of the probability obtained for a two tailed test of the hypothesis that the estimated coefficient equals zero, based on the standard normal distribution.

TABLE VI

RE-ESTIMATION OF BARRO CROSS-COUNTRY GROWTH EQUATION
SAMPLE OF 98 COUNTRIES

Dependent variable: Growth of per-capita GDP between 1960 and 1985

Estimation technique	Ordinary least squares		Bivariate regression with selection	
Growth equation sample	Barro (98 countries)		Barro (98 countries)	
Growth Equation Explanatory variable	Estimated coefficient (s.e.)	Prob. *t*∃x (t)	Estimated coefficient (s.e.)	Prob *Z̦*∃z (Z)
CONSTANT	0.03 (0.007)	0.0 (4.7)	0.02 (0.007)	0.0 (3.3)
GDP60	-0.0075 (0.0012)	0.0 (-6.0)	-0.0071 (0.0018)	0.0 (-4.0)
SEC60	0.031 (0.01)	0.0 (2.8)	0.032 (0.02)	0.04 (2.0)
PRIM60	0.025 (0.006)	0.0 (3.9)	0.027 (0.007)	0.0 (3.6)
GCY	-0.12 (0.03)	0.0 (-4.3)	-0.12 (0.03)	0.0 (-4.1)
REV	-0.019 (0.007)	0.0 (-2.9)	-0.018 (0.007)	0.01 (-2.5)
ASSASS	-0.036 (0.018)	0.04 (-2.0)	-0.037 (0.017)	0.03 (-2.1)
PPI60DEV	-0.014 (0.005)	0.0 (-2.6)	-0.014 (0.005)	0.0 (-2.8)
Adjusted R^2	0.52		–	
Log-Likelihood	–		-242	

Note: Standard errors are in parentheses, below the estimated coefficients; the value of the t statistic is in parentheses, below the value obtained for a two tailed test of the hypothesis that the estimated coefficient equals zero (for the OLS regression, the probability is based on the "t" distribution, while for the selection-adjusted equations, the probability is based on the standard normal distribution).

TABLE VII

SELECTION PROBIT EQUATION, MAURO MODEL
SAMPLE OF 58 SAMPLE COUNTRIES PLUS 89 CENSORED COUNTRIES

Dependent variable: $S_{i,}$ equal to one if the country is in the growth regression sample and zero otherwise.

Estimation technique	Bivariate regression model with selection	
Growth equation sample	Mauro’s 58 countries	
Selection equation sample	147 countries	
Selection equation Explanatory variable	Estimated coefficient (s.e.)	Prob. ∗Z∗ ∃0 (Z)
CONSTANT	1.1973 (0.7149)	0.09 (1.7)
POOR	-1.3670 (0.6434)	0.03 (-2.1)
MIDDLE	-0.8289 (0.4454)	0.06 (-1.9)
HIGHERED	1.8085 (3.726)	0.63 (0.5)
URBAN	-0.0012 (0.0087)	0.89 (-0.1)
AUTHORITA	-0.1554 (0.0828)	0.06 (-1.9)
Rho (1,2)	-0.73 (0.32)	0.02 (-2.29)

Note: Standard errors are in parentheses, below the estimated coefficients; the second column of each equation presents the value of the probability obtained for a two tailed test of the hypothesis that the estimated coefficient equals zero, based on the standard normal distribution.

TABLE VIII

RE-ESTIMATION OF MAURO (1995) CROSS-COUNTRY GROWTH EQUATION

SAMPLE OF 58 COUNTRIES

Dependent variable: Growth of per-capita GDP between 1960 and 1995

Estimation technique	Ordinary least squares		Bivariate regression model with selection	
Growth equation sample	Mauro (58 countries)		Mauro (58 countries)	
Growth Equation Explanatory variable	Estimated coefficient (s.e.)	Prob. ∗t∗∃x (t)	Estimated coefficient (s.e.)	Prob ∗Ẓ∗∃z (Z)
CONSTANT	0.010 (0.010)	0.3 (1.0)	0.025 (0.018)	0.2 (1.3)
GDP60	-0.009 (0.002)	0.0 (-4.6)	-0.010 (0.002)	0.0 (-4.2)
SEC60	0.015 (0.017)	0.3 (0.9)	0.013 (0.027)	0.6 (0.5)
POP6085	-0.621 (0.303)	0.1 (-2.1)	-0.459 (0.459)	0.3 (-1.0)
BUREAU	0.006 (0.001)	0.0 (4.2)	0.005 (0.001)	0.0 (4.1)
Adjusted R^2	0.33		–	
Log-Likelihood	–		-90.6	

Note: Standard errors are in parentheses, below the estimated coefficients; the value of the t statistic is in parentheses, below the value obtained for a two tailed test of the hypothesis that the estimated coefficient equals zero (for the OLS regression, the probability is based on the “t” distribution, while for the selection-adjusted equations, the probability is based on the standard normal distribution).

REFERENCES

Barro, Robert J., "Economic Growth in a Cross-Section of Countries,"*Quarterly Journal of Economics,* Vol. 106, No. 2, 1991, 407-43.

Barro, Robert J., *Determinants of Economic Growth: A Cross-Country Empirical Study*, The MIT Press, Cambridge, 1997.

Barro, Robert J. and Jong-Wha Lee, "Data Set for a Panel of 138 Countries," mimeo., Harvard University, Cambridge, Massachusetts, January 1994.

Barro, Robert J. and Jong-Wha Lee, "International Data on Educational Attainment: Updates and Implications," Center for International Development, Harvard University, Cambridge, Massachusetts, April 2000.

Bils, Mark and Peter J. Klenow, "Does Schooling Cause Growth?," *American Economic Review*, Vol. 90, No. 5, 1160-1183, December 2000.

Durlauf, Steven N., "Econometric Analysis and the Study of Economic growth: A Skeptical Perspective," in R. Backhouse and A. Salanti, eds., *Macroeconomics and the Real World*, Oxford University Press, Oxford, 2000.

Easterly, William and Ross Levine, "Africa's Growth Tragedy: Policies and Ethnic Divisions," *Quarterly Journal of Economics*, Vol. 112, No. 4, 1997, 1203-50.

Freedom House, *Freedom in the World: The Annual Survey of Political Rights and Civil Liberties, 1996-1997*, Transaction Publishers, New Brunswick, 1998.

Greene, William H. *Econometric Analysis*, Prentice Hall Publishing Company, Upper Saddle River, New Jersey, 2000.

Greene, William H., *LIMDEP Version 7.0 User's Manual*, Econometric Software, Inc., Plainview, New York, 1998.

Hanushek, Eric A. and Dennis D. Kimko, "Schooling, Labor-Force Quality, and the Growth of Nations," *American Economic Review*, Vol. 90, No. 5, December 2000, 1184-1208.

Heckman, James, "Sample Selection Bias as a Specification Error," *Econometrica*, Vol. 47, 1979, 153-161.

Knack, Stephen and Philip Keefer, "Does Social Capital Have an Economic Payoff?: A Cross-Country Investigation," *Quarterly Journal of Economics*, Vol. 112, No. 4, December 1997, 1250-1288.

Krueger, Alan B. and Mikael Lindahl, "Education for Growth in Sweden and the World," National Bureau of Economic Research Working Paper No. 7190, Cambridge, September 1997.

Levine, Ross and D. Renelt, “A Sensitivity Analysis of Cross-Country Growth Regressions,” *American Economic Review*, Vol. 82, No. 4, September 1992, 942-963.

Levine, Ross and Sara Zervos, “Stock Markets, Banks and Economic Growth,” *American Economic Review*, Vol. 88, No. 3, June 1998, 537-558.

Maddala, G., *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, New York, 1983.

Mankiw, N. Gregory, David Romer and David N. Weil, “A Contribution to the Empirics of Economic Growth,” *Quarterly Journal of Economics*, Vol. 107, No. 2, 1992, 407-37.

Mauro, Paolo, “Corruption and Growth,” *Quarterly Journal of Economics*, Vol. 110, No. 3, 1995, 681-712.

Pritchett, Lant, "Where has All the Education Gone?," Policy Research Working Paper No. 1581, The World Bank, Washington, D.C., March 1996.

Rodrik, Dani, “Where did all the Growth Go?: External Shocks, Social Conflict and Growth Collapses,” National Bureau of Economic Research Working Paper No. 6350, January1998.

Sachs, Jeffrey and Andrew M. Warner, “Sources of Slow Growth in African Economies,” *Journal of African Economies,* December 1997, Vol. 6, No. 3, October 1997, 335-376.

Temple, Jonathan, “The New Growth Evidence,” *Journal of Economic Literature*, Vol. 37, No. 1, March 1999, 112-156.

World Bank, *World Development Report*, Oxford University Press, New York, 1981.

World Bank, *World Development Indicators 2000*, The World Bank, Washington, D.C., 2000(a).

World Bank, *Higher Education in Developing Countries: Peril and Promise*, The World Bank, Washington, D.C., 2000 (b).

APPENDIX I

VARIABLE DEFINITIONS AND SAMPLE MEANS:

TABLE AI-I

PROBIT SELECTIVITY EQUATION SAMPLE MEANS
MANKIW-ROMER-WEIL SAMPLES OF 75 AND 98 COUNTRIES

Variable		Sample Means	
	Overall Sample of 147 Countries	M-R-W Sample of 75 Countries	M-R-W Sample of 98 Countries
POOR (Dummy variable equal to 1 if the country has per-capita GDP of less than $1,600 in 1990 and zero otherwise)	0.358	0.266	0.418
MIDDLE (Dummy variable equal to 1 if the country has per-capita GDP between $1,600 and $8,600 in 1990 and zero otherwise)	0.455	0.413	0.337
HIGHERED (Proportion of the population who has some college education)	0.053	0.066	0.053
NONIMF (Dummy variable equal to 1 if the country was a member of the IMF in 1980 and zero otherwise)	0.186	0.013	0.031
URBAN (Percentage of the population residing in urban areas in 1980)	45.7	51.8	45.0
AUTHORITA (Value of Freedom House's classification of countries for 1980, on a scale of 1 to 7, with a lower value indicating greater political rights)	4.5	3.5	4.1

Source: Mankiw-Romer-Weil (1997) supplemented by data from Barro-Lee (1994, 2000), Freedom House (1998), IMF (1980) and World Bank (2000a,b).

TABLE AI-II

PROBIT SELECTIVITY EQUATION SAMPLE MEANS

BARRO SAMPLE OF 98 COUNTRIES

Variable	Sample Means	
	Overall Sample of 147 Countries	Barro Sample of 98 Countries
POOR (Dummy variable equal to 1 if the country has per-capita GDP of less than $1,600 in 1990 and zero otherwise)	0.358	0.316
MIDDLE (Dummy variable equal to 1 if the country has per-capita GDP between $1,600 and $8,600 in 1990 and zero otherwise)	0.455	0.418
HIGHERED (Proportion of the population who has some college education)	0.053	0.057
NONIMF (Dummy variable equal to 1 if the country was a member of the IMF in 1980 and zero otherwise)	0.186	0.031
URBAN (Percentage of the population residing in urban areas in 1980)	45.7	61.7
AUTHORITA (Value of Freedom House's classification of countries for 1980, on a scale of 1 to 7, with a lower value indicating greater political rights)	4.5	3.7

Source: Barro (1991) supplemented with data from Barro-Lee (1994,2000), Freedom House (1998), IMF (1980) and World Bank (2000a,b).

TABLE AI-III

PROBIT SELECTIVITY EQUATION SAMPLE MEANS

MAURO SAMPLE OF 58 COUNTRIES

Variable	Sample Means	
	Overall Sample of 147 Countries	Mauro Sample of 58 Countries
POOR (Dummy variable equal to 1 if the country has per-capita GDP of less than $1,600 in 1990 and zero otherwise)	0.358	0.190
MIDDLE (Dummy variable equal to 1 if the country has per-capita GDP between $1,600 and $8,600 in 1990 and zero otherwise)	0.455	0.414
HIGHERED (Proportion of the population who has some college education)	0.053	0.081
NONIMF (Dummy variable equal to 1 if the country was a member of the IMF in 1980 and zero otherwise)	0.186	0.034
URBAN (Percentage of the population residing in urban areas in 1980)	45.7	56.8
AUTHORITA (Value of Freedom House's classification of countries for 1980, on a scale of 1 to 7, with a lower value indicating greater political rights)	4.5	3.3

Source: Barro (1991) supplemented with data from Barro-Lee (1994,2000), Freedom House (1998), IMF (1980) and World Bank (2000a,b).

TABLE AI-IV

GROWTH REGRESSION EQUATIONS, SAMPLE MEANS

Variable	Sample Mean (Standard deviation)			
Mankiw-Romer-Weil				
	75 countries		**98 countries**	
y (log of real GDP in 1985 divided by the working-age population)	0.084	(0.09)	0.080	(0.10)
ln (I/GDP) (average share of real investment in real GDP from 1960 to 1985)	-1.73	(0.44)	-1.85	(0.51)
ln (n + g +.) (rate of growth of working-age population plus rate of tech. change plus rate of depreciation)	-2.64	(0.14)	-2.64	(0.13)
ln(SCHOOL) (average percentage of the working age population enrolled in secondary education between 1960 and 1985)	-2.94	(0.70)	-3.23	(0.91)
Barro (98 observations)				
GR6085 (average annual growth rate of real GDP per-capita from 1960 to 1985)	0.022	(0.018)		
GDP60 (real per-capita GDP in 1960, in thousands of 1980 $)	1.92	(1.81)		
SEC60 (secondary school enrollment rate in 1960)	0.23	(0.21)		
PRIM60 (primary school enrollment rate in 1960)	0.78	(0.31)		
GCY (average ratio of real government consumption to real GDP for 1970-1985)	0.11	(0.05)		
REV (number of revolutions and coups per year, 1960-1985)	0.18	(0.23)		
ASSASS (number of assassinations per million population per year)	0.03	(0.07)		
PPI60DEV (deviation of 1960 PPP investment deflator from the sample mean)	0.23	(0.25)		

TABLE AI-4, CONT.

GROWTH REGRESSION EQUATIONS, SAMPLE MEANS

Variable	Sample Mean (Standard deviation)	
Mauro (58 observations)		
GR6085 (average real per-capita GDP growth rate between1960 and 1985)	0.0025	(0.017)
GDP60 (real per-capita GDP in 1960, in thousands of 1980 US $)	2.37	(01.92)
SEC60 (Secondary school enrollment rate in 1960)	0.30	(0.23)
POP6085 (Population growth rate Between 1960 and 1985)	0.018	(0.010)
BUREAU (Bureaucratic efficiency index)	6.90	(2.16)

Sources: Mankiw-Romer-Weil (1992), Barro (1991), and Mauro (1995).

APPENDIX II

TABLE AII - I

HECKIT, TWO-STAGE MODEL

RE-ESTIMATION OF MANKIW-ROMER-WEIL CROSS-COUNTRY GROWTH EQUATION

SAMPLE OF 75 COUNTRIES

Dependent variable: log GDP per working-age person in 1985

Estimation technique	OLS		Heckman’s Two-Stage Selection Model	
Growth equation	Mankiw-Romer-Weil		Mankiw-Romer-Weil	
sample	Non-oil countries		Non-oil countries	
	(75 countries)		(75 countries)	
Growth Equation	Estimated	Prob.	Estimated	Prob
Explanatory variable	coefficient	*t*∃x	coefficient	*t*∃x
	(s.e.)	(t)	(s.e.)	(t)
CONSTANT	7.81	0.0	8.44	0.0
	(1.19)	(6.5)	(1.28)	(6.7)
ln (I/GDP)	0.70	0.0	0.63	0.0
	(0.13)	(4.7)	(0.14)	(4.6)
ln (n + g + δ)	-1.50	0.0	-1.19	0.2
	(0.42)	(-3.7)	(0.45)	(-2.7)
ln(SCHOOL)	0.73	0.0	0.61	0.0
	(0.07)	(7.7)	(0.10)	(6.3)
MILLS	–	–	-0.53	0.0
			(0.24)	(-2.2)
ADJUSTED R-SQ	0.78		0.79	

Note: Standard errors are in parentheses, below the estimated coefficients; the value of the t statistic is in parentheses, below the value obtained for a two tailed test of the hypothesis that the estimated coefficient equals zero.

TABLE AII - II

HECKIT TWO-STAGE MODEL

RE-ESTIMATION OF MANKIW-ROMER-WEIL CROSS-COUNTRY GROWTH EQUATION

SAMPLE OF 98 COUNTRIES

Dependent variable: log GDP per working-age person in 1985

Estimation technique	Ordinary least squares		Heckman's Two-Stage Selection Model	
Growth equation sample	Mankiw-Romer-Weil Non-oil countries (98 countries)		Mankiw-Romer-Weil Non-oil countries (98 countries)	
Growth Equation Explanatory variable	Estimated coefficient (s.e.)	Prob. *t*∃x (t)	Estimated coefficient (s.e.)	Prob *t*∃x (t)
CONSTANT	6.85 (1.18)	0.0 (5.8)	5.86 (1.20)	0.0 (4.9)
ln (I/GDP)	0.69 (0.13)	0.0 (5.2)	0.74 (0.13)	0.0 (5.7)
ln (n + g +.)	-1.74 (0.42)	0.0 (-4.2)	-2.03 (0.42)	0.0 (-4.8)
ln(SCHOOL)	0.65 (0.07)	0.0 (9.0)	0.62 (0.07)	0.0 (8.3)
MILLS	–	–	0.54 (0.18)	0.0 (3.1)
ADJUSTED R-SQ	0.78		0.79	

Note: Standard errors are in parentheses, below the estimated coefficients; the value of the t statistic is in parentheses, below the value obtained for a two tailed test of the hypothesis that the estimated coefficient equals zero.

TABLE AII - III

HECKIT TWO-STAGE MODEL
RE-ESTIMATION OF BARRO CROSS-COUNTRY GROWTH EQUATION
SAMPLE OF 98 COUNTRIES

Dependent variable: Growth of per-capita GDP between 1960 and 1985

Estimation technique	Ordinary least squares		Heckman's Two-Stage Selection Model	
Growth equation sample	Barro (98 countries)		Barro (98 countries)	
Growth Equation Explanatory variable	Estimated coefficient (s.e.)	Prob. *t*∃x (t)	Estimated coefficient (s.e.)	Prob *t*∃x (t)
CONSTANT	0.03	0.0	0.03	0.0
	(0.007)	(4.7)	(0.007)	(3.8)
GDP60	-0.0075	0.0	-0.0073	0.0
	(0.0012)	(-6.0)	(0.0012)	(-6.0)
SEC60	0.031	0.0	0.031	0.0
	(0.01)	(2.8)	(0.01)	(2.9)
PRIM60	0.025	0.0	0.026	0.0
	(0.006)	(3.9)	(0.006)	(4.4)
GCY	-0.12	0.0	-0.12	0.0
	(0.03)	(-4.3)	(0.03)	(-4.4)
REV	-0.019	0.0	-0.020	0.0
	(0.007)	(-2.9)	(0.006)	(-3.1)
ASSASS	-0.036	0.04	-0.035	0.04
	(0.018)	(-2.0)	(0.017)	(-2.0)
PPI60DEV	-0.014	0.0	-0.014	0.0
	(0.005)	(-2.6)	(0.005)	(-2.7)
MILLS	–	–	0.006	0.24
			(0.005)	(1.2)
ADJUSTED R-SQ.	0.52		0.52	

Note: Standard errors are in parentheses, below the estimated coefficients; the value of the t statistic is in parentheses, below the value obtained for a two-tailed test of the hypothesis that the estimated coefficient equals zero.

TABLE AII - IV

HECKIT TWO-STAGE MODEL
RE-ESTIMATION OF MAURO CROSS-COUNTRY GROWTH EQUATION
SAMPLE OF 58 COUNTRIES

Dependent variable: Growth of per-capita GDP between 1960 and 1995

Estimation technique	Ordinary least squares		Heckman’s Two-Stage Selection Model	
Growth equation	Mauro		Mauro	
sample	(58 countries)		(58 countries)	
Growth Equation	Estimated	Prob.	Estimated	Prob
Explanatory variable	coefficient	*t*∃x	coefficient	*Z̧*∃z
	(s.e.)	(t)	(s.e.)	(Z)
CONSTANT	0.010	0.3	0.040	0.0
	(0.010)	(1.0)	(0.017)	(2.3)
GDP60	-0.009	0.0	-0.010	0.0
	(0.002)	(-4.6)	(0.002)	(-4.6)
SEC60	0.015	0.3	0.0001	0.9
	(0.017)	(0.9)	(0.021)	(0.01)
POP6085	-0.621	0.05	-0.482	0.2
	(0.303)	(-2.1)	(0.355)	(-1.4)
BUREAU	0.006	0.0	0.005	0.0
	(0.001)	(4.2)	(0.001)	(3.5)
MILLS	–	–	-0.021	0.0
			(0.010)	(-2.20)
ADJUSTED R-SQ	0.33		0.38	

Note: Standard errors are in parentheses, below the estimated coefficients; the value of the t statistic is in parentheses, below the value obtained for a two-tailed test of the hypothesis that the estimated coefficient equals zero.