ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Deckers, Thomas; Hanck, Christoph

Conference Paper

Multiple Testing in Growth Econometrics

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2010: Ökonomie der Familie - Session: Analysing Macroeconomic Panel Data Sets, No. B2-V3

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Deckers, Thomas; Hanck, Christoph (2010) : Multiple Testing in Growth Econometrics, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2010: Ökonomie der Familie - Session: Analysing Macroeconomic Panel Data Sets, No. B2-V3, Verein für Socialpolitik, Frankfurt a. M.

This Version is available at:
https://hdl.handle.net/10419/37534

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Multiple Testing in Growth Econometrics*

Thomas Deckers† Universität Bonn

Christoph Hanck‡ Rijksuniversiteit Groningen

February 19, 2010

Abstract

This paper discusses two longstanding questions in growth econometrics which involve multiple hypothesis testing. In cross sectional GDP growth regressions many variables are simultaneously tested for significance. Similarly, when investigating pairwise convergence of output in panel data sets of n countries, $n(n-1)/2$ tests are performed. We propose to control the false discovery rate (FDR) so as not to erroneously declare variables significant in these multiple testing situations only because of the large number of tests performed. Doing so, we provide a simple new way to robustly select variables in economic growth models. We find that few other variables beyond the initial GDP level are needed to explain growth. We also show that convergence in panels of per capita output using a time series definition with the necessary condition of no unit root in the log per-capita output gap of two economies does not appear to hold.

Keywords: Growth Empirics, Panel Data, Multiple Testing, Convergence, Bootstrap
JEL-Codes: O47, C12

*We thank Franz Palm, Stephan Smeekes and Jean-Pierre Urbain for useful comments that helped improve the paper. Part of the paper was written while the authors were at Maastricht University, whose hospitality is gratefully acknowledged. All errors are ours. The data and programs used in this paper are available upon request.

†Department of Economics, thomas.deckers@uni-bonn.de.

‡Department of Economics and Econometrics, Nettelbosje 2, 9747AE Groningen, Netherlands. +31 (0)50 363 3836, c.h.hanck@rug.nl.

1 Introduction

Why do some countries grow faster than others? What determines long-run economic growth? These intriguing questions have been part of economic research since its beginnings. Answering them would permit more stable, long-run oriented policy making. Moreover, knowing what truly determines growth would help tackle the inequality in living standards between countries. The importance of this issue is reflected in the extensive research in this area. Unfortunately, different authors find quite diverse components to determine economic growth (Durlauf, Kourtellos and Tan, 2008). A far from exhaustive list includes initial GDP, fertility rates, high school enrolment rates, public debt and quality of institutions.

In judging whether a variable is significant in an economic growth model, researchers often rely on traditional significance tests. That is, one tests each coefficient individually at some level α using appropriate p-values. Given the large set of possible explanatory variables in a growth regression, one simultaneously tests a large number of hypotheses. Given some α, the probability of committing *at least one* type I error is arbitrarily larger than α. To see this, note the event of a rejection is a Bernoulli random variable with "success" probability α if the null is true. Hence, assuming (for illustration only) that all hypotheses are true and independent, P_l, the probability of finding l rejections in k tests, corresponding to k possible regressors, is the probability mass function of a Binomial random variable,

$$P_l = \left(\begin{array}{c} k \\ l \end{array} \right) \alpha^l (1-\alpha)^{k-l}.$$

Therefore, the probability of at least one erroneous rejection for $\alpha = 0.05$ and $k = 50$ equals

$$P_{l \geqslant 1} = \sum_{j=1}^{50} \left(\begin{array}{c} 50 \\ j \end{array} \right) 0.05^j (1-0.05)^{50-j} = 0.9231.$$

Hence, one is bound to erroneously declare irrelevant variables to explain growth. This is of substantial policy relevance. Growth regressions are used to identify effective policy measures to boost growth. Now, if variables are only spuriously found related to growth, ineffective public expenditures may well arise.

The issue is thus related to, but different from, data mining, where, in the words of Lovell (1983) "a data miner uncovers t-statistics that appear significant at the 0.05 level by running a large number of alternative regressions on the same body of data." Doing so, "the probability of a type I error of rejecting the null hypothesis when it is true is much greater than the claimed 5%."

Another topic extensively researched in growth econometrics is whether per-capita outputs of different economies converge. Investigating convergence is relevant for example in the context of European integration. Of the different notions of convergence proposed in the literature[1] we will focus on convergence *across* economies using a *time-series* definition proposed by Pesaran (2007a). There, two economies converge only if a unit root test on the output gap of two economies,

[1] For a survey see for example Islam (2003).

i.e. the difference of the log per-capita income, rejects. These pairwise tests with the null of no convergence are conducted for all different combinations of n countries. This results in $n(n-1)/2$ simultaneous tests. Again, given the high number of simultaneous tests, even if no country pair converges one is bound to falsely reject the null of no convergence for several pairs.

We propose to tackle these problems by using multiple testing techniques. These take the multiplicity of tests performed explicitly into account. One way to achieve such multiplicity control is to only declare a variable significant if its p-value satisfies $p_j \leqslant \alpha_j$ for some suitably chosen cutoff $\alpha_j \leqslant \alpha$. Such multiple testing techniques are routinely applied in many areas of applied statistics that involve multiple hypothesis testing, e.g. genomics (e.g., Dudoit and van der Laan, 2007). Following Romano, Shaikh and Wolf (2008b), we argue that selecting important regressors in this fashion can fruitfully be used as a simple-to-use model selection device. In that way, we aim to find a new way to robustly select explanatory variables in economic growth models and to shed some further light on the question whether economies converge.

We will focus on controlling the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). The FDR is defined as the expected value of the number of falsely rejected hypotheses divided by the overall number of rejections. It is thus an extension of the notion of a type I error to the multiple testing situation. For the first application, rejection means declaring some determinant to be significant in an economic growth model. In the second application, rejection means rejection of the null of no convergence.

The rejective power of the FDR controlling procedures might differ substantially (Romano, Shaikh and Wolf, 2008a). Thus, we first conduct an extensive Monte Carlo study judging the effectiveness of the procedures under different settings relevant to the present testing problems.

We find that few other variables beyond initial GDP level are needed to explain growth. We also show that output convergence using a time series definition with the necessary condition of no unit root in the output gap of two economies does not seem to hold.

This paper is organized as follows. Section 2 surveys previous approaches in cross-sectional growth models as well as in the convergence literature. Section 3 describes the FDR controlling procedures that we employ. Section 4 assesses their quality in a Monte Carlo study. Section 5 examines cross sectional growth regressions using FDR controlling techniques and compares the results to other approaches used in the literature. Section 6 conducts FDR controlling pairwise tests for output convergence. The final section concludes.

2 Literature Review

2.1 Growth Regressions

We follow standard practice and regress the real per capita output on two sets of variables. The first set includes levels of state variables measuring the initial position of an economy. The second set uses variables accounting for the difference in steady states across economies. Such

a specification is consistent with a variety of neoclassical growth models that have as possible solution a log-linearization around the steady state of the form (Barro and Sala-i-Martin, 1995)

$$\log y_T - \log y_0 = -(1 - e^{-\lambda T}) \log y_0 + (1 - e^{-\lambda T}) \log y^*, \tag{1}$$

where $\log y_t$ is the logarithm of the per capita gross domestic product (GDP for short) at time t, $\log y^*$ is its steady-state value, and λ is the convergence rate. If all economies have the same steady state, the regression corresponding to (1) for n countries and time periods T and 0 reads:

$$\log(y_{iT}/y_{i0}) = \mu + \delta \log(y_{i0}) + u_i, \quad i = 1, \ldots, n, \tag{2}$$

where u_i is a disturbance term. Here, δ is expected to have a negative sign, indicating that poor economies tend to grow faster and converge to rich ones. This is called *absolute convergence* (Barro and Sala-i-Martin, 1995). But when taking (2) to the data, this effect can only be seen when fitting the regression to a set of relatively homogeneous economies. This is because economies generally differ in their steady state. For heterogeneous economies, the sign of δ becomes ambiguous (Barro and Sala-i-Martin, 1995). The reason is that heterogeneous economies have different individual-specific parameters that determine their steady states. Omitting these from (2) leads to a bias. This motivates the concept of *conditional convergence*. To incorporate this into (2), one adds additional control variables $\boldsymbol{x}_i$, measuring the steady state determinants of an economy. The equation then becomes:

$$\log(y_{iT}/y_{i0}) = \mu + \delta \log(y_{i0}) + \boldsymbol{x}_i'\boldsymbol{\beta} + u_i, \quad i = 1, \ldots, n, \tag{3}$$

Although specifications like (3) are very widely used in the literature, there is little agreement on which variables to include in the vector $\boldsymbol{x}_i$.

Barro (1991) performs growth regressions with data for 98 countries from 1960 to 1985. Next to the negative relation to initial GDP, he finds a positive relation of growth to initial human capital. Furthermore, growth appears to be inversely related to the share of government consumption and a proxy for market distortions, but positively related to measures of political stability. Following this seminal work, the literature pointed out many different variables that would be significant in explaining growth. However, few authors perform multiplicity control in spite of often using a number of candidate variables of the same order of magnitude as n.

Recognizing this problem, Levine and Renelt (1992) apply the extreme bound theory to find a robust selection of variables in empirical growth models. But since this method is quite severe, they only find the initial level of income, the investment rate, the secondary school enrolment rate and the rate of population growth to have explanatory power. Sala-i-Martin (1997) uses an approach where he aims at assigning some level of certainty to each variable. Doing so, he finds variables like geography, measures of political quality, shares of certain religions or whether a country is a former Spanish colony to influence growth. Fernandez, Ley and Steel (2001), Sala-i-Martin, Doppelhofer and Miller (2004) and Eicher, Papageorgiou and Raftery (2010) use Bayesian Model Averaging (BMA) techniques to account for the fact that growth theories are not mutually exclusive and that even if one knew the true theories, it would be unclear which variable

to include for each. When accounting for this model uncertainty, Sala-i-Martin *et al.* (2004) find 11 variables to robustly explain growth, of which initial GDP level again has the strongest impact. The related work of Durlauf *et al.* (2008) aims at finding relevant growth theories rather than single variables. The results are still comparable to the aforementioned papers in that they declare a growth theory to be important if at least one variable associated with that theory appears in the final model. The robustness of these studies using BMA is discussed in Ley and Steel (2009) and Eicher *et al.* (2010). They show that the choices made concerning prior distributions strongly affect the number of variables declared significant. Thus the results are sensitive to difficult and somewhat subjective user choices.

2.2 Pairwise Testing for Output Convergence

The literature discusses different definitions for output convergence and resulting tests; see Islam (2003) for a survey. We are concerned with pairwise output convergence across countries. We work with the time-series definition of Pesaran (2007a). He assumes the following common factor model for the GDP of country i at time t, y_{it},[2]

$$y_{it} = c_i + g_i t + \boldsymbol{\theta}_i' \mathbf{f}_t + \epsilon_{it} + \eta_{it} \quad \text{for } i = 1, 2, \ldots, n. \tag{4}$$

Here, η_{it} is a zero-mean stationary process, but $\mathbf{f}_t$ and ϵ_{it} could be non-stationary. Following the definition in Bernard and Durlauf (1995), countries i and j converge if

$$\lim_{k\to\infty} \mathrm{E}(y_{i,t+k} - y_{j,t+k}|F_t) = 0 \quad \text{at any fixed time } t, \tag{5}$$

where F_t is an information set containing at least the current and past output series $y_{i,t-s}$ for $i = 1, 2, \ldots, n$ and $s = 1, 2, \ldots, t$. Bernard and Durlauf (1995) show that for (5) to hold, a necessary, though not sufficient, condition is that the GDPs are cointegrated with cointegrating vector $(1, -1)'$. Substituting (4) into (5) the condition for convergence reads

$$\begin{aligned}\lim_{k\to\infty} (c_i - c_j) + (g_i - g_j)(t+k) + (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)' \,\mathrm{E}(\mathbf{f}_{t+k}|F_t) + \\ \mathrm{E}(\epsilon_{i,t+k} - \epsilon_{j,t+k}|F_t) + \mathrm{E}(\eta_{i,t+k} - \eta_{j,t+k}|F_t) = 0.\end{aligned}$$

Since η_{it} is assumed to be stationary, so is the difference $\eta_{it} - \eta_{jt}$. Thus, we have

$$\lim_{k\to\infty} \mathrm{E}(\eta_{i,t+k} - \eta_{j,t+k}|F_t) = \mathrm{E}(\eta_{it} - \eta_{j,t}) = 0.$$

Moreover, for convergence it has to hold that ϵ_{it} is $I(0)$, since otherwise $\lim_{k\to\infty} \mathrm{E}(\epsilon_{i,t+k} - \epsilon_{j,t+k}|F_t) \neq 0$. We now have to consider two cases for the order of integration of $\boldsymbol{\theta}_i'\mathbf{f}_t$, namely $\boldsymbol{\theta}_i'\mathbf{f}_t \sim I(0)$ and $\boldsymbol{\theta}_i'\mathbf{f}_t \sim I(1)$. Under the former, convergence according to (5) requires that $c_i = c_j$ and $g_i = g_j$. If $\boldsymbol{\theta}'\mathbf{f}_t \sim I(1)$ we must also have that $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$.

Pesaran (2007a) argues that of these three conditions, the most unlikely to hold is $c_i = c_j$. This would require the two converging countries to be identical in nearly every respect, including for example their initial endowment. Hence, he employs a less stringent definition of convergence.

[2]For a detailed derivation and discussion of the corresponding assumptions refer to Pesaran (2007a).

The definition is based on the idea that for two countries i and j to converge, the output gap $y_{it} - y_{jt}$ should not fall outside a pre-specified interval C with high probability π. Formally,

$$\Pr\{|y_{i,t+s} - y_{j,t+s}| < C|F_t\} > \pi, \tag{6}$$

for all $s = 1, 2, \ldots, \infty$. This less stringent definition of pairwise convergence requires $y_{it} - y_{jt}$ not to have a stochastic or deterministic trend. But contrary to (5), (6) does not require $c_i = c_j$ for two countries to converge.

Using a cointegration framework and (6), the GDPs of two countries i and j should be cointegrated and cotrended with vectors $(1, -1)'$ for them to converge. Pesaran (2007a) demonstrates that this condition can be checked by directly testing each output gap for the absence of a unit root and a linear trend. In this setting, the absence of a unit root can be seen as a necessary condition and the additional absence of a linear trend ($g_i = g_j$) as a sufficient condition.

Pesaran (2007a) tests for the absence of a unit root for all possible $n(n-1)/2$ pairs of countries. For large n and T, if no country pair converges, the null of no convergence should only be rejected for a fraction of pairs equal to the significance level α of the individual tests applied. In this case, the null could have been rejected by chance for pairs found to be convergent. On the other hand, if all country pairs converge, for large n and T, the fraction of country pairs found convergent should tend to 1. Using data from the Penn World Tables 6.1 he only rejects the null for a fraction of country pairs roughly equal to α. Hence, he finds no evidence for overall convergence. However, since his approach makes no statements about individual country pairs it cannot say whether the fraction of rejections consists entirely of type I errors or possibly does contain some correct rejections.

Further studies using a time-series framework and similar definitions of convergence are also largely in disfavor of the overall convergence hypothesis. Bernard and Durlauf (1995) apply cointegration tests in a panel setting to test for convergence. They fail to find convergence for a set of 15 OECD countries when testing for the cointegration vector $(1, -1)'$. Nevertheless they find cointegration relationships of the form $(1, -a)$ among different countries. This indicates, following their definition, conditional convergence, a concept less strict than the one employed here. Pesaran (2007a) shows that this approach can only handle a limited number of countries simultaneously. Earlier studies using univariate time series techniques (e.g., Campbell and Mankiw, 1989; Quah, 1990) also fail to find evidence for overall convergence. A problem inherent to their approaches is the choice of a 'reference country' to which convergence of the other economies is tested.

This paper tests the null of a unit root in each output gap for all possible $n(n-1)/2$ pairs vs. the alternative of no unit root. Unlike Pesaran (2007a) we directly account for the multiplicity of tests, thereby allowing us to make statements about individual country pairs rather than fractions of rejections. We thus test for the necessary condition for convergence of two countries, controlling the FDR. Doing so, we clarify if the fraction of convergent pairs in Pesaran (2007a) is spurious or if there is some evidence of 'true convergence'. Our results in section 6 show that not even the necessary condition holds such that we do not investigate the sufficient one $g_i = g_j$.

3 Controlling the FDR

The false discovery rate (FDR) as a desirable measurement of type I errors in multiple testing situations is introduced by Benjamini and Hochberg (1995). They also provide a step-up procedure that is shown to control the FDR at a desired level γ. A step-up method considers the hypotheses ordered from most significant to least significant based on their p-values. The idea is then, beginning with the least significant hypothesis, to accept hypotheses up to a certain point and reject the remaining ones.

We now present four methods controlling the FDR in more detail. Concretely, we employ the methods from Benjamini and Hochberg (1995), Storey, Taylor and Siegmund (2004), Benjamini, Krieger and Yekutieli (2006) and the bootstrap method of Romano *et al.* (2008a). We will frequently refer to these as BH method, Storey method, BKY algorithm and bootstrap method.

Adapting a notation similar to Benjamini and Hochberg (1995) and referring to Table 1, there are m hypotheses to be tested simultaneously out of which m_0 are true. $\mathbf{R}$ is an observable random variable, whereas $\mathbf{U}$, $\mathbf{F}$, $\mathbf{S}$ and $\mathbf{T}$ are unobservable random variables. The proportion of falsely rejected null hypotheses can be described by $\mathbf{Q} = \mathbf{F}/(\mathbf{F}+\mathbf{S})$. Naturally, if $\mathbf{F}+\mathbf{S} = 0$, we take $\mathbf{Q} = 0$. The FDR is then defined as $\mathrm{E}(\mathbf{Q}) = \mathrm{E}(\mathbf{F}/(\mathbf{F}+\mathbf{S})) = \mathrm{E}(\mathbf{F}/\mathbf{R})$. Lower-case letters denote the realizations of the underlying random variable.

3.1 BH Method

In the FDR controlling method suggested in Benjamini and Hochberg (1995) one first chooses a level γ at which to control the FDR. Let $\widehat{p}_{(1)} \leq \ldots \leq \widehat{p}_{(m)}$ be the ordered p-values and $H_{(1)}, \ldots, H_{(m)}$ the corresponding null hypotheses, such that the hypotheses are now ordered from most to least significant. For $1 \leq j \leq m$, let $\gamma_j = \frac{j}{m}\gamma$. Then the method rejects $H_{(1)}, \ldots, H_{(j^*)}$, where j^* is the largest j such that $\widehat{p}_{(j)} \leq \gamma_j$. If no such j exists, no hypothesis is rejected. The intuition behind the procedure is the following: When testing each hypothesis at some level α, we have that $\mathrm{E}(\mathbf{F}) \leq \alpha m$. So a valid estimate for the upper bound of the expected value of $\mathbf{Q}$ is $\alpha m/r(\alpha)$, where $r(\alpha)$ denotes the number of rejected hypotheses when testing each at level α. Given the observed p-values, α can now be chosen to maximize the observed number of rejections $r(\alpha)$ given the constraint that FDR $\leq \gamma$. Given FDR $\leq \gamma$, maximizing the observed number of

Table 1: Number of decisions made when testing m null hypotheses

	Declared non-significant	Declared significant	Total
True null hypotheses	$\mathbf{U}$	$\mathbf{F}$	m_0
Non-true null hypotheses	$\mathbf{T}$	$\mathbf{S}$	$m - m_0$
	m-$\mathbf{R}$	$\mathbf{R}$	m

rejections is desirable, since this identifies as many false hypotheses as possible. Here this means that one finds as many relevant variables as possible. Moreover, Benjamini and Yekutieli (2001) show control of the FDR under positive regression dependency, which under certain conditions includes coefficient test statistics in static regressions. The Monte Carlo study in section 4 shows that the FDR is also controlled under plausible assumptions about e.g. the DGP of a cross-sectional growth regression.

3.2 Storey Method

Even if the conditions of the BH method are met the method is conservative as it can be shown that FDR $\leq \frac{m_0}{m}\gamma$. So, unless $m_0 = m$, the power of the method can be improved by redefining γ_j as $\gamma_j = \frac{j}{m_0}\gamma$. Since m_0 is unknown, it has to be estimated. Storey *et al.* (2004) propose to estimate m_0 by $\widehat{m}_0 = \frac{\#[\widehat{p}_j > \lambda]+1}{1-\lambda}$, where $\lambda \in (0,1)$ is a user-specified parameter. The idea behind this estimator is the following: p-values corresponding to true null hypotheses approximately follow a uniform$[0,1]$ distribution. Therefore one would expect $m_0(1-\lambda)$ of these to lie in the interval $(\lambda, 1]$. When replacing m in the BH method by $\hat{m}_0$, Storey *et al.* (2004) show that this procedure typically controls the FDR whenever the BH method does. The Storey procedure can however be quite liberal under constant positively dependent p-values.

3.3 BKY Algorithm

Benjamini *et al.* (2006) propose another improvement of the original BH method. It comes in the form of a two step algorithm. First, one applies the BH procedure at a level $\gamma^* = \gamma/(1+\gamma)$. If $r = 0$, no hypothesis is rejected and the procedure stops. Likewise, if $r = m$, all hypotheses are rejected. Otherwise, the procedure continues using the BH method with γ_j replaced by $\frac{j}{\hat{m}_0}\gamma^*$, where $\hat{m}_0 = m - r$. Benjamini *et al.* (2006) prove that this method works under independent test statistics and also provide simulations suggesting that it works under dependent test statistics.

3.4 Bootstrap Method

The bootstrap method is a step-down rather than a step-up procedure as the three previously described methods. Assume without loss of generality that a hypothesis $H_{(i)}$ is rejected for large values of its corresponding test statistic T_i. Further, order the test statistics from smallest to largest, i.e. $T_1 \leq T_2 \leq \ldots \leq T_m$, and let $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ denote the corresponding hypotheses. A step-down procedure then compares the largest test statistic T_m with a suitable critical value c_m. If $T_m < c_m$ the procedure rejects no hypothesis; otherwise it rejects $H_{(m)}$ and steps down to T_{m-1}. The procedure continues in this fashion until it either rejects $H_{(1)}$ or does not reject the current hypothesis. More formally, a step-down procedure rejects hypotheses

$$H_{(m)}, H_{(m-1)}, \ldots, H_{(m-j^*)},$$

where j^* is the largest integer j satisfying

$$T_m \geq c_m, T_{m-1} \geq c_{m-1}, \ldots, T_{m-j} \geq c_{m-j}.$$

If no such j exists, the method does not reject any hypotheses.

3.4.1 Intuition

Recall the FDR is the expected value of falsely rejected hypotheses over the total number of rejected hypotheses. For any step-down procedure this expected value can be calculated as

$$\begin{aligned} \text{FDR} &= E\left[\frac{\mathbf{F}}{\max\{\mathbf{R}, 1\}}\right] = \sum_{1 \leq r \leq m} \frac{1}{r} E[\mathbf{F}|\mathbf{R} = r] P\{\mathbf{R} = r\} \\ &= \sum_{1 \leq r \leq m} \frac{1}{r} E[\mathbf{F}|\mathbf{R} = r] \\ &\quad \times P\{T_m \geq c_m, \ldots, T_{m-r+1} \geq c_{m-r+1}, T_{m-r} < c_{m-r}\}, \end{aligned}$$

where the event $T_{m-r} < c_{m-r}$ is defined to be true when $r = m$.

Assume without loss of generality that the true hypotheses correspond to the indices $\{1, \ldots, m_0\}$. Under weak assumptions relating to test consistency, all other (i.e. the false) hypotheses are rejected with probability tending to one. Moreover, let $T_{r:t}$ denote the rth smallest of the test statistics $T_1, \ldots, T_t$. In particular, when $t = m_0$, $T_{r:m_0}$ denotes the rth smallest test statistic of all true hypotheses. Then, with probability approaching one,

$$\begin{aligned} \text{FDR} &= \sum_{m - m_0 + 1 \leq r \leq m} \frac{r - m + m_0}{r} \\ &\quad \times P\{T_{m_0:m_0} \geq c_{m_0}, \ldots, T_{m-r+1:m_0} \geq c_{m-r+1}, T_{m-r:m_0} < c_{m-r}\}. \end{aligned} \tag{7}$$

Again the event $T_{m-r} < c_{m-r}$ is defined to be true when $r = m$. The intuition behind (7) is the following: Given that all false hypotheses are rejected with probability tending to one, one only needs to search among the hypotheses $\{H_1, \ldots, H_{m_0}\}$ how many of these are falsely declared significant. Then, summing over every possible proportion of falsely rejected hypotheses weighted by its probability gives the expected value of this, i.e. the FDR.

Since m_0 is unknown, one has to ensure that (7) is bounded above by γ for every possible m_0. That is exactly the condition used to recursively determine the critical values. For $m_0 = 1$, (7) simplifies to

$$\text{FDR} = \frac{1}{m} P\{T_{1:1} \geq c_1\}.$$

Hence one can calculate the first critical value from

$$c_1 = \inf\left\{x \in \mathbb{R} : \frac{1}{m} P\{T_{1:1} \geq x\} \leq \gamma\right\}$$

If $m\gamma > 1$, c_1 is set to $-\infty$. For $m_0 = 2$, (7) equals

$$\frac{1}{m-1} P\{T_{2:2} \geq c_2, T_{1:2} < c_1\} + \frac{2}{m} P\{T_{2:2} \geq c_2, T_{1:2} \geq c_1\} \tag{8}$$

Hence, having determined c_1, c_2 then simply is the smallest number for which (8) is bounded above by γ. In general, having determined $c_1, \ldots, c_{j-1}$, c_j is the smallest value for which the following expression is bounded above by γ,

$$\begin{aligned} \text{FDR} &= \sum_{m-j+1 \leq r \leq m} \frac{r-m+j}{r} \\ &\quad \times P\left\{T_{j:j} \geq c_j, \ldots, T_{m-r+1:j} \geq c_{m-r+1}, T_{m-r:j} < c_{m-r}\right\} \end{aligned} \tag{9}$$

For $j = m$, (9) simplifies to

$$P\left\{T_{m:m} \geq c_m\right\}.$$

Thus, one can find the mth critical value via the following minimization:

$$c_m = \inf\left\{x \in \mathbb{R} : P\left\{T_{m:m} \geq x\right\} \leq \gamma\right\}.$$

In practice this choice of critical values is infeasible, since the probability measure P is unknown. Hence, P is approximated using bootstrap techniques as presented in the following.

3.4.2 Procedure

The idea is to replace P by a suitable $\hat{P}$ using bootstrap techniques. The exact choice of bootstrap depends of course on the nature of the data and the problem that is analyzed. We will present the two bootstrap procedures used later. Here, it is only required that $\hat{P}$ estimates P such that T_j^*, the bootstrapped test statistic, is a good approximation of T_j whenever the corresponding null hypothesis is true (Romano *et al.*, 2008a).

Given $\hat{P}$, the critical values are defined recursively as follows: Having determined $\hat{c}_1, \ldots, \hat{c}_{j-1}$, the jth critical value is determined using the minimization rule (Romano *et al.*, 2008a):

$$\begin{aligned} \hat{c}_j &= \inf\Bigg\{c \in \mathbb{R} : \sum_{m-j+1 \leq r \leq m} \frac{r-m+j}{r} \\ &\quad \times \hat{P}\left\{T_{j:j}^* \geq c, \ldots, T_{m-r+1:j}^* \geq \hat{c}_{m-r+1}, T_{m-r:j}^* < \hat{c}_{m-r}\right\} \leq \gamma\Bigg\} \end{aligned} \tag{10}$$

Here, it is crucial to understand the meaning of $T_{r:t}^*$. The index t stems from the ordering of the original test statistics, whereas r corresponds to the bootstrapped test statistics. So $T_{r:t}^*$ has the following meaning: Out of the t smallest original test statistics pick the rth smallest of the corresponding bootstrap test statistics.

3.4.3 Linear Regression Model

For the application to cross sectional growth regressions we consider the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u},$$

where $\boldsymbol{X}$ is a $n \times (k+1)$ matrix containing k possible regressors and a constant. All vectors are of dimension n. We use the following semi-parametric bootstrap:

1. Estimate $\boldsymbol{\beta}$ using $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.
2. Calculate the residuals $\hat{\boldsymbol{u}}$ using $\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and their demeaned version $\tilde{\boldsymbol{u}} = \hat{\boldsymbol{u}} - \boldsymbol{\iota}(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\hat{\boldsymbol{u}}$, where $\boldsymbol{\iota}$ is a vector of ones.
3. For each element of $\boldsymbol{\beta}$, calculate the t-statistic $\frac{\hat{\beta}_i}{\sqrt{s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{ii}}}$ for $\mathrm{H}_0 : \beta_i = 0$ against $\mathrm{H}_1 : \beta_i \neq 0$. Here $s^2 = \sum_i \frac{\hat{u}_i^2}{n-(k+1)}$.
4. Resample non-parametrically with replacement from $\tilde{\boldsymbol{u}}$ to obtain the bootstrap residuals u_i^* and build the bootstrap sample
$$y_i^* = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + u_i^*$$
5. Calculate $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}^*$ and $\boldsymbol{u}^* = \boldsymbol{y}^* - \boldsymbol{X}\hat{\boldsymbol{\beta}}^*$.
6. For each element of $\hat{\boldsymbol{\beta}}^*$ construct the bootstrapped version of each individual t-statistic using the formula $\frac{\hat{\beta}_i^* - \hat{\beta}_i}{\sqrt{s^{2*}(\boldsymbol{X}'\boldsymbol{X})^{-1}_{ii}}}$, where $s^{2*} = \sum_i \frac{u_i^{*2}}{n-(k+1)}$. Repeat steps 4 – 6 B times.
7. Apply the rule (10) with $m = k+1$ to calculate the critical values.
8. Use the critical values from 7 and compare them in the previously described step-down fashion (10) to the t-statistics found in 2.

We thus bootstrap by estimating the original model under the alternative and calculating the bootstrap t-statistic accordingly. This has two major advantages. First this leads to an enormous gain in computation time of the algorithm, since one only needs to rebuild the DGP once and not $k+1$ times. Secondly, and more importantly, this preserves the dependence of the test statistics in each bootstrap iteration. This is important, since when applying the rule (10), one makes statements about the joint distribution of the test statistics.

3.4.4 Pairwise Testing for Output Convergence

When applying the bootstrap procedure to pairwise testing of output convergence we consider n different countries leading to $n(n-1)/2$ different country pairs. Previous applications of the Romano *et al.* (2008a) approach tested stationary variables. We now describe how to extend their approach to the present testing problem on nonstationary time series. We apply the following semi-parametric bootstrap unit root test procedure developed in Smeekes (2009).

1. Calculate the output gap $d_{ijt} = y_{it} - y_{jt}$ for each country pair, i.e. for $i = 1, \ldots, n-1$, $j = 2, \ldots, n$ and $t = 1, \ldots, T$. Do the next steps simultaneously for all $i = 1, \ldots, n-1$ and $j = 2, \ldots, n$.
2. Detrend the output gap d_{ijt}. We consider two detrending schemes. Either calculate $d_{ijt}^d = d_{ijt} - \hat{\boldsymbol{\phi}}'\boldsymbol{z}_t$, where $\boldsymbol{z}_t = (1, t)'$. Here, $\hat{\boldsymbol{\phi}}$ is the usual OLS estimator, $\hat{\boldsymbol{\phi}} = \left(\sum_{t=1}^T \boldsymbol{z}_t\boldsymbol{z}_t'\right)^{-1} \times \left(\sum_{t=1}^T \boldsymbol{z}_t d_{ijt}\right)$. We also consider GLS detrending. Elliot, Rothenberg and Stock (1996) show that GLS detrending may result in higher power of the ADF test against local alternatives of the form $\rho = 1 + \bar{c}T^{-1}$. Let $\boldsymbol{z}_{1\bar{c}} = \boldsymbol{z}_1$ and $\boldsymbol{z}_{t\bar{c}} = \boldsymbol{z}_t - (1 + \bar{c}T^{-1})\boldsymbol{z}_{t-1}$ for $t = 1, 2, \ldots, T$.

Likewise, define $d_{ij1\bar{c}} = d_{ij1}$ and $d_{ijt\bar{c}} = d_{ijt} - (1 + \bar{c}T^{-1})d_{ij,t-1}$ for $t = 1, 2, \ldots, T$. Then, calculate

$$\hat{\boldsymbol{\phi}}_{\bar{c}} = \left(\sum_{t=1}^{T} \boldsymbol{z}_{t\bar{c}} \boldsymbol{z}'_{t\bar{c}} \right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{z}_{t\bar{c}} d_{ijt\bar{c}} \right).$$

Finally calculate $d^d_{ijt} = d_{ijt} - \hat{\boldsymbol{\phi}}'_{\bar{c}} \boldsymbol{z}_t$.

3. Estimate an ADF regression of order p for d^d_{ijt} and calculate the residuals as

$$\hat{\epsilon}_{ijt} = \Delta d^d_{ijt} - \hat{\alpha} d^d_{ij,t-1} - \sum_{j=1}^{p} \hat{\psi}_j \Delta d^d_{ij,t-j}.^{3}$$

Calculate the demeaned residuals $\tilde{\epsilon}_{ijt} = \hat{\epsilon}_{ijt} - \frac{1}{n-p-1} \sum_t \hat{\epsilon}_{ijt}$. Also calculate the ADF test statistic $t_{\hat{\alpha}}$ for $\hat{\alpha}$ and $\mathrm{ADF}_{-1} = (-1)t_{\hat{\alpha}}$, so as to reject for large values of the test statistic as assumed in the derivation of critical values (10).

4. Resample $\tilde{\epsilon}_{ijt}$ non-parametrically with replacement to obtain the bootstrap residuals ϵ^*_{ijt}.
5. Build u^*_{ijt} recursively as $u^*_{ijt} = \sum_{j=1}^{p} \hat{\psi}_j u^*_{ij,t-j} + \epsilon^*_{ijt}$. Then build $d^*_{ijt} = d^*_{ij,t-1} + u^*_{ijt}$.[4]
6. Detrend d^*_{ijt} as in step 2 to obtain d^{d*}_{ijt}.
7. Estimate by OLS the ADF regression

$$\Delta d^{d*}_{ijt} = \hat{\alpha}^* d^{d*}_{ij,t-1} + \sum_{j=1}^{p} \hat{\psi}_j \Delta d^{d*}_{ij,t-j} + \hat{\epsilon}^*_{ijt}$$

and calculate the ADF test statistic for $\hat{\alpha}^*$ and ADF^*_{-1}. Repeat steps 2 – 7 B times.

8. Apply the rule (10) with $m = n(n-1)/2$ and ADF^*_{-1} to calculate the critical values.
9. Use the critical values from 8 and compare them to ADF_{-1} from 2 using the step-down procedure (10).

The bootstrap method is consistent, i.e. satisfies $\limsup_{T\to\infty} \mathrm{FDR} \leqslant \alpha$ under a set of weak conditions (see Romano *et al.*, 2008a, Thm. 1). Concretely, these are (i) continuous marginal distributions of the test statistics, (ii) connected support of the joint distribution of the test statistic, (iii) the test statistics form an exchangeable sequence, (iv) availability of consistent estimators of the standard errors of coefficient estimators, e.g. $\hat{\alpha}$ and (v) weak convergence of bootstrap distributions to the true one as $T \to \infty$. Conditions (i) and (iv) are well-known to hold in the unit root literature (Phillips, 1987), while (ii) is a regularity condition. Pesaran (2007b) shows (iii) to hold in the panel unit root testing framework. Bootstrap consistency results for nonstationary panels (v) are provided by Smeekes (2009).

[3] If the estimated AR process is explosive we impose a root bound as suggested in Burridge and Taylor (2004).

[4] Following Smeekes (2009), we do not add deterministic components to the bootstrapped series for simplicity.

4 Monte Carlo Study

We now shed some light on the performance of the FDR controlling techniques described above. We compare the four procedures with each other and with the classical approach to hypothesis testing (i.e. rejecting H_i if $p_i \leq \alpha$). The first performance criterion is the average of the proportion of falsely rejected hypotheses. This will, as the number of simulations grows, converge to the FDR. The second criterion is the average number of rightly rejected null hypotheses. This provides insights on the rejective power of the procedures. Section 4.1 considers significance tests in the linear regression model, corresponding to cross sectional growth regressions. Section 4.2 investigates multiple unit root tests as in the tests for output convergence.

4.1 Linear Regression Model

We simulate a data generating process (DGP) of the following form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}.$$

In the basic setup, $\boldsymbol{X} = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_k')'$ is a $n \times k$ matrix of regressors, where each row $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{ki})$ is distributed multivariate normal with mean zero, variance one and common correlation ρ. The elements u_i of the vector $\boldsymbol{u}$ are $iid\ \mathrm{N}(0,1)$. The vector $\boldsymbol{\beta}$ is $k \times 1$. In each setup, $n = 100$ and $k = 50$.[5] We consider $\rho = \{0, 0.3, 0.5\}$ and investigate three scenarios for the vector $\boldsymbol{\beta}$:

- Every tenth $\beta_i = 0.5$, and the remaining $\beta_i = 0$, such that there are 5 false hypotheses.[6]
- Every fifth $\beta_i = 0.5$, and the remaining $\beta_i = 0$, such that there are 10 false hypotheses.
- Every second $\beta_i = 0.5$, and the remaining $\beta_i = 0$, such that there are 25 false hypotheses.

For the case of five false hypotheses, Table 2 shows that all multiple testing procedures control the FDR for any correlation ρ. Classical testing does not; in general FDRs substantially higher than the nominal level of the individual tests result. The violation of the FDR increases in ρ. One reason for this is that the larger ρ, the larger the problem of multicollinearity among the regressors. Hence, the individual t-tests become less reliable. This can also be seen from the fact that the number of right rejections declines for higher ρ. Unsurprisingly, the number of right rejections using classical testing is above that of the multiple testing procedures. But this comes at the price of unacceptable many wrong rejections. Moreover, unlike with multiple testing procedures, the FDR of classical testing is unknown in practice. Hence, the user does not know how many hypotheses are rightfully rejected. Among the FDR controlling procedures the bootstrap yields the highest number of right rejections. This likely is due to the bootstrap method taking the dependence between the test statistics into account. Hence the bootstrap appears to slightly outperform the other FDR controlling procedures.

[5] Unreported experiments draw qualitatively similar pictures for different n and k.

[6] The value 0.5 is chosen so as to be able to discriminate between the rejective power of the procedures. An extremely high value for example, would have resulted in all false hypotheses being rejected by all procedures. A low value would have resulted in the opposite.

Table 2: Linear regression model with 5 false hypotheses

	# false hypotheses: 5		Sample size: 100		# Regressors: 50	
	$\rho = 0$					
	1%		5%		10%	
	FDR	Right Rejections	FDR	Right Rejections	FDR	Right Rejections
Classical	0.083	3.96	0.273	4.61	0.426	4.81
BH	0.010	2.27	0.038	3.35	0.086	3.81
Storey	0.013	2.32	0.049	3.40	0.102	3.84
BKY	0.010	2.29	0.040	3.36	0.086	3.79
Bootstrap	0.011	2.36	0.052	3.45	0.010	3.94
	$\rho = 0.3$					
	1%		5%		10%	
	FDR	Right Rejections	FDR	Right Rejections	FDR	Right Rejections
Classical	0.097	3.08	0.302	4.12	0.457	4.49
BH	0.007	1.21	0.043	2.29	0.092	2.85
Storey	0.008	1.26	0.051	2.35	0.113	2.91
BKY	0.007	1.22	0.043	2.28	0.090	2.80
Bootstrap	0.009	1.22	0.050	2.39	0.102	2.95
	$\rho = 0.5$					
	1%		5%		10%	
	FDR	Right Rejections	FDR	Right Rejections	FDR	Right Rejections
Classical	0.125	2.250	0.336	3.48	0.473	3.98
BH	0.009	0.547	0.042	1.32	0.080	1.80
Storey	0.011	0.594	0.050	1.35	0.096	1.87
BKY	0.009	0.549	0.042	1.30	0.078	1.75
Bootstrap	0.010	0.592	0.048	1.35	0.092	1.88

This table shows the results of a Monte Carlo simulation as specified in section 4.1 with 2000 simulations and the indicated parameter settings. The procedures controlling the FDR were applied as described in section 3. For the Storey procedure, $\lambda = 0.5$.

When either 10 or 25 coefficients are significant, the picture is qualitatively the same.[7] Again, all procedures except classical testing control the FDR at the desired γ. Moreover, the bootstrap approach again appears to be most powerful. The non-control of the FDR using classical testing becomes less severe as the number of false hypotheses increases. In view of the definition of the FDR, $\mathrm{E}\left(\frac{\mathbf{F}}{\mathbf{F+S}}\right)$, this should be no surprise. As more variables are rightly declared significant, keeping the total number of false rejections, **F**, constant, the FDR is lower by construction.

4.2 Pairwise Testing for Output Convergence

As in Pesaran (2007a) we use the following DGP

$$y_{it} = \gamma_i f_t + \epsilon_{it},$$

where

$$f_t = f_{t-1} + v_t, \quad v_t = \rho_v v_{t-1} + e_t, \quad e_t \sim iid\, \mathrm{N}(0, 1 - \rho_v^2)$$

and

$$\epsilon_{it} = \rho_i \epsilon_{i,t-1} + \nu_{it}, \quad \nu_{it} \sim iid\, \mathrm{N}(0, \sigma_{\nu i}^2(1 - \rho_i^2)), \quad \sigma_{\nu i}^2 \sim iid\, \mathrm{uniform}[0.5, 1.5],$$

for $i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, T$. We consider $n = 10$ and set $T = 50$ and $T = 100$.[8] When generating the autoregressive processes we start at $t = -49$ and discard the first 50 draws. As in

[7] Results are available upon request. **For the referees: Please refer to Tables A.1 and A.2.**

[8] Due to the large computation time required to find the bootstrap critical values in Algorithm 3.4.4 we only conduct limited experiments for larger n, for which we find a qualitatively similar pattern. Results are available upon request. **For the referees: Please refer to Table A.6.**

Table 3: Pairwise unit root tests with 10 false hypotheses and $T = 50$

$n = 10$[45pairs]	$T = 50$		false hypotheses: 5[10pairs]		
OLS detrending			GLS detrending		
$p = 0$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.203	9.95	Classical approach	0.184	9.94
BH	0.160	9.54	BH	0.142	9.48
Storey	0.281	9.45	Storey	0.179	9.38
BKY	0.171	9.62	BKY	0.150	9.58
Bootstrap	0.168	9.62	Bootstrap	0.154	9.70
$p = 1$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.173	9.47	Classical approach	0.159	9.35
BH	0.119	7.00	BH	0.115	7.24
Storey	0.316	7.78	Storey	0.284	7.86
BKY	0.126	7.24	BKY	0.123	7.43
Bootstrap	0.115	7.15	Bootstrap	0.123	7.68
$p = 3$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.144	6.69	Classical approach	0.166	6.48
BH	0.0668	1.85	BH	0.0867	2.26
Storey	0.279	4.63	Storey	0.291	4.82
BKY	0.0648	1.94	BKY	0.0853	2.33
Bootstrap	0.0599	1.89	Bootstrap	0.0652	2.14
$p = 4$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.158	5.10	Classical approach	0.172	4.94
BH	0.0505	0.866	BH	0.0636	1.08
Storey	0.272	3.89	Storey	0.266	3.81
BKY	0.0472	0.910	BKY	0.0600	1.13
Bootstrap	0.0454	0.989	Bootstrap	0.0478	0.978
$p = 5$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.175	4.24	Classical approach	0.183	4.24
BH	0.0467	0.557	BH	0.0643	0.769
Storey	0.277	3.62	Storey	0.259	3.45
BKY	0.0450	0.637	BKY	0.0630	0.845
Bootstrap	0.0386	0.650	Bootstrap	0.0276	0.592
$p = 10$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.216	1.59	Classical approach	0.232	1.73
BH	0.0287	0.0730	BH	0.0498	0.154
Storey	0.190	1.81	Storey	0.162	1.15
BKY	0.0278	0.0840	BKY	0.0469	0.193
Bootstrap	0.0178	0.112	Bootstrap	0.00911	0.0770

This table shows the results of a Monte Carlo simulation as specified in section 4.2 with 1000 simulations and the indicated parameter settings. Tests were conducted for $\alpha = \gamma = 0.05$. The FDR controlling procedures applied are described in section 3. For the Storey procedure, $\lambda = 0.5$.

Pesaran (2007a) we take $\rho_v = 0.6$ and $\rho_i \sim iid$ uniform$[0.2, 0.6]$. We only consider one single pool of convergent countries, i.e. all countries within the pool converge with each other and all others do not. For all converging pairs we set $\gamma_i = \gamma_j = 1$. For all others, $\gamma_i \sim iid\ \chi^2_{\kappa_i}$, where κ_i is drawn with replacement from integers 1 to 10. We consider two scenarios:

- For 3/10 of all countries $\gamma_i = 1$. For $n = 10$ this amounts to 3 convergent pairs.
- For 1/2 of all countries $\gamma_i = 1$. For $n = 10$ this amounts to 10 convergent pairs.

When applying the ADF regression we consider $p = 0$ and $p = 1$ as well as deterministic lag length choices as suggested in Schwert (1989). We do not use model selection criteria such as AIC, since Leeb and Pötscher (2008) show that such an approach can distort subsequent tests. Rules we consider are: $p = 4(T/100)^{1/4}$, $p = 5(T/100)^{1/4}$, $p = 6(T/100)^{1/4}$ and $p = 12(T/100)^{1/4}$. These

rules are shown to work well in Demetrescu, Hassler and Kuzin (2008). We employ both OLS and GLS detrending in the ADF test (cf. sec. 3.4.4). We do not include a time trend, since the model is simulated without such. The null distribution when using GLS detrending asymptotically follows the standard ADF-distribution with no deterministics included. However, the finite sample distribution differs substantially from those. We therefore simulate MacKinnon-type finite sample distributions to get more accurate finite sample p-values.

Table 3 shows the results for $n = 10$, so 45 hypotheses are tested, $T = 50$ and 10 pairs are significant, i.e. contain no unit root. At any lag length p, classical testing results in a severe violation of the FDR, irrespective of whether OLS or GLS detrending is used. All the other procedures except Storey control the FDR for sufficiently high p.[9] A high number of lags is plausible for this DGP, as the d_{ijt} are sums of $AR(1)$ processes, that generally follow an $ARMA(2,1)$ (Granger and Morris, 1976), or $AR(\infty)$, process, that has to be approximated with a high number of lags. Moreover, that FDR control can only be attained for high p is in line with Pesaran (2007a), who finds size distortions of the individual tests when $p < 4$. Given the required p for each FDR controlling procedure, the bootstrap is most powerful. For the BH and the BKY method, GLS detrending yields somewhat higher power than OLS detrending.

Varying T and the number of false hypotheses leaves the results qualitatively the same.[10] Classical testing results in a violation of the FDR in all scenarios for OLS as well as GLS detrending. Again, the Storey procedure does not control the FDR. The main difference for $T = 100$ is that FDR control of the BH, BKY and bootstrap procedure is only attained when $p = 6$ and $p = 12$ (corresponding to the rules $p = 6(T/100)^{1/4}$ and $p = 12(T/100)^{1/4}$).

5 Empirical Growth Models Revisited

We now apply the above techniques to two widely used data sets. We first discuss the data, their sources and prior use and then present the results obtained for each. We estimate equation (3) by OLS. Of course, OLS generally produces inconsistent estimators under e.g. endogeneity and/or measurement errors. However, Hauk Jr. and Wacziarg (2009) have recently shown that these and other effects plaguing cross-sectional growth regressions largely cancel out in OLS estimation, making it a superior choice to more complicated estimators like Arellano and Bond's (1991).

In a third section we compare the results to previous studies that have used these data sets. The first data set is from Fernandez, Ley and Steel (2001) (FLS). It also constitutes the main reference point, since many authors have used this data set to whose results we can compare ours. The second data set was first used in Masanjala and Papageorgiou (2005) (MP).[11]

[9] We even find that all FDR controlling procedures also control the Familywise Error Rate (FWER), defined as the probability of one or more false rejections, a stricter criterion than the FDR.

[10] Results are available upon request. **For the referees: Please refer to Tables A.3, A.4 and A.5.**

[11] These data are available at http://econ.queensu.ca/jae/. We also investigated the data set of Sala-i-Martin, Doppelhofer and Miller (2004), but only found one variable ("Fraction Hindu") to be significant at the 5% level using classical testing. Controlling the FDR at any level, no variable was found to explain growth.

Table 4: Results for the FLS Data Set

	Regressor	$\hat{\beta}_i$	p-value	Classical	BH	Storey	BKY	Boot
1	GDP level 1960	−0.0170	0.00001	1%	1%	1%	1%	1%
2	Fraction Confucian	0.0748	0.00002	1%	1%	1%	1%	1%
3	Life expectancy	0.000891	0.00303	1%	5%	5%	5%	5%
4	Equipment investment	0.127	0.00778	1%	5%	5%	5%	5%
5	Sub-Saharan dummy	−0.0201	0.00589	1%	5%	5%	5%	5%
6	Fraction Muslim	0.0107	0.227	-	-	-	-	-
7	Rule of Law	0.0116	0.0679	10%	-	-	-	-
8	Number of years open economy	−0.00269	0.620	-	-	-	-	-
9	Degree of Capitalism	0.00111	0.284	-	-	-	-	-
10	Fraction Protestant	−0.00280	0.677	-	-	-	-	-
11	Fraction GDP in mining	0.0401	0.00769	1%	5%	5%	5%	5%
12	Non-Equipment investment	0.0367	0.0811	10%	-	-	-	-
13	Latin American dummy	−0.0127	0.0391	5%	-	10%	10%	10%
14	Primary School Enrollment, 1960	0.0202	0.0455	5%	-	10%	10%	10%
15	Fraction Buddhist	0.00734	0.277	-	-	-	-	-
16	Black-market premium	−0.00690	0.0752	10%	-	-	-	-
17	Fraction Catholic	0.00307	0.593	-	-	-	-	-
18	Civil Liberties	−0.00242	0.322	-	-	-	-	-
19	Fraction Hindu	−0.0967	0.000540	1%	1%	1%	1%	1%
20	Political Rights	0.000162	0.934	-	-	-	-	-
21	Primary Exports, 1970	−0.00550	0.421	-	-	-	-	-
22	Exchange rate distortions	−0.00002	0.538	-	-	-	-	-
23	Age	−0.000009	0.774	-	-	-	-	-
24	War dummy	−0.00144	0.548	-	-	-	-	-
25	Size labor force	0.0000003	0.00411	1%	5%	5%	5%	5%
26	Fraction speaking foreign language	−0.00246	0.468	-	-	-	-	-
27	Fraction of Pop speaking English	−0.00707	0.132	-	-	-	-	-
28	Ethnologic fractionalization	0.0137	0.0122	5%	5%	5%	5%	5%
29	Spanish Colony dummy	0.0131	0.0225	5%	10%	10%	10%	10%
30	SD of black-market premium	−0.000001	0.892	-	-	-	-	-
31	French Colony Dummy	0.00894	0.0378	5%	-	10%	10%	10%
32	Absolute latitude	−0.00009	0.521	-	-	-	-	-
33	Ratio of workers to population	−0.000520	0.945	-	-	-	-	-
34	Higher education enrollment	−0.129	0.00214	1%	5%	5%	5%	5%
35	Population Growth	−0.114	0.609	-	-	-	-	-
36	British Colony dummy	0.00680	0.0726	10%	-	-	-	-
37	Outward orientation	−0.00453	0.0364	5%	-	10%	10%	10%
38	Fraction Jewish	−0.000774	0.942	-	-	-	-	-
39	Revolutions and coups	0.00321	0.504	-	-	-	-	-
40	Public Education Share	0.138	0.250	-	-	-	-	-
41	Area (Scale Effect)	0.0000003	0.638	-	-	-	-	-
42	Intercept	0.0207	0.000	1%	1%	1%	1%	1%

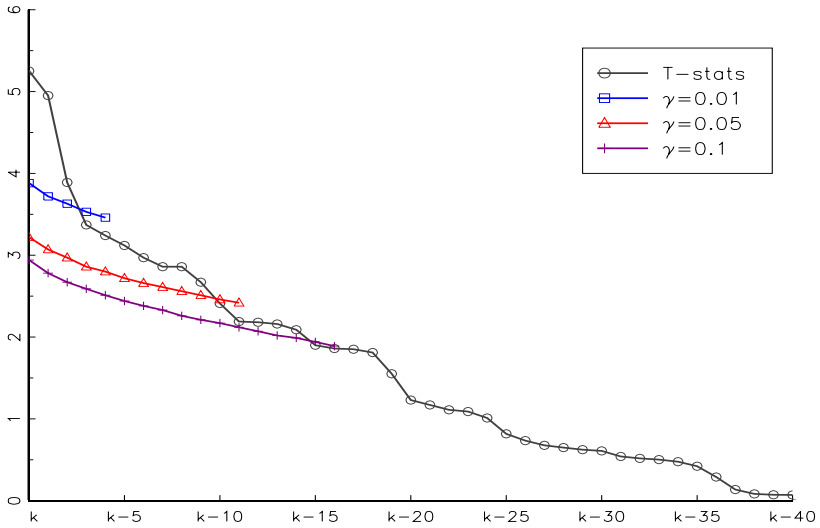
For every regressor in the FLS data set this table shows whether the variable is found to be significant when controlling the FDR at the indicated level. The procedures are described in section 3. We work with 5000 bootstrap iterations. In the Storey approach $\lambda = 0.5$.

5.1 FLS

5.1.1 Data

The FLS data set is a version of that used in Sala-i-Martin (1997). The latter includes $n = 140$ countries for which GDP growth is computed over the period 1960-1992, and 62 regressors. Fernandez *et al.* (2001) only include those countries for which observations on all 25 variables flagged important by Sala-i-Martin (1997) are available. This amounts to $n = 72$. They then add all variables which do not induce any missing observations in these countries. This yields 41 variables. Using this data set allows us to compare our findings to those of Fernandez *et al.* (2001), Sala-i-Martin (1997), Hendry and Krolzig (2004) and Ley and Steel (2009).

Figure 1: Test statistics and bootstrap critical values for the FLS data



The test statistics T_i and the relevant bootstrap critical values $\hat{c}_{\gamma,j}$ (i.e. those until and around $T_i < \hat{c}_{\gamma,i}$) for different γ plotted against k, the ranks of the T_i.

5.1.2 Results

Table 4 shows the results of the application of the different techniques to the FLS data set. The results reported are obtained with usual OLS standard errors. We also calculate results using the heteroscedasticity robust standard errors HC_2 and HC_3 (MacKinnon and White, 1985). Results using HC_2 are similar to the OLS results; using HC_3 resulted in substantially fewer rejections.[12]

Classical testing declares 20 variables (excluding the intercept) to be significant at the 10% level, 15 at the 5% level and 9 at the 1% significance level. The picture is different when accounting for multiple testing by controlling the FDR. The most rejective and still reliable procedure according to the Monte Carlo study, the bootstrap method, declares 15 variables significant at $\gamma = 0.1$, 10 at $\gamma = 0.05$ and when controlling the FDR at the 1% level, only 3 variables are found to be significantly related to GDP growth. (See Figure 1 for a graphical illustration.) We will elaborate on the three variables later. Hence, this simple comparison reveals that several variables appear to be only spuriously declared significant because of the large number of hypothesis tests performed. Unreported additional simulations[13] as in section 4 with 10 (out of 50) false hypotheses find an FDR of classical testing of around 1/3. To the extent that the MC study of sec. 4 is representative for the present empirical application, we should therefore expect roughly 1 false for every 3 rejections when testing at $\alpha = 0.1$. This would mean that we should expect around $14(= 20 - 6)$ of the 20 rejections to be correct. This is in line with the bootstrap results, where around

[12]Performing a White type test for heteroscedasticity would be pointless here, since the null of no heteroscedasticity would never be rejected because of the high number of variables included in the model. The 95% critical value of the 800 degrees of freedom test is 866.9, whereas $n \cdot R^2$ of the test regression is bounded above by 72. **For the referees: Please refer to Tables A.7 and A.8.**

[13]**For the referees: Please refer to Table A.1.**

Table 5: Comparison FLS Data Set

	Regressor	Bootstrap	BMA posterior prob.	Sala-i-Martin	Hendry & Krolzig
1	GDP level 1960	1%	1.000	1.000^{**}	yes
2	Fraction Confucian	1%	0.995	1.000^{*}	yes
3	Life expectancy	5%	0.946	0.999^{**}	yes
4	Equipment investment	5%	0.942	1.000^{*}	yes
5	Sub-Saharan dummy	5%	0.757	0.997^{*}	yes
6	Fraction Muslim	-	0.656	1.000^{*}	-
7	Rule of Law	-	0.516	1.000^{*}	-
8	Number of years open economy	-	0.502	1.000^{*}	yes
9	Degree of Capitalism	-	0.471	0.987^{*}	-
10	Fraction Protestant	-	0.461	0.966^{*}	-
11	Fraction GDP in mining	5%	0.441	0.994^{*}	yes
12	Non-Equipment investment	-	0.431	0.982^{*}	-
13	Latin American dummy	10%	0.190	0.998^{*}	yes
14	Primary School Enrollment, 1960	10%	0.184	0.992^{**}	yes
15	Fraction Buddhist	-	0.167	0.964^{*}	-
16	Black-market premium	-	0.157	0.825	-
17	Fraction Catholic	-	0.110	0.963^{*}	-
18	Civil Liberties	-	0.100	0.997^{*}	-
19	Fraction Hindu	1%	0.097	0.654	yes
20	Political Rights	-	0.071	0.990^{*}	-
21	Primary Exports, 1970	-	0.069	0.998^{*}	-
22	Exchange rate distortions	-	0.060	0.968^{*}	-
23	Age	-	0.058	0.903	-
24	War dummy	-	0.052	0.984^{*}	-
25	Size labor force	5%	0.047	0.835	yes
26	Fraction speaking foreign language	-	0.047	0.831	-
27	Fraction of Pop speaking English	-	0.047	0.910^{*}	-
28	Ethnologic fractionalization	5%	0.035	0.643	yes
29	Spanish Colony dummy	10%	0.034	0.938^{*}	yes
30	SD of black-market premium	-	0.031	0.993^{*}	-
31	French Colony Dummy	10%	0.031	0.702	yes
32	Absolute latitude	-	0.024	0.980^{*}	-
33	Ratio of workers to population	-	0.024	0.766	-
34	Higher education enrollment	5%	0.024	0.579	yes
35	Population Growth	-	0.022	0.807	-
36	British Colony dummy	-	0.022	0.579	yes
37	Outward orientation	10%	0.021	0.634	-
38	Fraction Jewish	-	0.019	0.747	-
39	Revolutions and coups	-	0.017	0.995^{*}	-
40	Public Education Share	-	0.016	0.580	-
41	Area (Scale Effect)	-	0.016	0.532	-

"Bootstrap" denotes significance when controlling the FDR at the indicated level; "BMA post. prob." denotes the marginal posterior probability of inclusion found in Fernandez *et al.* (2001); "Sala-i-Martin" shows the CDF(0), explained in section 5.1.3, found in Sala-i-Martin (1997), where $**$ indicate variables always included and $*$ indicate variables found significantly related to growth; "Hendry and Krolzig" shows a "yes" if the variable is included following the procedure in Hendry and Krolzig (2004).

$(1-0.1)\times 15 \approx 13-14$ of the rejections can be expected to be correct. Of course, such corrective reasoning for classical testing hinges on Monte Carlo designs that may or may not yield estimated FDRs representative for the empirical study under consideration. A key advantage of multiple testing techniques is that FDR control obtains under much more general conditions, as evidenced by our and many other studies and theoretical contributions.

Here, the bootstrap method does not lead to more rejections than the Storey procedure or the BKY algorithm. Although the Monte Carlo study draws a somewhat different picture, one should note that the power differences were modest. The three variables found most robustly related to per capita growth, i.e. the ones with significance at $\gamma = 0.01$, are "GDP level 1960", "Fraction Confucian" and "Fraction Hindu" (Table 4). The rationale for the impact of initial GDP on economic growth is well established since the work of Solow (1956) and Swan (1956). The signifi-

cance of the two other variables might indicate that some religions support a growth stimulating environment (Confucian) and some harm it (Hindu). On the other hand, these variables might also capture regional heterogeneity, since the data set contains no regional dummies.

5.1.3 Comparison to Previous Studies

Fernandez *et al.* (2001) Table 5 compares our findings to those of Fernandez *et al.* (2001). To be able to better compare the findings, we briefly explain the concept of marginal inclusion probability, which is key in Fernandez *et al.* (2001) to judge the importance of a variable in a regression.

Fernandez *et al.* (2001) use BMA techniques to account for the model uncertainty in the growth regression context. Given the 41 possible explanatory variables there are 2^{41} possible correct models, assuming that these nest the true DGP. One now has to assign a 'prior distribution' to each model as well as to the inclusion of a certain variable in each model. We will elaborate on the influence of these choices in section 5.1.3. Using Bayes' law Fernandez *et al.* (2001) then compute the 'posterior inclusion probability' of each of the 2^{41} possible models: "The marginal posterior probability of including a certain variable is simply the sum of the posterior probabilities of all models that contain this regressor." The Bayesian framework provides no cut-off value for the marginal inclusion probability of a variable to declare it (non-)significant. Nevertheless, one gets a good feeling of the importance of each variable.

Table 5 shows that the five variables with the highest marginal posterior probability of inclusion in Fernandez *et al.* (2001) are also significant according to the bootstrap method at $\gamma = 0.05$. But there are also variables which are significantly related to growth using our approach which have a marginal posterior probability below 0.1. For example, "Fraction Hindu" is significant at $\gamma = 0.05$ while its marginal posterior probability is only 0.097. Hence, there are non-negligible differences between the two approaches.

Sala-i-Martin (1997) We now compare the results of the FLS data set with those of Sala-i-Martin (1997). His approach is as follows: In every regression run, there are three sets of variables, $\mathbf{w}$, v and $\mathbf{q}$. $\mathbf{w}$ consists of the three variables, namely "GDP level in 1960", "Life expectancy" and "Primary school enrolment". These three variables are included in every regression. v is the variable under study and $\mathbf{q}$ is a trio of the other remaining 58 variables. He then computes the cumulative distribution function (CDF) of the coefficient corresponding to v at zero by averaging the coefficient of v obtained in the regressions for all combinations of variables in $\mathbf{q}$.[14] Since a variable can have a positive or negative coefficient he sets $\text{CDF}(0) = \max\{\text{CDF}(0), 1 - \text{CDF}(0)\}$. He then flags a variable significantly related to growth if its $\text{CDF}(0) \leq 0.91$. The outcome of this approach depends on a large set of misspecified regressions, in particular via the omitted variable bias inherent when conducting many overly short regressions.

[14] This results in 455,126 different models. The two million regressions in his title arise because any model is counted four times, since he distinguishes whether a variable is in v or in $\mathbf{q}$.

Sala-i-Martin (1997) finds the same five variables to be significant that were already significant in both the bootstrap approach and in Fernandez *et al.* (2001) (Table 5). But beyond that, there seems to be little agreement concerning significance of variables in Sala-i-Martin (1997) and in our study. Out of the 25 variables[15] flagged significant by Sala-i-Martin (1997), we can only confirm 9 using an FDR up to 10%.

Hendry and Krolzig (2004) Hendry and Krolzig (2004) also use the FLS data. They argue that if the general unrestricted model (GUM) provided by all available explanatory variables nests a good approximation of the DGP, then a "General to simple" (Gets) selection based approach would suffice to find the best model. Of course, this abstracts from problems like data accuracy, exogeneity of regressors or constancy of the parameters across observations. But this abstraction is inherent to all approaches presented here, including the one we propose. Hendry and Krolzig (2004) further note that if all regressors were mutually orthogonal, then selection based on the ordered squared t-statistics from the GUM, say $t^2_{(1)} \geq t^2_{(2)} \geq \ldots \geq t^2_{(k)}$, could be validly used for model selection. One would then declare variables $1, \ldots, \tilde{k}$ with $t^2_{(\tilde{k})} \geq c_\alpha$ significant and the remaining ones insignificant, where $t^2_{(\tilde{k}+1)} \leq c_\alpha$. They choose $\alpha = 0.025$ so as to only select one variable by chance out of the $k = 41$ regressors. Since t-values in fact are not mutually orthogonal, a multi-path search introduced in Hoover and Perez (1999) is used. This approach is somewhat similar to the one employed here, where decisions are also based on individual t-statistics and their dependence is taken into account using the bootstrap approach described in section 3.3. The bootstrap approach may be more transparent for the user—all one has to choose is some γ. Transparent cutoffs $\hat{c}_j$ are then provided, and no multi-path search is needed.

The similarity of the two approaches can be seen when comparing the variables found significant (Table 5). The number of significant variables in Hendry and Krolzig (2004) is 16 compared to 15 for our approach, out of which 14 coincide. (A J-test rejects both models, although Hendry and Krolzig's model is rejected only at larger significance values.) Again, the five variables already jointly significant in Fernandez *et al.* (2001) and Sala-i-Martin (1997) are also significant in Hendry and Krolzig (2004). Hence, we are quite confident regarding the importance of "GDP level 1960", "Fraction Confucian", "Life expectancy", "Equipment investment" and "Sub-Saharan dummy".

Ley and Steel (2009) and Eicher *et al.* (2010) Results from BMA techniques to account for model uncertainty in growth regression are potentially sensitive to user choices. This issue is identified among others in Ley and Steel (2009). They consider BMA on linear regression models M_j with $0 \leq k_j \leq k$ regressors grouped in $\boldsymbol{X}_j$ leading to

$$\boldsymbol{y}|a, \boldsymbol{\beta}_j, \sigma \sim \mathrm{N}(a\boldsymbol{\iota} + \boldsymbol{X}_j\boldsymbol{\beta}_j, \sigma^2\boldsymbol{I}),$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^{k_j}$ and $\sigma \in \mathbb{R}_+$ is a scale parameter.

One then needs to specify the prior distribution of each parameter for model M_j as well as the prior distribution of the inclusion of model M_j. For the prior density of the parameters, Ley and

[15] He actually finds 22 significant variables, but he also *assumes* significance of the three variables in $\mathbf{w}$.

Steel (2009) use a combination of a so called 'non-informative' improper priors on a and σ as well as a 'g-prior' on $\boldsymbol{\beta}$. Ley and Steel (2009) advocate choosing $g = 1/\max\left\{n, k^2\right\}$. The prior model probabilities are often specified as $P(M_j) = \theta^{k_j}(1-\theta)^{k-k_j}$, assuming that each regressor enters a model with equal probability and independently of the others. Fernandez *et al.* (2001) choose $\theta = 0.5$, implying an expected model size of $m = k/2$. Ley and Steel (2009) also consider choosing θ randomly. In that case, one first fixes the expected model size m and then determines the interval in which θ can vary.

Ley and Steel (2009) show that when using the above priors in the FLS data set, the posterior mean model size ranges from 6.03 with $m = k/2 = 20.5$, random θ and $g = 1/k^2$ to 19.84 for $m = k/2 = 20.5$, fixed θ and $g = 1/n$. Their Table II shows the ranking of the marginal posterior probability of including a certain variable to also be highly sensitive to the prior settings. Comparing cases with $g = 1/n$ and fixed θ they note: "Fraction Hindu, the Labor force size, and Higher education enrolment go from virtually always included with $m = 20.5$ to virtually never included with $m = 7$."

Recent related work by Eicher *et al.* (2010) shows that twelve different plausible 'default' priors can lead to rather different growth models using the FLS data, with "as few as three and as many as 22 regressors" being found related to growth. They recommend a 'unit information prior'. It will be interesting to see whether the BMA literature will henceforth adopt this choice, or whether different models continue to be put forward using different variants of BMA.

These findings might help explain the differences between the bootstrap approach and the marginal posterior inclusion probabilities found in Fernandez *et al.* (2001). Moreover, it might also be a reason for the differences between the latter and the variable selection in Sala-i-Martin (1997). The findings of Ley and Steel (2009) imply that the robustness of BMA techniques must be interpreted with care. A transparent and robust selection of variables might be obtained using e.g. the FDR controlling bootstrap approach presented here.

5.2 MP Data Set

5.2.1 Data

The data set covers 93 countries, for which average GDP growth was calculated from 1960 to 1992. It was developed and first used in Masanjala and Papageorgiou (2005). The data set consists of 32 basic variables out of which 22 are also combined with an interaction dummy for African countries. Thus, $k = 54$ here. To tackle the endogeneity issues Masanjala and Papageorgiou (2005) devote careful attention to only including predetermined variables. Furthermore, the interaction dummies address possible parameter heterogeneity.

Table 6: MP Data Set

	Regressor	$\hat{\beta}_i$	p-value	Classical	BH	Storey	BKY	Boot
1	ln GDP per capita, 1960	−1.80	0.000960	1%	5%	5%	5%	5%
2	Life expectancy, 1960	0.196	0.000101	1%	1%	1%	1%	1%
3	A∗Mining	10.7	0.000504	1%	1%	5%	1%	1%
4	South-East Asia	2.15	0.0240	5%	-	-	-	-
5	Fraction muslim	2.62	0.0174	5%	-	-	-	-
6	A∗Land area	−0.00113	0.0744	10%	-	-	-	-
7	OECD	1.48	0.0337	5%	-	-	-	-
8	A∗Primary export, 1970	−4.73	0.0141	5%	-	-	-	-
9	Primary export, 1970	0.0783	0.942	-	-	-	-	-
10	A∗European language	3.98	0.278	-	-	-	-	-
11	European language	0.955	0.0575	10%	-	-	-	-
12	Fraction Confucian	3.74	0.0835	10%	-	-	-	-
13	A∗Colony	1.98	0.604	-	-	-	-	-
14	A∗Tropical fraction	−2.90	0.570	-	-	-	-	-
15	A∗Landlocked	−0.199	0.849	-	-	-	-	-
16	Latin America	−0.227	0.778	-	-	-	-	-
17	A∗Labor force, 1960	0.000	0.755	-	-	-	-	-
18	Malaria prevalence, 1960	0.335	0.723	-	-	-	-	-
19	Landlocked	−0.436	0.427	-	-	-	-	-
20	Fraction Catholic	0.835	0.383	-	-	-	-	-
21	A∗Malaria prevalence, 1960	−1.77	0.664	-	-	-	-	-
22	A∗Primary school, 1960	4.39	0.0574	10%	-	-	-	-
23	A∗Fraction Muslim	0.905	0.577	-	-	-	-	-
24	Fraction Protestant	0.534	0.619	-	-	-	-	-
25	A∗ln GDP per capita, 1960	0.300	0.725	-	-	-	-	-
26	Tropical fraction	0.559	0.519	-	-	-	-	-
27	Fraction Buddhist	0.707	0.539	-	-	-	-	-
28	Primary school, 1960	−1.42	0.400	-	-	-	-	-
29	A∗British colony	−0.483	0.887	-	-	-	-	-
30	Sub Saharan Africa	7.68	0.235	-	-	-	-	-
31	A∗Fraction urban pop, 1960	−6.49	0.105	-	-	-	-	-
32	A∗Life expectancy, 1960	−0.0356	0.666	-	-	-	-	-
33	A∗Secondary school, 1960	1.13	0.961	-	-	-	-	-
34	Mining	−0.319	0.892	-	-	-	-	-
35	Fraction Hindu	2.82	0.0957	10%	-	-	-	-
36	A∗Fraction Catholic	−2.43	0.753	-	-	-	-	-
37	Distance from equator	−0.00667	0.800	-	-	-	-	-
38	Land area	0.000161	0.0583	10%	-	-	-	-
39	Colony	−1.21	0.258	-	-	-	-	-
40	A∗Ethnoliguistic fractionalization	−0.559	0.645	-	-	-	-	-
41	Spanish Colony	0.895	0.324	-	-	-	-	-
42	Secondary school, 1960	0.725	0.622	-	-	-	-	-
43	A∗French colony	−1.03	0.751	-	-	-	-	-
44	A∗Distance from equator	−0.103	0.142	-	-	-	-	-
45	Tertiary education, 1960	−4.50	0.401	-	-	-	-	-
46	British colony	0.554	0.582	-	-	-	-	-
47	A∗Fraction Protestant	−2.73	0.639	-	-	-	-	-
48	A∗Fraction English speaking	−3.71	0.847	-	-	-	-	-
49	French colony	0.861	0.369	-	-	-	-	-
50	Ethnoliguistic fractionalization	−0.107	0.876	-	-	-	-	-
51	Fraction English speaking	0.281	0.693	-	-	-	-	-
52	Labor force, 1960	−0.000002	0.810	-	-	-	-	-
53	Fraction Jewish	2.00	0.166	-	-	-	-	-
54	Fraction urban pop, 1960	0.0889	0.924	-	-	-	-	-
	Intercept	1.73	0.000	1%	1%	1%	1%	1%

For every regressor in the MP data set this table shows whether the variable is found to be significant when controlling the FDR at the indicated level. The procedures are described in section 3. We work with 5000 bootstrap iterations. In the Storey approach $\lambda = 0.5$.

5.2.2 Results

Table 6 shows the results when accounting for multiple testing by controlling the FDR in the MP data set.[16] Only three variables are found to be significantly related to GDP growth, namely "ln GDP per capita, 1960" (at $\gamma = 0.05$), "Life expectancy"(at $\gamma = 0.01$) and "A∗Mining"(at $\gamma = 0.01$) using the bootstrap method.[17] Again the difference to the number of significant

[16]We also calculate robust standard errors. The results are qualitatively the same as for the FLS data set.

[17]For the MP data set, the BH method and the BKY algorithm yield the same results; the Storey method finds significance of "A*Mining" only at a FDR of 5%.

Table 7: Countries Included

Argentina	Finland	Mexico	Spain
Australia	France	Morocco	Sri Lanka
Austria	Guatemala	Netherlands	Sweden
Belgium	Honduras	New Zealand	Switzerland
Bolivia	Iceland	Nicaragua	Thailand
Brazil	India	Nigeria	Trinidad & Tobago
Canada	Ireland	Norway	Turkey
Colombia	Israel	Pakistan	Uganda
Costa Rica	Italy	Panama	United Kingdom
Denmark	Japan	Peru	United States
Egypt	Kenya	Philippines	Uruguay
El Salvador	Luxembourg	Portugal	Venezuela
Ethiopia	Mauritius	South Africa	

variables using classical testing is substantial. Classical testing finds 13 significant variables at $\alpha = 0.1$. Compared to the FLS data, a possible reason for the larger difference between classical testing and FDR controlling testing is the larger number of explanatory variables. This results in stricter FDR controlling tests. (Of course, it is also simply other data.)

5.2.3 Comparison to Previous Studies

Masanjala and Papageorgiou (2005) also use BMA techniques to account for model uncertainty. Out of their three variables with a marginal posterior probability of inclusion of 100%, we can confirm two. On the other hand, "A*Mining" only has marginal posterior probability of 76.1, yet it is found significant by the FDR controlling techniques. Hence, again the BMA results differ somewhat from the multiple testing results.[18] Ley and Steel (2009) also use this data set to investigate the influence of the priors on the outcome of BMA. Again, depending on the different choices, the average posterior model size ranges from 5.42 ($m = 7$, random θ, $g = 1/k^2$) to 17.90 ($m = 27$, fixed θ, $g = 1/n$).

6 Output Convergence Revisited

We now employ the FDR controlling techniques described in section 3 to a data set of $n = 51$ countries ranging from 1950 to 2003, so $T = 54$. The resulting number of different country pairs is 1275. The data is from the Penn World Tables Version 6.2 (Heston, Summers and Aten, 2006) and includes all countries for which data on per-capita output was available for the indicated time span.[19] We employ standard ADF tests and ADF-GLS tests. The deterministic component also includes a time trend. We choose the lag length p according to $p = 5(T/100)^{1/4}$ and $p = 6(T/100)^{1/4}$ as those yielded the highest number of right rejections while still controlling the FDR for the bootstrap procedure when $T = 50$, which is close to the actual T. For the present data this results in $p = 4$ and $p = 5$. Critical values and p-values are adjusted to sample size.

[18] **For the referees:** For a comparison of the marginal posterior probability to our results refer to Table A.9.

[19] Table 7 gives a complete list of countries included in the analysis.

Table 8: Pairwise Test for Output convergence

OLS detrending			GLS detrending		
$p = 4$					
$\alpha = \gamma = 0.01$					
	# pairs	% of pairs		# pairs	% of pairs
Classical approach	2	0.16	Classical approach	3	0.24
BH	0	0	BH	0	0
BKY	0	0	BKY	0	0
Bootstrap	0	0	Bootstrap	0	0
$\alpha = \gamma = 0.05$					
	# pairs	% of pairs		# pairs	% of pairs
Classical approach	40	3.13	Classical approach	23	1.80
BH	0	0	BH	0	0
BKY	0	0	BKY	0	0
Bootstrap	0	0	Bootstrap	0	0
$\alpha = \gamma = 0.1$					
	# pairs	% of pairs		# pairs	% of pairs
Classical approach	83	6.51	Classical approach	47	3.68
BH	0	0	BH	0	0
BKY	0	0	BKY	0	0
Bootstrap	0	0	Bootstrap	0	0
$p = 5$					
$\alpha = \gamma = 0.01$					
	# pairs	% of pairs		# pairs	% of pairs
Classical approach	10	0.78	Classical approach	8	0.63
BH	0	0	BH	0	0
BKY	0	0	BKY	0	0
Bootstrap	0	0	Bootstrap	0	0
$\alpha = \gamma = 0.05$					
	# pairs	% of pairs		# pairs	% of pairs
Classical approach	45	3.53	Classical approach	39	3.06
BH	0	0	BH	0	0
BKY	0	0	BKY	0	0
Bootstrap	0	0	Bootstrap	0	0
$\alpha = \gamma = 0.1$					
	# pairs	% of pairs		# pairs	% of pairs
Classical approach	96	7.53	Classical approach	73	5.49
BH	0	0	BH	0	0
BKY	0	0	BKY	0	0
Bootstrap	0	0	Bootstrap	0	0

"# pairs" shows the number of country pairs for which the null of a unit root is rejected. "% of pairs" shows the proportion of those pairs compared to the total number of pairs. The procedures are described in section 3. We use 5000 bootstrap iterations. In the Storey approach, $\lambda = 0.5$.

Table 4.2 shows the results of applying the FDR controlling techniques to testing for pairwise convergence. We corroborate the results of Pesaran (2007a) that the null of no convergence is only rejected for a fraction of countries less than or equal to the individual significance level of the unit root test applied. When accounting for the multiplicity of tests performed, for no level γ do we find any rejection of the null; this also holds in particular for the most powerful FDR controlling procedure, the bootstrap method using either OLS or GLS detrending.[20]

At first glance, this finding might seem a bit surprising. But this finding clarifies the results in Pesaran (2007a).[21] His approach does not allow to say whether rejection of the null for some country pairs was spurious or not. Finding no converging pairs when employing a more appropriate testing framework for individual tests (rather than for fractions of rejections), the confidence in Pesaran's no time-series convergence finding is strengthened.

Nevertheless, we emphasize that cross-country convergence was tested using a very strict defini-

[20] As Pesaran (2007a), we also perform analogous exercises for subgroups like European or American countries. Again, our findings coincide with his in that we do not find individual converging country pairs.

[21] As we have shown in the Monte Carlo study, we also control the FWER at γ. Hence, the probability of even a single false rejection is bounded by γ.

tion (Islam, 2003). We do not rule out cross-country convergence using for example conditional definitions nor do we make any statements about 'within convergence' here, i.e. whether countries converge to a steady state output. Indeed we found evidence for 'within convergence' in section 5 as initial GDP was always included in the final model. What is more, our empirical results are consistent with recent theoretical work of Pampel (2009), who shows that the two notions of convergence may well conflict under plausible assumptions. Hence, we only claim that convergence across economies using a time series definition with the necessary condition of no unit root in the log per-capita output gap of two economies does not appear to hold.

7 Conclusion

This paper highlights the importance of accounting for the multiplicity of tests performed in two applications to growth econometrics. Controlling the FDR, we have shown how to robustly select explanatory variables in cross-sectional growth regressions. Among others, this was done using a bootstrap approach which takes the dependence structure of the test statistics into account and thus has high power. The outcome of other approaches in cross-sectional growth regressions, such as Bayesian Model Averaging, is sensitive to user choices and hence may not be robust. In addition to the robustness of our bootstrap approach, we believe the selection of variables using FDR controlling techniques to be very transparent for the user. The only choice to make is the level at which one wants to control the FDR.

The actual variables selected support the neoclassical growth theory and its implied convergence of countries to a steady state output. The second application investigates cross-country convergence using pairwise unit-root tests. Controlling the FDR, we find no evidence of this type of convergence. This clarifies the results of Pesaran (2007a), whose framework does not allow to say whether the fraction of rejected pairs is spurious or not.

There are more fields in econometrics where accounting for multiplicity is important. For example, modeling returns to schooling involves a large number of candidate explanatory variables. As we have shown in the application to cross-sectional growth regressions, classical significance tests produce a non-negligible number of variables which appear to be only spuriously significant because of the large number of tests performed. As such, the techniques studied here may prove fruitful in many other literatures.

References

Arellano M, Bond S. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* **58**: 277–297.

Barro RJ. 1991. Economic growth in a cross section of countries. *The Quarterly Journal of Economics* **106**: 407–443.

Barro RJ, Sala-i-Martin XX. 1995. *Economic Growth*. McGraw Hill.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B* **57**: 289–300.

Benjamini Y, Krieger AM, Yekutieli D. 2006. Adaptive linear step-up procedure that control the false discovery rate. *Biometrika* **93**: 491–507.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**: 1165–1188.

Bernard A, Durlauf S. 1995. Convergence in international output. *Journal of Applied Econometrics* **10**: 97–108.

Burridge P, Taylor AMR. 2004. Bootstrapping the HEGY seasonal unit root tests. *Journal of Econometrics* **123**: 67–87.

Campbell Y, Mankiw N. 1989. International evidence on the persistence of economic fluctuations. *Journal of Monetary Economics* **23**: 319–333.

Demetrescu M, Hassler U, Kuzin V. 2008. Pitfalls of post-model-selection testing: Experimental quantification. *mimeo* .

Dudoit S, van der Laan MJ. 2007. *Multiple Testing Procedures and Applications to Genomics*. Springer Series in Statistics, Berlin: Springer.

Durlauf SN, Kourtellos A, Tan CM. 2008. Are any growth theories robust? *The Economic Journal* **118**: 329–346.

Eicher TS, Papageorgiou C, Raftery AE. 2010. Default priors and predictive performance in bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics* .

Elliot G, Rothenberg T, Stock J. 1996. Efficient tests for an autoregressive unit root. *Econometrica* **64**: 813–836.

Fernandez C, Ley E, Steel M. 2001. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* **16**: 563–576.

Granger CWJ, Morris MJ. 1976. Time Series Modelling and Interpretation. *Journal of the Royal Statistical Society. Series A (General)* **139**: 246–257.

Hauk Jr WR, Wacziarg R. 2009. A Monte Carlo study of growth regressions. *Journal of Economic Growth* **14**: 103–147.

Hendry DH, Krolzig HM. 2004. We ran one regression. *Oxford Bulletin of Economics and Statistics* **66**: 799–810.

Heston A, Summers R, Aten B. 2006. Penn World Table Version 6.2, center for International Comparisons of Production, Income and Prices at the University of Pennsylvania, September.

Hoover KD, Perez SJ. 1999. Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *Econometrics Journal* **2**: 167–191.

Islam N. 2003. What have we learnt from the convergence debate? *Journal of Economic Surveys* **17**: 312–355.

Leeb H, Pötscher B. 2008. Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* **24**: 338–367.

Levine R, Renelt D. 1992. A sensitivity analysis of cross-country growth regressions. *American Economic Review* **82**: 942–963.

Ley E, Steel M. 2009. On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics* **24**: 651–674.

Lovell MC. 1983. Data mining. *The Review of Economics and Statistics* **65**: 1–12.

MacKinnon JG, White H. 1985. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* **29**: 305–325.

Masanjala W, Papageorgiou C. 2005. Initial conditions, European colonialism and Africas growth. *Mimeo, Department of Economics, Louisiana State University, Baton Rouge* .

Pampel T. 2009. On the dynamics of basic growth models: Ratio stability vs. convergence and divergence in state space. *German Economic Review* **10**: 384–400.

Pesaran M. 2007a. A pair-wise approach to testing for output and growth convergence. *Journal of Econometrics* **138**: 312–355.

Pesaran MH. 2007b. A simple panel unit root test in the presence of cross section dependence. *Journal of Applied Econometrics* **22**: 265–312.

Phillips PCB. 1987. Towards a unified asymptotic theory for autoregression. *Biometrika* **74**: 535–547.

Quah D. 1990. International patterns of growth I: Persistence in cross-country disparities, working Paper, MIT.

Romano JP, Shaikh AM, Wolf M. 2008a. Control of the false discovery rate under dependence using the bootstrap

and subsampling. *Test* **17**: 417–442.

Romano JP, Shaikh AM, Wolf M. 2008b. Formalized data snooping based on generalized error rates. *Econometric Theory* **24**: 404–447.

Sala-i-Martin XX. 1997. I just ran two million regressions. *American Econimc Review* **87**: 178–183.

Sala-i-Martin XX, Doppelhofer G, Miller R. 2004. Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* **94**: 813–835.

Schwert G. 1989. Tests for unit roots: A Monte Carlo investigation. *Journal of Businnes and Economic Statistics* **7**: 5–17.

Smeekes S. 2009. Detrending bootstrap unit root tests. METEOR Research Memorandum 09/56, Maastricht University.

Solow RM. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* **70**: 65–94.

Storey JD, Taylor JE, Siegmund D. 2004. Strong control, conservative point estimation and siumltaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society. Series B* **66**: 187–205.

Swan TW. 1956. Economic growth and capital accumulation. *Economic Record* **32**: 334–361.

Appendix A Additional Tables—Not For Publication

Table A.1: Linear regression model with 10 false hypotheses

	# false hypotheses: 10		Sample size: 100		# Regressors: 50	
	$\rho = 0$					
	1%		5%		10%	
	FDR	Right Rejections	FDR	Right Rejections	FDR	Right Rejections
Classical	0.039	7.94	0.156	9.26	0.268	9.61
BH	0.008	5.39	0.040	7.68	0.079	8.50
Storey	0.011	5.65	0.053	7.88	0.101	8.66
BKY	0.009	5.52	0.045	7.78	0.087	8.57
Bootstrap	0.011	5.50	0.050	7.89	0.101	8.67
	$\rho = 0.3$					
	1%		5%		10%	
	FDR	Right Rejections	FDR	Right Rejections	FDR	Right Rejections
Classical	0.047	6.19	0.176	8.27	0.284	8.95
BH	0.008	2.96	0.038	5.45	0.076	6.77
Storey	0.011	3.22	0.053	5.76	0.104	7.05
BKY	0.009	3.03	0.042	5.54	0.083	6.83
Bootstrap	0.010	3.10	0.046	5.71	0.096	7.03
	$\rho = 0.5$					
	1%		5%		10%	
	FDR	Right Rejections	FDR	Right Rejections	FDR	Right Rejections
Classical	0.062	4.57	0.199	6.90	0.299	8.00
BH	0.006	1.41	0.036	3.27	0.072	4.63
Storey	0.009	1.61	0.050	3.57	0.097	4.97
BKY	0.006	1.42	0.037	3.30	0.073	4.62
Bootstrap	0.008	1.52	0.044	3.42	0.091	4.74

This table shows the results of a Monte Carlo simulation as specified in section 4.1 with 2000 simulations and the indicated parameter settings. The procedures controlling the FDR were applied as described in section 3. For the Storey procedure, $\lambda = 0.5$.

Table A.2: Linear regression model with 25 false hypotheses

	# false hypotheses: 25		Sample size: 100		# Regressors: 50	
	$\rho = 0$					
	1%		5%		10%	
	FDR	Right Rejections	FDR	Right Rejections	FDR	Right Rejections
Classical	0.011	19.83	0.048	23.11	0.090	24.01
BH	0.005	16.67	0.025	21.61	0.050	23.07
Storey	0.001	18.69	0.051	22.86	0.100	23.93
BKY	0.007	17.88	0.043	22.63	0.087	23.81
Bootstrap	0.008	18.15	0.046	22.80	0.095	23.93
	$\rho = 0.3$					
	1%		5%		10%	
	FDR	Right Rejections	FDR	Right Rejections	FDR	Right Rejections
Classical	0.013	15.39	0.054	20.61	0.098	22.38
BH	0.004	10.18	0.026	17.34	0.050	20.10
Storey	0.010	12.78	0.051	19.68	0.102	22.01
BKY	0.006	11.03	0.039	18.74	0.080	21.37
Bootstrap	0.008	11.32	0.042	19.31	0.087	21.82
	$\rho = 0.5$					
	1%		5%		10%	
	FDR	Right Rejections	FDR	Right Rejections	FDR	Right Rejections
Classical	0.017	11.32	0.063	17.27	0.103	19.98
BH	0.004	5.16	0.024	11.81	0.048	15.77
Storey	0.008	7.21	0.047	14.67	0.093	18.65
BKY	0.004	5.49	0.032	12.78	0.065	17.02
Bootstrap	0.006	5.60	0.034	13.18	0.075	17.27

This table shows the results of a Monte Carlo simulation as specified in section 4.1 with 2000 simulations and the indicated parameter settings. The procedures controlling the FDR were applied as described in section 3. For the Storey procedure, $\lambda = 0.5$.

Table A.3: Pairwise unit root tests with 3 false hypotheses and $T = 50$

$n = 10$[45pairs]	$T = 50$		false hypotheses: 3[3pairs]		
OLS detrending			GLS detrending		
$p = 0$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.432	2.98	Classical approach	0.431	2.98
BH	0.311	2.57	BH	0.326	2.64
Storey	0.400	2.54	Storey	0.343	2.52
BKY	0.314	2.56	BKY	0.328	2.64
Bootstrap	0.329	2.67	Bootstrap	0.332	2.77
$p = 1$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.355	2.84	Classical approach	0.361	2.78
BH	0.208	1.59	BH	0.228	1.64
Storey	0.393	1.92	Storey	0.380	1.94
BKY	0.209	1.59	BKY	0.227	1.64
Bootstrap	0.211	1.55	Bootstrap	0.225	1.77
$p = 3$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.318	1.99	Classical approach	0.329	1.95
BH	0.0809	0.338	BH	0.116	0.458
Storey	0.330	1.18	Storey	0.327	1.21
BKY	0.0791	0.346	BKY	0.112	0.484
Bootstrap	0.0768	0.409	Bootstrap	0.110	0.444
$p = 4$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.295	1.53	Classical approach	0.319	1.56
BH	0.0556	0.185	BH	0.0906	0.253
Storey	0.317	1.04	Storey	0.319	1.06
BKY	0.0523	0.189	BKY	0.0891	0.271
Bootstrap	0.0610	0.211	Bootstrap	0.0593	0.231
$p = 5$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.305	1.26	Classical approach	0.318	1.27
BH	0.0486	0.137	BH	0.0569	0.147
Storey	0.322	1.04	Storey	0.302	0.936
BKY	0.0460	0.153	BKY	0.0552	0.156
Bootstrap	0.0434	0.138	Bootstrap	0.0430	0.157
$p = 10$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.277	0.544	Classical approach	0.318	0.512
BH	0.0313	0.0340	BH	0.0517	0.0530
Storey	0.220	0.594	Storey	0.232	0.492
BKY	0.0279	0.0510	BKY	0.0464	0.0680
Bootstrap	0.0176	0.0360	Bootstrap	0.0202	0.0260

This table shows the results of a Monte Carlo simulation as specified in section 4.2 with 1000 simulations and the indicated parameter settings. Tests were conducted for $\alpha = \gamma = 0.05$. The FDR controlling procedures applied are described in section 3. For the Storey procedure, $\lambda = 0.5$.

Table A.4: Pairwise unit root tests with 3 false hypotheses and $T = 100$

$n = 10$[45pairs]	$T = 100$		false hypotheses: 3[3pairs]		
OLS detrending			GLS detrending		
$p = 0$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.452	3.00	Classical approach	0.425	3.00
BH	0.363	3.00	BH	0.337	2.96
Storey	0.449	3.00	Storey	0.364	2.94
BKY	0.366	3.00	BKY	0.340	2.96
Bootstrap	0.360	3.00	Bootstrap	0.335	2.96
$p = 1$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.390	3.00	Classical approach	0.350	2.94
BH	0.281	2.99	BH	0.251	2.64
Storey	0.457	2.98	Storey	0.427	2.71
BKY	0.284	2.99	BKY	0.254	2.64
Bootstrap	0.280	2.98	Bootstrap	0.260	2.68
$p = 4$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.261	2.90	Classical approach	0.286	2.40
BH	0.140	1.71	BH	0.146	1.33
Storey	0.396	2.10	Storey	0.417	1.88
BKY	0.139	1.70	BKY	0.145	1.33
Bootstrap	0.135	1.83	Bootstrap	0.148	1.39
$p = 5$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.236	2.77	Classical approach	0.268	2.11
BH	0.116	1.19	BH	0.131	0.898
Storey	0.384	1.78	Storey	0.395	1.62
BKY	0.113	1.18	BKY	0.128	0.903
Bootstrap	0.125	1.35	Bootstrap	0.113	0.917
$p = 6$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.242	2.51	Classical approach	0.261	1.84
BH	0.107	0.780	BH	0.0933	0.603
Storey	0.420	1.61	Storey	0.355	1.35
BKY	0.106	0.784	BKY	0.0915	0.617
Bootstrap	0.101	0.926	Bootstrap	0.0889	0.677
$p = 12$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.205	1.25	Classical approach	0.266	0.879
BH	0.0293	0.112	BH	0.0437	0.116
Storey	0.301	1.02	Storey	0.345	0.983
BKY	0.0291	0.122	BKY	0.0420	0.153
Bootstrap	0.0220	0.154	Bootstrap	0.0248	0.0900

This table shows the results of a Monte Carlo simulation as specified in section 4.2 with 1000 simulations and the indicated parameter settings. Tests were conducted for $\alpha = \gamma = 0.05$. The FDR controlling procedures applied are described in section 3. For the Storey procedure, $\lambda = 0.5$.

Table A.5: Pairwise unit root tests with 10 false hypotheses and $T = 100$

$n = 10$[45pairs]	$T = 100$		false hypotheses: 5[10pairs]		
OLS detrending			GLS detrending		
$p = 0$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.207	10.0	Classical approach	0.175	9.99
BH	0.166	10.0	BH	0.140	9.92
Storey	0.271	10.0	Storey	0.189	9.90
BKY	0.178	10.0	BKY	0.147	9.93
Bootstrap	0.190	10.0	Bootstrap	0.156	9.94
$p = 1$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.169	10.0	Classical approach	0.154	9.86
BH	0.126	9.99	BH	0.115	9.40
Storey	0.328	9.99	Storey	0.311	9.51
BKY	0.137	10.0	BKY	0.123	9.48
Bootstrap	0.132	9.99	Bootstrap	0.132	9.44
$p = 4$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.126	9.67	Classical approach	0.141	7.92
BH	0.0803	7.75	BH	0.0931	5.57
Storey	0.331	8.40	Storey	0.331	7.09
BKY	0.0853	7.97	BKY	0.0974	5.74
Bootstrap	0.0751	8.15	Bootstrap	0.0781	5.76
$p = 5$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.114	9.25	Classical approach	0.127	7.03
BH	0.0743	6.16	BH	0.0760	4.10
Storey	0.305	7.31	Storey	0.325	6.23
BKY	0.0780	6.35	BKY	0.0774	4.24
Bootstrap	0.0671	6.55	Bootstrap	0.0766	4.42
$p = 6$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.111	8.48	Classical approach	0.127	6.28
BH	0.0661	4.12	BH	0.0688	2.86
Storey	0.291	6.08	Storey	0.309	5.33
BKY	0.0675	4.26	BKY	0.0679	2.94
Bootstrap	0.0540	4.69	Bootstrap	0.0496	3.03
$p = 12$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.111	4.25	Classical approach	0.152	3.15
BH	0.0222	0.493	BH	0.0484	0.439
Storey	0.266	3.63	Storey	0.238	2.69
BKY	0.0208	0.518	BKY	0.0467	0.515
Bootstrap	0.0147	0.551	Bootstrap	0.0110	0.315

This table shows the results of a Monte Carlo simulation as specified in section 4.2 with 1000 simulations and the indicated parameter settings. Tests were conducted for $\alpha = \gamma = 0.05$. The FDR controlling procedures applied are described in section 3. For the Storey procedure, $\lambda = 0.5$.

Table A.6: Pairwise unit root tests for $n = 20$ (190 pairs)

$n = 20$[190pairs]			$T = 50$		
false hypotheses: 4[6pairs]			false hypotheses: 10[45pairs]		
$p = 3$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.509	3.91	Classical approach	0.159	29.3
BH	0.137	0.446	BH	0.0904	8.06
Storey	0.370	2.12	Storey	0.290	21.0
BKY	0.131	0.469	BKY	0.0907	8.27
Bootstrap	0.162	0.466	Bootstrap	0.0709	7.11
$p = 4$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.514	3.06	Classical approach	0.159	22.6
BH	0.0913	0.195	BH	0.0586	2.78
Storey	0.328	1.87	Storey	0.253	15.5
BKY	0.0846	0.229	BKY	0.0578	2.93
Bootstrap	0.0816	0.181	Bootstrap	0.0460	2.44
$p = 5$					
	FDR	Right Rejections		FDR	Right Rejections
Classical approach	0.502	2.59	Classical approach	0.184	18.8
BH	0.0772	0.154	BH	0.0572	1.65
Storey	0.330	1.87	Storey	0.258	14.7
BKY	0.0731	0.216	BKY	0.0566	1.85
Bootstrap	0.0412	0.0950	Bootstrap	0.0329	1.05

This table shows the results of a Monte Carlo simulation as specified in section 4.2 with 1000 simulations and the indicated parameter settings. Tests were conducted for $\alpha = \gamma = 0.05$. The FDR controlling procedures applied are described in section 3. For the Storey procedure, $\lambda = 0.5$. Detrending is done using GLS.

Table A.7: Results for the FLS Data Set using HC_2 standard errors

	Regressor	$\hat{\beta}_i$	p-value	Classical	BH	Storey	BKY	Boot
1	GDP level 1960	−0.0170	0.0000001	1%	1%	1%	1%	
2	Fraction Confucian	0.0748	0.00002	1%	1%	1%	1%	
3	Life expectancy	0.000891	0.00309	5%	5%	5%	5%	
4	Equipment investment	0.127	0.00559	5%	5%	5%	5%	
5	Sub-Saharan dummy	−0.0201	0.00165	5%	1%	1%	1%	
6	Fraction Muslim	0.0107	0.295	-	-	-	-	
7	Rule of Law	0.0116	0.0247	10%	10%	10%	10%	
8	Number of years open economy	−0.00269	0.651	-	-	-	-	
9	Degree of Capitalism	0.00111	0.317	-	-	-	-	
10	Fraction Protestant	−0.00280	0.679	-	-	-	-	
11	Fraction GDP in mining	0.0401	0.0120	10%	5%	5%	5%	
12	Non-Equipment investment	0.0367	0.0850	-	-	-	-	
13	Latin American dummy	−0.0127	0.0637	-	10%	-	-	
14	Primary School Enrollment, 1960	0.0202	0.0537	-	10%	10%	10%	
15	Fraction Buddhist	0.00734	0.410	-	-	-	-	
16	Black-market premium	−0.00690	0.0347	-	10%	10%	10%	
17	Fraction Catholic	0.00307	0.550	-	-	-	-	
18	Civil Liberties	−0.00242	0.258	-	-	-	-	
19	Fraction Hindu	−0.0967	0.0001	1%	1%	1%	1%	
20	Political Rights	0.000162	0.926	-	-	-	-	
21	Primary Exports, 1970	−0.00550	0.457	-	-	-	-	
22	Exchange rate distortions	−0.00002	0.457	-	-	-	-	
23	Age	−0.000009	0.786	-	-	-	-	
24	War dummy	−0.00144	0.509	-	-	-	-	
25	Size labor force	0.0000003	0.000360	1%	1%	1%	1%	
26	Fraction speaking foreign language	−0.00246	0.393	-	-	-	-	
27	Fraction of Pop speaking English	−0.00707	0.134	-	-	-	-	
28	Ethnologic fractionalization	0.0137	0.0172	10%	5%	10%	10%	
29	Spanish Colony dummy	0.0131	0.0279	10%	10%	10%	10%	
30	SD of black-market premium	−0.000002	0.905	-	-	-	-	
31	French Colony Dummy	0.00894	0.0431	-	10%	10%	10%	
32	Absolute latitude	−0.00009	0.538	-	-	-	-	
33	Ratio of workers to population	−0.000520	0.950	-	-	-	-	
34	Higher education enrollment	−0.129	0.00185	5%	1%	1%	1%	
35	Population Growth	−0.114	0.559	-	-	-	-	
36	British Colony dummy	0.00680	0.0417	-	10%	10%	10%	
37	Outward orientation	−0.00453	0.0444	-	10%	10%	10%	
38	Fraction Jewish	−0.000774	0.927	-	-	-	-	
39	Revolutions and coups	0.00321	0.534	-	-	-	-	
40	Public Education Share	0.138	0.110	-	-	-	-	
41	Area (Scale Effect)	0.0000003	0.616	-	-	-	-	
42	Intercept	0.0207	–	1%	1%	1%	1%	

For every regressor in the FLS data set this table shows whether the variable is found to be significant when controlling the FDR at the indicated level. The procedures are described in section 3. We work with 5000 bootstrap iterations. In the Storey approach $\lambda = 0.5$.

Table A.8: Results for the FLS Data Set using HC_3 standard errors

	Regressor	$\hat{\beta}_i$	p-value	Classical	BH	Storey	BKY	Boot
1	GDP level 1960	−0.0170	0.00383	10%	10%	10%	-	
2	Fraction Confucian	0.0748	0.0616	-	-	-	-	
3	Life expectancy	0.000891	0.0670	-	-	-	-	
4	Equipment investment	0.127	0.118	-	-	-	-	
5	Sub-Saharan dummy	−0.0201	0.0772	-	-	-	-	
6	Fraction Muslim	0.0107	0.527	-	-	-	-	
7	Rule of Law	0.0116	0.205	-	-	-	-	
8	Number of years open economy	−0.00269	0.806	-	-	-	-	
9	Degree of Capitalism	0.00111	0.549	-	-	-	-	
10	Fraction Protestant	−0.00280	0.804	-	-	-	-	
11	Fraction GDP in mining	0.0401	0.243	-	-	-	-	
12	Non-Equipment investment	0.0367	0.299	-	-	-	-	
13	Latin American dummy	−0.0127	0.275	-	-	-	-	
14	Primary School Enrollment, 1960	0.0202	0.222	-	-	-	-	
15	Fraction Buddhist	0.00734	0.662	-	-	-	-	
16	Black-market premium	−0.00690	0.299	-	-	-	-	
17	Fraction Catholic	0.00307	0.720	-	-	-	-	
18	Civil Liberties	−0.00242	0.498	-	-	-	-	
19	Fraction Hindu	−0.0967	0.0159	-	-	-	-	
20	Political Rights	0.000162	0.958	-	-	-	-	
21	Primary Exports, 1970	−0.00550	0.659	-	-	-	-	
22	Exchange rate distortions	−0.00002	0.686	-	-	-	-	
23	Age	−0.000009	0.863	-	-	-	-	
24	War dummy	−0.00144	0.693	-	-	-	-	
25	Size labor force	0.0000003	0.0583	-	-	-	-	
26	Fraction speaking foreign language	−0.00246	0.616	-	-	-	-	
27	Fraction of Pop speaking English	−0.00707	0.388	-	-	-	-	
28	Ethnologic fractionalization	0.0137	0.125	-	-	-	-	
29	Spanish Colony dummy	0.0131	0.209	-	-	-	-	
30	SD of black-market premium	−0.000002	0.945	-	-	-	-	
31	French Colony Dummy	0.00894	0.245	-	-	-	-	
32	Absolute latitude	−0.00009	0.714	-	-	-	-	
33	Ratio of workers to population	−0.000520	0.970	-	-	-	-	
34	Higher education enrollment	−0.129	0.111	-	-	-	-	
35	Population Growth	−0.114	0.743	-	-	-	-	
36	British Colony dummy	0.00680	0.236	-	-	-	-	
37	Outward orientation	−0.00453	0.229	-	-	-	-	
38	Fraction Jewish	−0.000774	0.995	-	-	-	-	
39	Revolutions and coups	0.00321	0.710	-	-	-	-	
40	Public Education Share	0.138	0.390	-	-	-	-	
41	Area (Scale Effect)	0.0000003	0.762	-	-	-	-	
42	Intercept	0.0207	0.000	1%	1%	1%	1%	

For every regressor in the FLS data set this table shows whether the variable is found to be significant when controlling the FDR at the indicated level. The procedures are described in section 3. We work with 5000 bootstrap iterations. In the Storey approach $\lambda = 0.5$.

Table A.9: Comparison MP Data Set

	Regressor	Bootstrap approach	BMA post. prob.
1	ln GDP per capita, 1960	5%	100.0
2	Life expectancy, 1960	1%	100.0
3	A∗Mining	1%	76.1
4	South-East Asia	-	93.9
5	Fraction muslim	-	83.5
6	A∗ Land area	-	62.8
7	OECD	-	45.3
8	A∗Primary export, 1970	-	43.1
9	Primary export, 1970	-	49.9
10	A∗European language	-	51.4
11	European language	-	24.4
12	Fraction Confucian	-	21.7
13	A∗Colony	-	54.6
14	A∗Tropical fraction	-	100.0
15	A∗Landlocked	-	72.2
16	Latin America	-	7.5
17	A∗Labor force, 1960	-	10.6
18	Malaria prevalence, 1960	-	12.8
19	Landlocked	-	7.1
20	Fraction Catholic	-	7.2
21	A∗Malaria prevalence, 1960	-	78.9
22	A∗Primary school, 1960	-	3.2
23	A∗Fraction Muslim	-	1.4
24	Fraction Protestant	-	3.8
25	A∗ln GDP per capita, 1960	-	2.8
26	Tropical fraction	-	3.5
27	Fraction Buddhist	-	2.0
28	Primary school, 1960	-	4.4
29	A∗British colony	-	1.1
30	Sub Saharan Africa	-	2.7
31	A∗Fraction urban pop, 1960	-	1.1
32	A∗Life expectancy, 1960	-	1.5
33	A∗Secondary school, 1960	-	0.9
34	Mining	-	4.9
35	Fraction Hindu	-	1.0
36	A∗Fraction Catholic	-	0.5
37	Distance from equator	-	0.5
38	Land area	-	0.6
39	Colony	-	0.9
40	A∗Ethnoliguistic fractionalization	-	1.0
41	Spanish Colony	-	0.7
42	Secondary school, 1960	-	1.1
43	A∗French colony	-	5.1
44	A∗Distance from equator	-	0.2
45	Tertiary education, 1960	-	0.2
46	British colony	-	0.3
47	A∗Fraction Protestant	-	0.2
48	A∗Fraction English speaking	-	47.1
49	French colony	-	0.3
50	Ethnoliguistic fractionalization	-	0.2
51	Fraction English speaking	-	0.1
52	Labor force, 1960	-	0.1
53	Fraction Jewish	-	0.1
54	Fraction urban pop, 1960	-	0.3

"Bootstrap approach" denotes significance when controlling the FDR at the indicated level; "BMA post. prob." denotes the marginal posterior probability of inclusion found in Masanjala and Papageorgiou (2005).