ECONOMETRIC STUDIES ON FLEXIBLE MODELING OF DEVELOPING COUNTRIES IN GROWTH ANALYSIS

Dissertation

Presented for the Degree of Doctor rerum politicarum
at the Faculty of Economic Sciences
of the Georg-August-Universität Göttingen

by
Max Köhler
from
Göttingen, Germany

Göttingen, 2012

First Examiner: Prof. Dr. Stefan Sperlich
Second Examiner: Prof. Inmaculada Martínez-Zarzoso, Ph.D.

Contents

List of Figures

List of Tables

Acknowledgements

I thank Prof. Dr. Stefan Sperlich, for his support, patience and helpful comments and suggestions. Furthermore, I would like to mention Prof. Dr. Thomas Kneib, whom I not only thank for his support, but also for his willingness to act as examiner. I also thank Prof. Inmaculada Martínez-Zarzoso (Ph.D), for giving helpful comments and for acting as an examiner.
Lastly, I would like to thank my girlfriend for her loving support and patience and my parents for their unconditional support and encouragement throughout my academic career.

Chapter 1

Introduction and Overview

> *"There is no evidence that God ever intended the United States of America to have a higher per capita income than the rest of the world for eternity."*
>
> Robert M. Solow

This thesis is structured as a cumulative dissertation and combines three papers which are treated seperately in this introduction.
The first paper is concerned with nonparametric regression. The world of regression is basically divided into two different approaches. On the one hand, the parametric approach, in which a model that follows a given family of functional forms is adapted to the data. The disadvantage of this approach is obvious, since the optimal chosen representative can be far away from what really generates the data. For example, a linear function is not able to adapt to a potential curvature of the underlying data. When wanting to select the appropriate parametric form, one is faced to the problem of choosing among infinitely many different functional forms. Needless to say that this choice critical. Therefore, more flexible methods have been proposed, which gets us to the approaches of nonparametric regression for which no assumption about a specific functional form is needed, except of smoothness. One of these approaches is the nonparametric kernel regression. We face the situation that we have the data $(X_1,Y_1),\ldots,(X_n,Y_n)$, $n \in \mathbb{N}$, following
$Y_i = m(X_i) + error_i,\ i = 1,\ldots,n$. The task is to estimate the functional value $m(x)$, where x is somewhere between the smallest and the largest X_i. Thereby, $m(x)$ is estimated by a weighted average of the Y_i's, where the weight function is called kernel and the width of the interval over which averaging is performed is called bandwidth. The choice of the bandwidth is a trade-off situation. On the one hand, a larger bandwidth provides more data for the estimation, resulting in a smaller variance; on the other hand, data that are far away from the regression point give less credible information about what happens at x, resulting in a larger bias. One could say that the selection of the bandwidth is one of the

fundamental model selection problems of nonparametric kernel regression. Bandwidth selection methods deal with the sensitive choice of balancing variance and bias by estimating the bandwidth from the data. The task of chapter (2) is to give a review that explains and compares the bandwidth selection methods which are available in literature. We discuss, implement and compare almost twenty selectors, complete by again almost 20 linear combinations of two seemingly negatively correlated groups of selectors of which the six best are presented. By this means, we observe which methods behave similar and find a certain ranking of methods, although no bandwidth selector performed uniformly best. The paper was submitted in a statistical journal, coauthord with Stefan Sperlich and Anja Schindler.

The second and third papers in this thesis are concerned with growth regressions. The central motivation of the growth literature is to explain differences in the country's growth paths. The growth regressions therein usually inhabit proximate determinants of economic growth and, depending on the paper and the special question its author wants to investigate, some more determinants. The typically used and theoretically well justified proximate determinants are the initial level of income, the share of capital being invested in physical capital, a measure of human capital and the population growth. The list of additional explanatory variables is a non-exhaustive enumeration. These variables could for example be ethnic homogeneity, political freedom, political stability, foreign direct investment or trade-policy openness, to mention a few. As a result, growth can be seen as a theory of everything, producing contradictory results. As we find that this development is critical, we stick very close to the aforementioned proximate determinants.

Obviously, the reason for incorporating the additional explanatory variables in growth regressions is that the proximate determinants do not suffice to explain growth. One famous example is that the growth performance of the sub-group of sub-Saharan African countries is significantly worse than that of all other countries. This is especially surprising, as a prominent stylized fact about economic growth is that when comparing two otherwise similar countries, the one with the lower initial mean income will tend to see the higher rate of growth. When wanting to explain this phenomenon, authors usually add more variables to the growth regression and find that the coefficient of the dummy variable identifying the group of sub-Saharan African countries, namely the Africa-Dummy, loses its significance. Then it is concluded, that the set of extra variables must be the missing variables in growth models and must therefore be added to growth regressions to explain the real growth performance. However, it remains unclaer if the special unique output of these variables only identifies sub-Saharan African countries and therefore acts like a dummy, or if it really drives growth. Therefore, we find it necessary to derive statistical facts about the Africa-Dummy. Chapter (3) deals with this. We develop a statistical method that is able to identify the Africa-Dummy and can moreover be extended to derive empirical facts about it. Open questions are: How does the Africa-Dummy interact with

the other explanatory variables? To what extend is the parametric linear structure of growth regressions responsible for the significance of the Africa-Dummy? How does the Africa-Dummy evolve over time? Answers to these questions can be found in chapter (3). Moreover, it gives a detailed introduction to the methodology of growth regressions, explaining the advantages of some methods compared to others. The chapter is supposed to be published together with Stefan Sperlich in a statistical journal.

The appearance of the Africa-Dummy already motivates the third paper. Basically, the coefficient of the Africa-Dummy is a correction of the intercept for sub-Saharan African countries. But what about the other coefficients? For example: Is there a reason to believe that a poor country has the same returns to investments in physical capital than a rich country? Let's not only focus on the two distinct groups of sub-Saharan African and other countries and instead, consider more generally the individual countries in the world and concentrate on their output of measures of the income distribution; namely on poverty, inequality and the share of income earned by their middle class. The literature shows that these variables affect economic growth. Thereby, it is argued that a poor country behaves different than a rich country. But this different bahaviour is not accounted for when estimating mean coefficinets. Estimating mean coefficnets reveals more problems as the following simplified example shows. Consider a growth regression of the form

$$growth = \beta * (growth\ driver) + error$$

and consider that the sample is clearly divided into poor and rich countries. First, it is very likely to hold that $\beta_{poor} \neq \beta_{rich} \neq \beta_{mean}$. Therefore, the mean coefficient only reflects a theoretical situation that might not be fulfilled in any of the country groups. Second, this situation already indicates an endogeneity problem. For example, if poor countries have a smaller return to the growth driver than the rich countries, this difference is very likely to move simultaneously with the growth performance, as there must be some reason for that the poor countries are poor and that the rich countries are rich. Third, there are problems when putting the model to data. Poor countries have systematically weaker databases and therefore, the estimation of β_{mean} is highly suspicious to suffer from a sample selection bias. All these problems are not present if we separate the two coefficients β_{poor} and β_{rich} from the beginning.

Chapter (4) deals with these problems. We formulate and apply a variable-coefficients model, allowing for the possibility of a "continuous transition" from poor to rich. This transition is explained by the country's individual levels of poverty, inequality and the share earned by its middle class in each year. Note that in this situation, the set of explanatory variables is not extended, as the extra variables only explain the coefficients of the proximate determinants. We investigate how these coefficients differ. The analysis is conducted for the growth rate of the mean income, for that of the poorer twenty per cent of the society and that of the richer twenty per cent. The chapter is supposed to be published together with Stefan Sperlich in a statistical journal.

Chapter 2

A Review and Comparison of Bandwidth Selection Methods for Kernel Regression

Over the last four decades, several methods for selecting the smoothing parameter, generally called the bandwidth, have been introduced in kernel regression. They differ quite a bit, and although there already exist more selection methods than for any other regression smoother we can still see coming up new ones. Given the need of automatic data-driven bandwidth selectors for applied statistics, this review is intended to explain and compare these methods.

2.1 Introduction

Today, kernel regression is a common tool for empirical studies in many research areas. This is partly a consequence of the fact that nowadays kernel regression curve estimators are provided by many software packages. Even though for explorative nonparametric regression the most popular and distributed methods are based on P-spline smoothing, kernel smoothing methods are still common in econometric standard methods, for example for estimation of the scedasticity function, estimation of robust standard errors in time series and panel regression models. Still quite recently, kernel regression has experienced a kind of revival in the econometric literature on treatment effect estimation and impact evaluation, respectively. Nevertheless, until today the discussion about bandwidth selection has been going on - or at least not be closed with a clear device or suggestion for practitioners. Typically, software implementations apply some defaults which in many cases are questionable, and new contributions provide simulations limited to show that the

own invention outperforms existing methods in particularly designed cases. An explicit review or comparison article can be found only about bandwidth selection for density estimation, see Heidenreich, Schindler and Sperlich (2010) and references therein.

There are many, quite different approaches dealing with the problem of bandwidth selection for kernel regression. One family of selection methods is based on the corrected ASE criterion and uses ideas from model selection to choose an optimal bandwidth. To the best of our knowledge this was first introduced by Rice (1984). A second family has become quite popular under the name of cross-validation (CV) going back to Clark (1977). A disadvantage of the CV approach is that it can easily lead to highly variable bandwidths, see Härdle, Hall and Marron (1988). A recently studied way to improve it is the one-sided cross-validation (OSCV) method proposed by Hart and Yi (1998). Alternatives to the ASE minimizing and CV approaches are the so-called plug-in methods. They look rather at the asymptotic mean integrated squared error where the unknown quantities, depending on the density of the covariate, $f(x)$, the regression function $m(x)$, and the variance (function) of the conditional response, are replaced by pre-estimates or priors, cf. for example Ruppert, Sheather and Wand (1995). Finally, there exist various bootstrap approaches but mainly focusing on the local optimal bandwidth for which reason they a comparison is hardly possible. Cao-Abad and González-Manteiga (1993) proposed a smoothed bootstrap, and González-Manteiga, Martínez Miranda and Pérez González (2004) a wild bootstrap procedure, both requiring a pilot bandwidth to be plugged in. As it is the case for the aforementioned plug-in methods, if we have an appropriate pilot or pre-estimator, then the performance of these methods is typically excellent, else not. Asymptotics including the rate of convergence of these methods was first studied by Hall, Marron and Park (1992).

We review a big set of existing selection methods for regression and compare them on a set of different data for which we vary the variances of the residuals, the sparseness of the design and the smoothness of the underlying curve. For different reasons we concentrate on small and moderate samples and restrict to global bandwidths. Due to the complexity of the problem we have had to be rather restrictive and decided to concentrate on designs and models which we believe are interesting (with regard to their smoothness and statistical properties rather than the specific functional form) for social and economic sciences. We are aware that neither the set of methods nor the comparison study can be comprehensive but hope it nevertheless may serve as a fair guide for applied researchers. Note that most of them cannot be found in any software package. We are probably the first who implemented all the here reviewed selection methods.

Suppose we have random pairs $(X_1,Y_1),\ldots,(X_n,Y_n)$, $n \in \mathbb{N}$, where the X_i's are explanatory variables drawn from a continuous distribution with density function f. Without loss of generality, we assume $X_1 < X_2 < \ldots < X_n$. The Y_i's are response variables generated by the

following model:

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

with i.i.d. random variables ε_i with mean zero and unit variance. Further, $\sigma^2(x) = var(Y|x)$ is finite, and the ε_i are independent of all X_j. Assume one aims to estimate $m(x) = E(Y \mid X = x)$ for an arbitrary point $x \in \mathbb{R}$.

Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function that fulfills $\int_{-\infty}^{\infty} K(u)\,\mathrm{d}u = 1$, $\int_{-\infty}^{\infty} uK(u)\,\mathrm{d}u = 0$ and $\int_{-\infty}^{\infty} u^2K(u)\,\mathrm{d}u =: \mu_2(K) < \infty$. Furthermore, denote $K_h(u) := \frac{1}{h}K(u/h)$, where $h \in \mathbb{R}^+$ is our bandwidth and or smoothing parameter. When speaking of kernel regression, there exist slightly different approaches for estimating $m(x)$. The maybe most popular ones are the Nadaraya-Watson estimator proposed by Nadaraya (1964) and Watson (1964) and the local linear estimator. Thinking of least squares estimation, the first one approximates $m(x)$ locally by a constant, whereas the latter one approximates $m(x)$ locally by a linear function. Before the local linear or more generally, the local polynomial smoother became popular, a well known alternative to the Nadaraya-Watson estimator was the so-called Gasser-Müller estimator, see Gasser and Müller (1979), which is an improved version of the kernel estimator proposed by Priestley and Chao (1972). Fan (1992) presents a list of the biases and variances of each estimator, see that paper also for more details. It is easy to see that the bias of the Nadaraya-Watson estimator is large when $|f'(x)/f(x)|$ is large, e.g. for clustered data, or when $|m'(x)|$ is large. The bias of the Gasser-Müller estimator looks simpler, does not have these drawbacks and is design-independent so that the function estimation in regions of sparse observations is improved compared to the Nadaraya-Watson estimator. On the other hand, the variance of the Gasser-Müller estimator is 1.5 times larger than that of the Nadaraya-Watson estimator. The local linear estimator has got the same variance as the Nadaraya-Watson estimator and the same bias as the Gasser-Müller estimator. When approximating $m(x)$ with higher order polynomials, a further reduction of the bias is possible but these methods require mode assumptions - and in practice also larger samples. For implementation, these methods are less attractive when facing multivariate regression, and several considered bandwidth selection methods are not made for these extensions. Most of these arguments hold also for higher order kernels. When comparing the local linear with the Gasser-Müller and the Nadaraya-Watson estimator, both theoretical approaches and simulation studies show that the local linear estimator in most cases corrects best for boundary effects, see also Fan and Gijbels (1992) or Cheng, Fan and Marron (1997). Moreover, in econometrics it is preferred to use models that nest the linear model without bias and directly provides the marginal impact and elasticities, i.e. the first derivatives. All this is provided automatically by the local linear but unfortunately not by the Nadaraya-Watson estimator. Consequently, we will concentrate in the following on the local linear estimator. More precisely, consider

$$\min_{\beta_0,\, \beta_1 \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1(X_i - x))^2 K_h(x - X_i) \tag{2.2}$$

where the minimizer can be expressed as a weighted sum of the Y_i, i.e. $1/n\sum_{i=n}^{n} W_{h,i}(x)Y_i$. Denote $S_{h,j} = \sum_{i=1}^{n} K_h(x-X_i)(X_i-x)^j$ and consider the following two cases:

- If
$$\det\begin{pmatrix} S_{h,0}(x) & S_{h,1}(x) \\ S_{h,1}(x) & S_{h,2}(x) \end{pmatrix} = S_{h,0}(x)S_{h,2}(x) - (S_{h,1}(x))^2 \neq 0 \tag{2.3}$$
the minimizer of (2.2) is unique and given below.
- If $S_{h,0}(x)S_{h,2}(x) - (S_{h,1}(x))^2 = 0$ we distinguish between
  - $\diamond$ $x = X_k$ for a $k \in \{1,\ldots,n\}$ but X_k does not have its neighbors close to it such that $K_h(X_k - X_i) = 0$ for all $i \neq k$ such that $S_{h,1}(x_k) = S_{h,2}(x_k) = 0$. In this case, the minimizing problem (2.2) is solved by $\beta_0 = Y_k$, and β_1 can be chosen arbitrarily.
  - $\diamond$ $x \neq X_k$ for all $k \in \{1,\ldots,n\}$. Then the local linear estimator is simply not defined as there are no observations close to x.

Summarizing, for our purpose we define the local linear estimator by

$$\hat{m}_h(x) = \frac{1}{n}\sum_{i=1}^{n} W_{h,i}(x)Y_i \tag{2.4}$$

with weights

$$W_{h,i}(x) = \begin{cases} \frac{nS_{h,2}(x)K_h(x-X_i) - nS_{h,1}(x)K_h(x-X_i)(X_i-x)}{S_{h,0}(x)S_{h,2}(x) - S_{h,1}(x)^2} & \text{, if } S_{h,0}(x)S_{h,2}(x) \neq S_{h,1}(x)^2, \\ n & \text{, if } S_{h,0}(x)S_{h,2}(x) = S_{h,1}(x)^2,\ x = x_i \\ 0 & \text{, else} \end{cases}$$

if $W_{h,i}(x) > 0$ for at least one i. If $W_{h,i}(x) = 0\ \forall\ i$ the local linear estimator is not defined. Note that the matrix with entrances $\{W_{h,i}(X_j)\}_{i,j}$ gives the so-called hat-matrix in kernel regression.

Thanks to the very limited set of assumptions, such a nonparametric regressor is most appropriate for explorative data analysis but also for further statistical inference when model specification is crucial for the question of interest, simply because model misspecification can be reduced here to a minimum. The main drawback is, however, that if the empirical researcher has no specific idea about the smoothness of $m(x)$ but - which is commonly the case - he does not know how to choose bandwidth h. Indeed, one could say that therefore the selection of smoothing parameters is one of the fundamental model selection problems of nonparametric statistics. For practitioners this bandwidth choice is probably the main reason for not using nonparametric estimation.

To the best of our knowledge there are hardly - and no recent - reviews available comparing either theoretically or numerically the different existing bandwidth selection

methods for regression. Some older studies to be mentioned are Rice (1984), Hurvich, Simonoff and Tsai (1998), or Hart and Yi (1998). Yang and Tschernig (1999) compared two plug-in methods for multivariate regression, and more recently, González-Manteiga, Martínez Miranda and Pérez González (2004) compared a new wild bootstrap and cross validation but with a focus on local bandwidths. None of these studies compared several global bandwidth selectors for random designs. The aim was typically to introduce a new methods and compare it with a standard method.

In the next section we briefly discuss three risk measures (or say objective functions) on which bandwidth selection could and should be based on. In Section (2.3) and Section (2.4) we introduce and discuss the various selection methods we could find in the literature, separately for the three different risk measures. In Section (2.5) we present in detail extensive simulation studies to compare all here discussed selection methods. Section (2.16) concludes.

2.2 Typically used Risk Measures

We now address the problem of which bandwidth h is optimal, beginning with the question what means 'optimal'. In order to do so let us consider the well known density weighted integrated squared error (dwISE) and the mean integrated squared error (MISE), i.e. the expectation of the dwISE, of the local linear estimator:

$$\begin{aligned} MISE(\hat{m}_h(x) \mid X_1,\ldots,X_n) &= E[\, dwISE \,] = E\left[\int \{\hat{m}_h(x) - m(x)\}^2 f(x)\, dx\right] \\ &= \frac{1}{nh}||K||_2^2 \int_S \sigma^2(x)dx \\ &+ \frac{h^4}{4}\mu_2^2(K)\int_S (m''(x))^2 f(x)dx + o_P\left(\frac{1}{nh} + h^4\right), \end{aligned}$$

where $f(x)$ indicates the density of X, $||K||_2^2 = \int K(u)^2 du$, $\mu_l(K) = \int u^l K(u)du$, and f the unknown density of the explanatory variable X with the compact support $S = [a,b] \subset \mathbb{R}$. Hence, assuming homoscedasticity, the AMISE (asymptotic MISE) is given by:

$$AMISE(\hat{m}_h(x) \mid X_1,\ldots,X_n) = \frac{1}{nh}||K||_2^2\sigma^2(b-a) + \frac{h^4}{4}\mu_2^2(K)\int_S (m''(x))^2 f(x)dx, \quad (2.5)$$

where the first summand is the mean integrated asymptotic variance, and the second summand the asymptotic mean integrated squared bias; cf. Ruppert, Sheather and Wand (1995). That is, we integrated squared bias and variance over the density of X, i.e. we weight the squared error by the design. Finding a reasonable bandwidth means to balance the variance and the bias part of (2.5). An obvious choice of defining an optimal bandwidth is to say choose h such that (2.5) is minimized. Clearly, the AMISE consists mainly of unknown functions and parameters. Consequently, the selection methods' main

challenge is to find appropriate substitutes or estimates. This will lead us either to the so-called plug-in methods or to bootstrap estimates of the AMISE.

For estimating a reasonable bandwidth from the data we have to find an error criterion that can be estimated in practice. Focusing on practical issues rises not only the question of how to get appropriate substitutes for the unknown functions and parameters of (2.5) but also the question of why we should look at the mean integrated squared error, i.e. a population oriented risk measure, when we just need a bandwidth for our particular sample at hand. If one does not take the expectation over the sample, i.e. considers the dwISE, one finds in the literature the so-called *ASE* (for average squared error) replacing the integration over the density of x by averaging over the sample. So this risk measure is a discrete approximation of the (density-weighted) integration of the squared deviation of our estimate from the true function. We define our *ASE* by

$$ASE(h) = \frac{1}{n}\sum_{j=1}^{n}(\hat{m}_h(X_j) - m(X_j))^2\, w(X_j), \tag{2.6}$$

where we introduced an additional trimming or weight function w to eliminate summands $(\hat{m}_h(X_j) - m(X_j))^2$ where X_j is near to the boundary. Having the explanatory variables ordered, we can simply set $w(X_j) = 1_{[X_{l+1}, X_{n-l}]}$ for a given l. By this means, we can reduce seriously the variability of the *ASE* score function, see Gasser and Müller (1979). Denote the minimizer of *ASE* by $\hat{h}_0$. Note that the *ASE* differs from the MISE in two points; first we do not integrate but average over the design, and second we do not take the expectation with respect to the estimator. If one wants to do the latter, one speaks of the *MASE* with optimal bandwidth h_0. A visual impression of what this function looks like is given in Figure (2.1). For the sake of illustration we have to anticipate here some definitions given in detail at the beginning of our simulation Section (2.5). When we refer here and in the following illustrations of this section to certain models, for details please consult Section (2.5).

For now we denote a minimizer of any other score function by $\hat{h}$. Following Shibata (1981), the bandwidth selection rule is called asymptotically optimal with respect to the *ASE* risk measure, if and only if

$$\lim_{n\to\infty}\frac{ASE(\hat{h})}{ASE(\hat{h}_0)} = 1 \tag{2.7}$$

almost surely. If (2.7) is fulfilled, it follows easily that

$$\frac{ASE(\hat{h})}{ASE(\hat{h}_0)} \xrightarrow{P} 1 \tag{2.8}$$

or nearly equivalently

$$\frac{\hat{h}}{\hat{h}_0} \xrightarrow{P} 1, \tag{2.9}$$

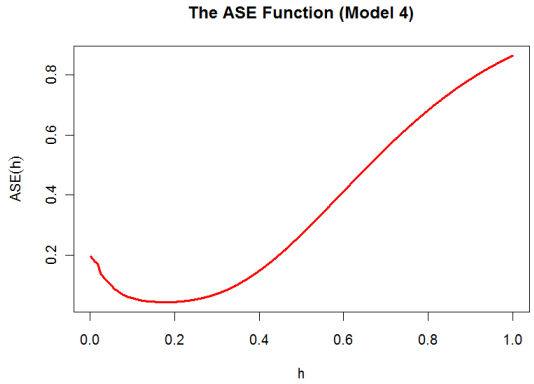

Figure 2.1: *ASE with* $w(X_j) = 1_{[X_6, X_{144}]}$ *for* $n = 150$ *simulated data following Model 3*

where $\xrightarrow{P}$ stands for convergence in probability. Note that optimality can also be defined with respect to the other risk measures like *MISE* or *MASE*.

Before we start, we should emphasize that we consider the ASE risk measure as our benchmark that should be minimized. All alternative criteria are typically motivated by the fact that asymptotically they are all the same. We believe that in explorative nonparametric fitting the practitioner is interested in finding the bandwidth that minimizes the (density weighted) integrated squared error for the given data, she/he is not interested in a bandwidth that minimizes the squared error for other samples or in average over all possible samples.

2.3 Choosing the smoothing parameter based on ASE

Having said that, it is intuitively obvious that one suggests to use ASE estimates for obtaining a good estimate of the 'optimal' bandwidth h. Therefore, all score functions introduced in this section are approaches to estimate the *ASE* function in practice when the true function m is not known. An obvious and easy approach for estimating the *ASE* function is plugging into (2.6) response Y_j for $m(X_j)$. This yields the substitution estimate

$$p(h) = \frac{1}{n} \sum_{j=1}^{n} (\hat{m}_h(X_j) - Y_j)^2 w(X_j). \tag{2.10}$$

It can easily be shown, that this is a biased estimator of $ASE(h)$, see for example Härdle (1992), chapter 5. One can accept a bias that is independent of h as in this case the minimizer of (2.10) is the same as that of (2.6). Unfortunately this is not the case for $p(h)$.

We present two approaches to correct for the bias. First the corrected ASE methods that penalizes each summand of (2.10) when choosing h too small, and second the cross

validation (CV) method that applies the leave one out estimator. Furthermore, we introduce the most recent one-sided cross validation (OSCV) method which is a remarkable enhancement of the classic CV.

2.3.1 The Corrected ASE

It is clear that $h \downarrow 0$ leads to interpolation, i.e. $\hat{m}_h(X_j) \to Y_j$, so that the function to be minimized, namely $p(h)$, could become arbitrarily small. On the other hand, this would surely cause a very large variance of $\hat{m}_h$ what indicates that such a criterion function would not balance bias and variance. Consequently, the corrected ASE penalizes when choosing h too small in an (at least asymptotically) reasonable sense. We define

$$G(h) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{m}_h(X_j))^2 \, \Xi\left(\frac{1}{n}W_{h,j}(X_j)\right) w(X_j), \tag{2.11}$$

where we use $w(X_j) = 1_{[X_{l+1}, X_{n-l}]}$ to trim near the boundary. $\Xi(.)$ is a penalizing function with first-order Taylor expansion

$$\Xi(u) = 1 + 2u + O(u^2) \ , \ u \to 0. \tag{2.12}$$

The smaller we choose bandwidth h the larger gets $W_{h,j}(X_j)$ and the penalizing factor $\Xi\left(\frac{1}{n}W_{h,j}(X_j)\right)$ increases. By conducting a first-order Taylor expansion of G and disregarding lower order terms it is easy to show that $G(h)$ is roughly equal to $ASE(h)$ up to a shift that is independent of h. The following list presents a number of proposed penalizing functions that satisfy the expansion $\Xi(u) = 1 + 2u + O(u^2) \ , \ u \to 0$:

- Shibata's model selector $\hat{h}_S = \underset{h \in \mathbb{R}^+}{\operatorname{argmin}} \, G_S(h)$, see Shibata (1981)

$$\text{with} \qquad \Xi_S(u) = 1 + 2u \ . \tag{2.13}$$

- Generalized cross validation (GCV) $\hat{h}_{GCV} = \underset{h \in \mathbb{R}^+}{\operatorname{argmin}} \, G_{GCV}(h)$, see Craven and Wahba (1979)

$$\text{with} \qquad \Xi_{GCV}(u) = (1-u)^{-2} \ . \tag{2.14}$$

- Akaikes information criterion (AIC) $\hat{h}_{AIC} = \underset{h \in \mathbb{R}^+}{\operatorname{argmin}} \, G_{AIC}(h)$, see Akaike (1974)

$$\text{with} \qquad \Xi_{AIC}(u) = \exp(2u) \ . \tag{2.15}$$

- The finite prediction error (FPE) $\hat{h}_{FPE} = \underset{h \in \mathbb{R}^+}{\operatorname{argmin}} \, G_{FPE}(h)$, see Akaike (1970)

$$\text{with} \qquad \Xi_{FPE}(u) = \frac{1+u}{1-u} \ . \tag{2.16}$$

- Rice's T (T) $\hat{h}_T = \underset{h\in\mathbb{R}^+}{\operatorname{argmin}}\, G_T(h)$, see Rice (1984)

$$\text{with} \qquad \Xi_T(u) = (1-2u)^{-1} \; . \tag{2.17}$$

All these corrected ASE bandwidth selection rules are consistent for $n \to \infty$ and $nh \to \infty$ as $h \downarrow 0$. In practice they certainly exhibit some deficiencies. To mitigate the problems that may occur for too small bandwidths, we will fix a data-adaptive lower bound for $\hat{h}$. Notice that for $h \leq h_{min,j} := \min\left\{X_j - X_{j-1}, X_{j+1} - X_j\right\}$ (recall that the explanatory variables are ordered for the sake of presentation), we get $\frac{1}{n}W_{h,j}(X_j) = 1$ and $\frac{1}{n}W_{h,i}(X_j) = 0$ for all $i \neq j$. In this case the j'th summand of (2.11) is not defined if we choose $\Xi(.) = \Xi_{GCV}(.)$ or $\Xi(.) = \Xi_{FPE}(.)$ but is $\Xi(1)$ finite for all other penalizing functions such that the j'th summand of (2.11) gets zero. This shows that for sufficient small bandwidths h the score function $G(h)$ is either not defined or can be arbitrarily small. This does surely not solve the problem of balancing bias and variance of the local linear estimator. Therefore, we first calculate the infimum of the set of all bandwidths for which (2.11) can be evaluated,

$$h_{min,G} = \max\left\{h_{min,l+1}, \ldots, h_{min,n-l}\right\}. \tag{2.18}$$

When minimizing $G(h)$ for any of the above listed criteria, we used only the bandwidths h that fulfill $h > h_{min,G}$, all taken from the grid in (2.18).

Figure (2.2) shows a plot of the corrected ASE score functions when using the Rice's T penalizing function. Not surprisingly, the optimal bandwidth that is related to the simulated smooth model 10 shows a clear optimum whereas the corrected ASE function corresponding to the rather wiggly regression $m(x)$ in model 4 takes it smallest value at the fixed (see above) minimum. However, even the smooth model might cause problems depending on how the minimum is ascertained: often one has at least two local minimums. These are typical problems of the corrected ASE bandwidth selection rules that we observed for almost all penalizing function. Recall that the models used for these calculations are specified in Section (2.5).

2.3.2 The Cross-Validation

In the following we present the CV method introduced by Clark (1977). To the best of our knowledge he was the first who proposed the score function

$$CV(h) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{m}_{h,-j}(X_j))^2 w(X_j) \; , \tag{2.19}$$

where $\hat{m}_{h,-j}(X_j)$ is the leave one out estimator which is simply the local linear estimator based on the data $(X_1,Y_1),\ldots(X_{j-1},Y_{j-1}),\ (X_{j+1},Y_{j+1}),\ldots,(X_n,Y_n)$. In analogy to the ASE function, the weights $w(\cdot)$ are used to reduce the variability of $CV(h)$. We again apply the

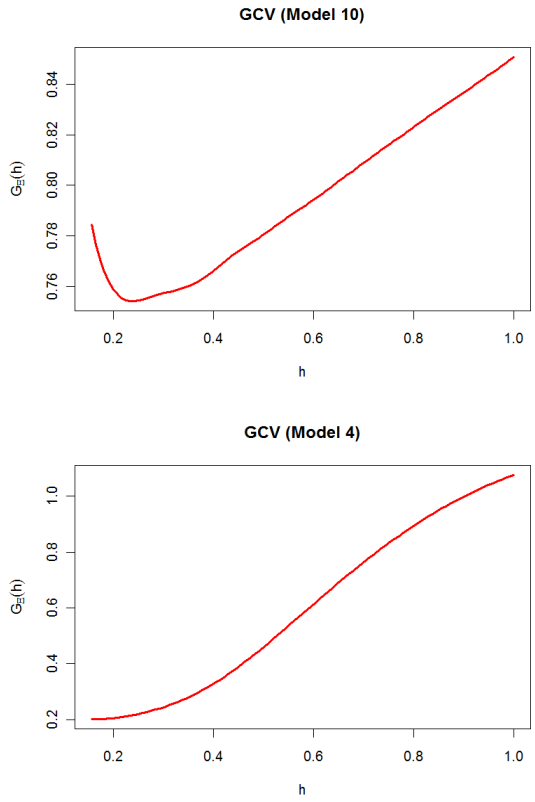

Figure 2.2: *The Corrected ASE Functions for* $n = 150$ *independent data following Model 4 and Model 10, respectively.*

trimming $w(X_j) = 1_{[X_{l+1}, X_{n-l}]}$ to get rid of boundary effects. It can easily be shown that this score function is a biased estimator of $ASE(h)$ but the bias is independent of h. This motivates the until today most popular data-driven bandwidth selection rule:

$$\hat{h}_{CV} = \underset{h \in \mathbb{R}^+}{\operatorname{argmin}}\, CV(h) \; . \tag{2.20}$$

As for the corrected ASE bandwidth selection rules, the CV bandwidth selection rule is consistent but in practice, curiously has especially serious problems as $n \to \infty$. The reason is that this criterion hardly stabilizes for increasing n and the variance of the resulting bandwidth estimate $\hat{h}$ is often huge. Clearly, for $h < h_{min,j} := \min\left\{X_j - X_{j-1}, X_{j+1} - X_j\right\}$ we have similar problems as for the corrected ASE methods as then the local linear estimator $\hat{m}_h(X_j)$ is not defined. Therefore, (2.19) is only defined if we fix $h > h_{min,CV}$ with

$$h_{min,CV} := \max\left\{h_{min,l+1}, \ldots, h_{min,n-l}\right\} \; . \tag{2.21}$$

Although this mitigates the problems at the lower bound of the bandwidth scale (i.e. for bandwidth approaching zero), Figure (2.3) exhibits similar problems for the CV as we saw

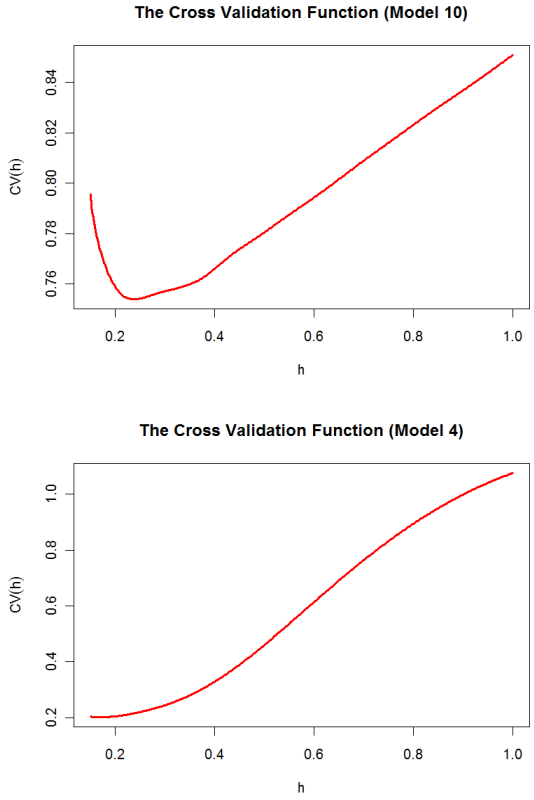

Figure 2.3: *The CV functions for $n = 150$ simulated data following Model 4 and Model 10, respectively.*

them for the corrected ASE criteria. Figure (2.3) shows the CV score functions when data followed model 10 and model 4. Again, for the wiggly model 4 we simply take the smallest possible bandwidth whereas for the smooth model 10 we seem to have a clear global minimum.

2.3.3 The One-Sided Cross-Validation

As mentioned above the main problem of CV is the lack of stability resulting in large variances of its estimated bandwidths. As has been already noted by Marron (1986), the harder the estimation problem the better CV works. Based on this idea, Hart and Yi (1998) developed a new modification of CV.

Consider the estimator $\hat{m}_{\hat{h}_{CV}}$ with kernel K with support $[-1,1]$ that uses the CV bandwidth $\hat{h}_{CV}$. Furthermore, we consider a second estimator $\tilde{m}_b$ with smoothing parameter b based

on a (selection) kernel L with support $[0,1]$. Then define

$$OSCV(b) = \frac{1}{n-2l} \sum_{i=l+1}^{n-l} (\tilde{m}_b^{-i}(X_i) - Y_i)^2, \tag{2.22}$$

where $\tilde{m}_b^{-i}(X_i)$ is the leave-one-out estimator based on kernel L. Note that l must be at least 2. This ensures that in each summand of (2.22) at least $l-1$ data points can be used. Denote the minimizer of (2.22) by $\hat{b}$. The OSCV method makes use of the fact that a transformation $h : \mathbb{R}^+ \to \mathbb{R}^+$ exists, such that $E(h(\hat{b})) \approx E(\hat{h}_{CV})$ and $Var(h(\hat{b})) < Var(\hat{h}_{CV})$. More precisely, (2.22) is an unbiased estimator of

$$\sigma^2 + E\left[\frac{1}{n-2l} \sum_{i=l+1}^{n-l} (\tilde{m}_b(X_i) - m(X_i))^2\right].$$

Therefore, minimizing (2.22) is approximately the same as minimizing

$$E\left[\frac{1}{n-2l} \sum_{i=l+1}^{n-l} (\tilde{m}_b(X_i) - m(X_i))^2\right]. \tag{2.23}$$

In almost the same manner it can be argued that minimizing $MASE(h)$ is approximately the same as minimizing $CV(h)$. We denote the minimizer of (2.23) by b_n and the $MASE(h)$ minimizer by h_n. Using the results in Fan (1992) for minimizing the MASE-expressions, dividing the minimizers and taking limits yields

$$\frac{h_n}{b_n} \to \left[\frac{||K||_2^2}{(\mu_2^2(K))^2} * \frac{(\mu_2^2(L))^2}{||L||_2^2}\right]^{1/5} =: C,$$

see Yi (2001). Note that the constant C only depends on known expressions of kernels K and L. One can therefore define the data driven bandwidth selector

$$\hat{h}_{OSCV} = C \cdot \hat{b}. \tag{2.24}$$

According to which selection kernel is used one gets different OSCV-values. A list of recommended and well studied selection kernels is given in Table (2.1), see also Figure (2.4). The transforming constants C of L_1 to L_4 are given together with the values $\mu_2^2(L_i)$ and $||L_i||_2^2$ in Table (2.2).

As for the corrected ASE and CV bandwidth selection rules, the OSCV bandwidth selection rule is consistent. Now consider the i'th summand of (2.22). Analogously to prior discussions, (2.22) is only defined if $b > b_{min,lOSCV} = \max\{X_{l+1} - X_l, \ldots, X_{n-l} - X_{n-l-1}\}$,

Table 2.1: *Selection kernels for left OSCV.*

Kernel	Formulae
One Sided Quartic	$L_1(x) = 15/8(1-x^2)^2 1_{[0,1]}$
Local Linear Epanechnikov	$L_2(x) = 12/19(8-15x)(1-x^2)1_{[0,1]}$
Local Linear Quartic	$L_3(x) = 10/27(16-35x)(1-x^2)^2 1_{[0,1]}$
opt. Kernel from Hart and Yi (1998)	$L_4(x) = (1-x^2)(6.92-23.08x+16.15x^2)1_{[0,1]}$

Table 2.2: *Properties of the selection kernels for left OSCV.*

Kernel	$\mu_2^2(L)$	$\|\|L\|\|_2^2$	C
L_1	0.148571	1.428571	0.8843141
L_2	-0.1157895	4.497982	0.6363232
L_3	-0.08862434	5.11357	0.5573012
L_4	-0.07692307	5.486053	0.5192593

so that for minimizing (2.22) we consider only bandwidths $b > h_{min,CV}$. Because of

$$\begin{aligned}
h_{min,G} &= h_{min,CV} \\
&= \max\{h_{min,l+1},\ldots,h_{min,m-l}\} \\
&= \max\{\min\{X_{l+1}-X_l, X_{l+2}-X_{l-1}\},\ldots,\min\{X_{n-l}-X_{n-l-1}, X_{n-l+1}-X_{n-l}\}\} \\
&\geq \max\{X_{l+1}-X_l,\ldots,X_{n-l}-X_{n-l-1}\} \\
&= b_{min,lOSCV} \\
&= 1/C * h_{min,lOSCV} \\
&\geq h_{min,lOSCV}
\end{aligned}$$

this problem is much less serious for the OSCV than for the other methods. Due to the fact that $\tilde{m}_b(x)$ uses only data that are smaller than the regression point x, the variance of $\tilde{m}_b(x)$ reacts much more sensitive when decreasing b. This makes it more likely that the true minimum of (2.22) is larger than $b_{min,lOSCV}$. And indeed, in our simulations the problem of not finding the true minimum did not occur. Clearly, the OSCV score functions show a wiggly behavior when choosing b small due to a lack of data when using data only from one side. Moreover, this selection rule overweights the variance reduction. Figure (2.5) demonstrates the problem: while for Model 4 we observe a clear minimum, for Model 10 we observe that the OSCV score function does not seem to visualize a punishment when b is chosen disproportionately large. In what follows we will deal with this problem and introduce modified OS kernels.

Note that the regression estimator used at the bandwidth selection stage, namely $\tilde{m}_b(x)$ in (2.22), uses only the data X_i that are smaller than the regression point x. This explains the notion left OSCV. For implementing the right OSCV, we use the kernel $R(u) := L(-u)$.

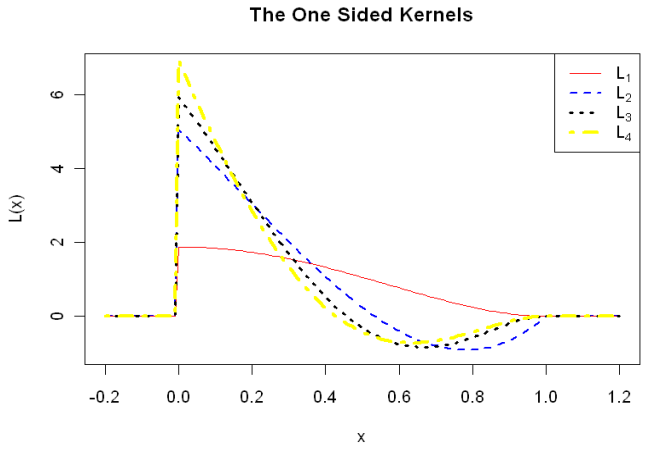

Figure 2.4: *The One Sided Selection Kernels used for left OSCV.*

Note that this kernel has support $[-1,0]$ and therefore $\tilde{m}_b(x)$ uses only data at the right side of x. The transforming constant C in (2.22) does not change. There is evidence that the difference of left and right sided OSCV is negligible. Hart and Yi (1998) considered the kernel estimator proposed by Priestley and Chao (1972) in an equidistant fixed and circular design setting and argued that the OSCV score function using any left sided kernel L is the same as the OSCV score function, when using its right sided version with kernel $L(-u)$. Furthermore, they conducted simulations with a fixed design setting using the local linear estimator and argued that in all simulations they had done, a correlation of the minimizers of the left and the right OSCV score function of larger than 0.9 was observed. Thus, in the theoretical considerations we only concentrate on the left sided OSCV and assume that the corresponding right sided OSCV has the same behavior.

When implementing the OSCV method one has to choose the one sided kernel L. Hart and Yi (1998) calculated the asymptotic relative efficiency, i.e.

$$ARE(K,L) = \lim_{n\to\infty} \frac{E((\hat{h}_{OSCV} - \hat{h}_0)^2)}{E((\hat{h}_{CV} - \hat{h}_0)^2)} \tag{2.25}$$

for different kernels for L. The setting was a fixed design using the kernel estimator for estimating m. They observed an almost twenty-fold reduction in variance compared to the CV method, when simply using the right kind of kernel L. They introduced two optimal kernels. One of them is the one sided local linear kernel based on Epanechnikov that is originally used for boundary correction in density estimation. For finding the optimal kernel in our case we conducted a simulation study, where we simulated 30 times the data $(X_1,Y_1),\ldots,(X_n,Y_n)$ for different data sets and different n. We compared the left OSCV methods, when using the kernels listed up in Table (2.4).

We calculated the bandwidths $(\hat{h}_0)_i$, $(\hat{h}_{CV})_i$ and $(\hat{h}_{OSCV})_i$ $(i = 1,\ldots,30)$ and then estimated

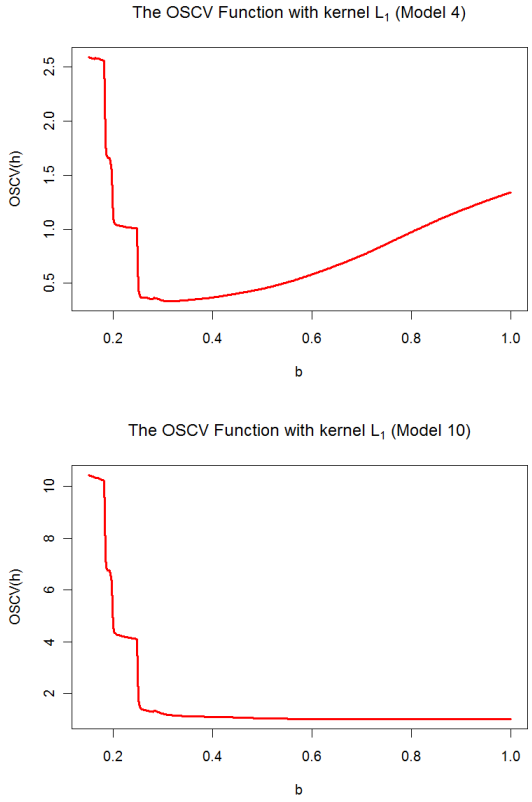

Figure 2.5: *The OSCV Functions based on* 150 *independent data* (X_i, Y_i).

$ARE(K,L)$ by

$$\widehat{ARE}(K,L) = \frac{\sum_{i=1}^{30}((\hat{h}_{OSCV})_i - (\hat{h}_0)_i)^2}{\sum_{i=1}^{30}((\hat{h}_{CV})_i - (\hat{h}_0)_i)^2}. \tag{2.26}$$

The results in the case of $n = 150$ are given in Table (2.3). We observed that in seven out of the twelve different cases using the kernel L_4 is best, in only three cases L_3 is best and kernel L_1 is only best in one case. When conducting the same simulation study with $n = 50$, $n = 100$ and $n = 200$ we observed very similar results. Therefore, we decided to use kernel L_4 in the following simulation studies.

A plot of the left OSCV Function, when using kernel L_4 is given in Figure (2.6). We observe that the OSCV functions are very wiggly when we use the kernel L_4 compared to using kernel L_1. The same wiggliness can be observed by using kernels L_2 and L_3. This behavior can also be observed when plotting the OSCV functions based on other data sets.

Even though one-sided cross validation from the left or from the right should not differ (from a theoretical point of view), in practice they do. To stabilize the behavior, Mammen, Martinez-Miranda, Nielsen and Sperlich (2011) proposed to merge them to a so-called double one-sided or simply do-validation (half from the left-sided, half from the

Table 2.3: *The estimated* $\widehat{ARE}(K,L_i)$ $i = 1,\ldots,4$ *and* $n = 150$.

Model	$\widehat{ARE}(K,L_1)$	$\widehat{ARE}(K,L_2)$	$\widehat{ARE}(K,L_3)$	$\widehat{ARE}(K,L_4)$	Best
1	5.828767	0.801370	0.915525	1.061644	L_2
2	96.290685	1.152327	19.722925	1.170663	L_2
3	6.928571	1.103896	1.032468	0.714286	L_4
4	2.051266	1.014796	1.013574	0.071266	L_4
5	1.541477	0.427530	0.427530	0.413856	L_4
6	2.025299	2.015951	1.000943	1.013723	L_3
7	2.674820	0.424460	0.250360	0.283453	L_3
8	1.519437	1.002538	0.998917	0.997350	L_4
9	3.474171	2.652201	2.651982	2.927879	L_3
10	3.945909	1.010591	1.000613	0.999650	L_4
11	47.943458	45.635282	38.257424	30.616100	L_4
12	1.484678	0.998468	0.524996	0.997636	L_3

right-sided OSCV bandwidth) for kernel density estimation and obtained amazingly good results with that procedure.

2.3.4 Notes on the Asymptotic Behavior

During the last two decades, a lot of asymptotic results for the corrected ASE methods and the CV method have been derived. Unfortunately, these asymptotic results are often only derived in the fixed and equidistant design case, when a kernel estimator or the Nadaraya-Watson estimator is considered. However, it is not hard to see that the results discussed in the following carry over to the local linear estimator which asymptotically can be considered as a Nadaraya-Watson estimator with higher order kernels.

Rice (1984) considered the kernel estimator

$$\hat{m}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i \tag{2.27}$$

proposed by Priestley and Chao (1972) in an equidistant and fixed design setting. Using Fourier-analysis, he analyzed the unbiased risk estimator of $p(h)$ introduced by Mallows (1976), and proved that its minimizer fulfills condition (2.9). He made some smoothness assumptions on K and m and considered bandwidths in the range of $H_n = \left[an^{-1/5}, bn^{-1/5}\right]$ for given a, b. Furthermore, he argued that this bandwidth selection rule is asymptotically equivalent to the corrected ASE and the CV selection rules and therefore, the minimizers of the corrected ASE functions also fulfill condition (2.9).

Härdle and Marron (1985) considered the Nadaraya-Watson estimator in a multivariate random design setting. They proved the optimality condition (2.7) for the minimizer of the

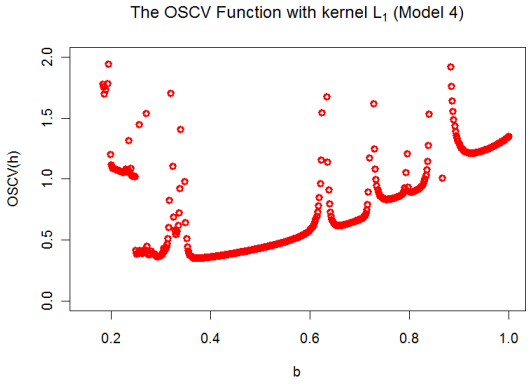

Figure 2.6: *The left OSCV function using kernel L_4.*

CV score function with respect to the ASE, ISE and MASE risk measures for the CV method. They made the assumption of h belonging to a range of possible bandwidths that is wider than $\left[an^{-1/5}, bn^{1/5}\right]$ so that the user of CV does not need to worry about the roughness of the underlying curve m. Further assumptions are the existence of the moments $E(Y^k|X=x)$, a Hölder continuous kernel K, i.e. $|K(u)-K(v)| \leq L||u-v||^{\xi}$ for a $\xi \in (0,1)$ and an $L>0$, $\int ||u||^{\xi}|K(u)|\,\mathrm{d}u < \infty$, the Hölder continuity of f and m and that the density f is bounded from below and compactly supported.

If conditions (2.8) and (2.9) are fulfilled for the bandwidth selection rules based on the CV and the corrected *ASE* score functions the question of the speed of convergence arises. Härdle, Hall and Marron (1988) considered the fixed and equidistant design case. They assumed i.i.d. errors ε_i for which all moments exist, a compactly supported kernel with Hölder continuous derivative and that the regression function has uniformly continuous integrable second derivative. Let $\hat{h}$ be any minimizer of a corrected *ASE* or the CV score function. Then, as $n \to \infty$,

$$n^{3/10}(\hat{h}-\hat{h}_0) \xrightarrow{\mathscr{L}} N(0,\sigma^2) \tag{2.28}$$

and

$$n^{3/10}(ASE(\hat{h})-ASE(\hat{h}_0)) \xrightarrow{\mathscr{L}} C\chi_1^2 \tag{2.29}$$

hold, where σ and C are constants depending on the kernel, the regression function and the observation error. It is interesting to observe that σ is independent of the particular penalizing function $\Xi()$ used. Taking the asymptotic rates of h's and *ASE*'s into account, one finds that condition (2.28) is of order $n^{1/10}$ and condition (2.29) is of order $n^{1/5}$. They also show that the differences $\hat{h}_0 - h_0$ and $ASE(\hat{h}_0) - ASE(h_0)$ have the same small rates of convergence. The authors conjecture that the slow rate of convergence of $\hat{h}$ and $\hat{h}_0$ is the best possible in the minimax sense.

Chiu (1990) considered the unbiased risk minimizer using the kernel estimator in an

equidistant, fixed design setting with periodic regression function (so-called circular design). He made the assumptions of independent errors ε_i for which all moments exist, some smoothness assumptions on the symmetric kernel K and m completed by technical conditions for the circular design. He only considered bandwidths belonging to a range that is slightly smaller than H_n. He pointed out that the normal distribution is not a good approximation for $\hat{h}$ because of its slow rate of convergence. Having finite samples in mind, he reasoned that

$$n^{3/10}(\hat{h}-h_0) \approx \sum_{j=1}^{\lfloor n/2 \rfloor} (V_j - 2) w_K(j), \tag{2.30}$$

where $V_1, \ldots, V_{\lfloor n/2 \rfloor}$ are i.i.d. χ^2_2-distributed random variables with weights $w_K(j)$ that only depend on the kernel K. This approximation has got interesting implications. Having in mind that the *MASE* minimizer is asymptotically the same as the *ASE* minimizer and that the unbiased risk minimizer is asymptotically the same as the minimizer of the corrected *ASE*'s and the CV score functions, it follows for example

$$n^{3/10}(\hat{h}_{CV}-h_0) \approx \sum_{j=1}^{\lfloor n/2 \rfloor} (V_j - 2) w_K(j). \tag{2.31}$$

When Hart and Yi (1998) computed the first twenty weights $w_K(j)$ $(j = 1,2,\ldots,20)$ and for the quartic kernel K and $n = 100$, they observed that $w_K(1)$ and $w_K(2)$ are large and negative but $w_K(3), \ldots, w_K(20)$ much smaller and mostly positive. This confirms that the distribution of $\hat{h}_{CV}$ is skewed to the left.

Assuming some further smoothness assumptions on the one sided selection kernel L and some technical conditions on L to be able to work with a circular design, they derived a similar result to (2.31) for OSCV, namely

$$n^{3/10}(\hat{h}_{OSCV}-h_0) \approx \sum_{j=1}^{\lfloor n/2 \rfloor} (V_j - 2) w_L(j). \tag{2.32}$$

When they calculated the weights $w_L(j)$ $(j = 1,2,\ldots,20)$ in (2.28) for L_4 and $n = 100$, they observed that these were now smaller in magnitude and almost symmetric around zero, indicating a symmetric distribution of $\hat{h}_{OSCV}$ with small(er) variance.

Yi (2001) proved the asymptotic stability of the OSCV selection rule. More precisely, let b_0 be the *MASE* optimal bandwidth using selection kernel L and $\hat{b}$ be the minimizer of the unbiased risk estimator. This is asymptotically the same as the minimizer of the OSCV score function, namely $\hat{b}_{CV}$. Then, for $Cb_0 - h_0 = o_P(\hat{b} - b_0)$ with constant C,

$$\lim_{n\to\infty} E((n^{3/10}(\hat{h}_{OSCV}-h_0))^2) = C^2 V(L), \tag{2.33}$$

where $V(L)$ is a constant that only depends on the selection kernel L. As before, he considered only an equidistant fixed design case, assumed normally distributed i.i.d.

errors, some smoothness for m, K and L with symmetric and compactly supported kernel K, and further technical conditions on m to be able to work with a circular design. Note that, when taking the rates of convergence of $\hat{h}_{OSCV}$ and h_0 into account, one finds, that his limit theorem (2.33) is of order $n^{1/5}$.

2.4 Choosing the smoothing parameter based on (A)MISE

In contrast to the cross-validation and corrected-ASE methods, the plug-in methods try to minimize the MISE or the AMISE. The conditional weighted AMISE of the local linear estimator $\hat{m}_h(x)$ was already given in (2.5). Minimizing w.r.t. h, leads to the AMISE-optimal bandwidth (h_{AMISE}), given by:

$$h_{AMISE} = \left(\frac{||K||_2^2 \cdot \int_S \sigma^2(x)\,dx}{\mu_2^2(K) \cdot \int_S (m''(x))^2 f(x) dx \cdot n} \right)^{1/5}, \tag{2.34}$$

where $S = [a,b] \subset \mathbb{R}$ is the support of the sample X of size n. One has the two unknown quantities, $\int_S \sigma^2(x)dx$ and $\int_S (m''(x))^2 f(x)dx$, that have to be replaced by appropriate estimates. Under homoscedasticity and using the quartic kernel, the h_{AMISE} reduces to:

$$h_{AMISE} = \left(\frac{35 \cdot \sigma^2 (b-a)}{\theta_{22} \cdot n} \right)^{1/5}, \qquad \theta_{rs} = \int_S m^{(r)}(x) m^{(s)}(x) f(x) dx, \tag{2.35}$$

where $m^{(l)}$ denotes the lth derivative of m.

The plug-in idea is to replace the unknown quantities by mainly three different strategies:

1. Rule-of-thumb bandwidth selector h_{rot}:
   The unknown quantities are replaced by parametric OLS estimators.

2. Direct-plug-in bandwidth selector h_{DPI}:
   Replace the unknown quantities by nonparametric estimates, where we need to choose 'prior (or pilot) bandwidths' for the two nonparametric estimators. In the second stage we use a parametric estimate for the calculation of these bandwidths.

3. Bootstrap based bandwidth selection h_{SB} and h_{WB}:
   The unknown expressions are estimated by bootstrap methods. In case of the smooth bootstrap (giving h_{SB}), again the unknown expressions in (2.35) are estimated, while the wild bootstrap method (h_{WB}) directly estimates the MISE of $\hat{m}_h$ and minimizes with respect to h. Both methods require a 'prior bandwidth'.

There also exists a bandwidth selector which does not require prior bandwidths but tries to solve numerically implicit equations. This procedure follows the solve-the-equation approach in kernel density estimation, see Park and Marron (1990) or Sheather and Jones

(1991). However, the results of this bandwidth selector are not uniformly better than those of the direct-plug-in approach (see Ruppert, Sheather and Wand (1995)) but require a much bigger computational effort, and are therefore quite unattractive in practice.

For the first two strategies a parametric pre-estimate in some stage is required. We have opted here for a piece-wise polynomial regression. For the sake of presentation assume the sample to be sorted in ascending order. The parametric OLS-fit is a blocked quartic fit, i.e. the sample of size n is divided in N blocks $\chi_j = \left(X_{\lfloor (j-1)n/N \rfloor + 1}, \ldots, X_{\lfloor jn/N \rfloor}\right)$, $(j = 1, \ldots, N)$. For each of these blocks we fit the model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i \quad i = \lfloor (j-1)n/N \rfloor + 1, \ldots, \lfloor jn/N \rfloor,$$

giving

$$\widehat{m}_{Q_j}(x) = \hat{\beta}_{0j} + \hat{\beta}_{1j} x_i + \hat{\beta}_{2j} x_i^2 + \hat{\beta}_{3j} x_i^3 + \hat{\beta}_{4j} x_i^4.$$

Then, the formula for the blocked quartic parametric estimator $\hat{\theta}_{rs}$, with $max(r,s) \leq 4$, is given by:

$$\hat{\theta}_{rs}^{Q}(N) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N} \widehat{m}_{Q_j}^{(r)}(X_i) \widehat{m}_{Q_j}^{(s)}(X_i) \mathbf{1}_{\{X_i \in \chi_j\}}.$$

Similarly, the blocked quartic estimator for σ^2 is

$$\hat{\sigma}_{Q}^{2}(N) = \frac{1}{n - 5N} \sum_{i=1}^{n} \sum_{j=1}^{N} (Y_i - \widehat{m}_{Q_j}(X_i))^2 \mathbf{1}_{\{X_i \in \chi_j\}}.$$

To choose N we follow Ruppert, Sheather and Wand (1995), respectively Mallows (1973): take the $\widehat{N}$ from $(1, 2, \ldots, N_{max})$ that minimizes

$$C_p(N) = \frac{RSS(N) \cdot (n - 5N_{max})}{RSS(N_{max})} - (n - 10N),$$

where $RSS(N)$ is the residual sum of squares of a blocked quartic N-block-OLS, and

$$N_{max} = max\left[min(\lfloor n/20 \rfloor, N^*), 1\right],$$

with $N^* = 5$ in our simulations. Another approach to the blocked parametric fit is to use nonparametric estimators for the unknown quantities in (2.35), see Subsection (2.4.2).

2.4.1 Rule-of-thumb plug-in bandwidth selection

The idea of the rule-of-thumb bandwidth selector is to replace the unknown quantities in (2.35) directly by parametric estimates, i.e. for θ_{22} use

$$\begin{aligned} \hat{\theta}_{22}^{Q}(N) &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N} \widehat{m}_{Q_j}^{(2)}(X_i) \widehat{m}_{Q_j}^{(2)}(X_i) \mathbf{1}_{\{X_i \in \chi_j\}} \\ &= \frac{1}{n} \sum_{i=(j-1)n/N+1}^{jn/N} \sum_{j=1}^{N} \left(2\hat{\beta}_{2j} + 6\hat{\beta}_{3j} x_i + 12\hat{\beta}_{4j} x_i^2\right)^2, \end{aligned}$$

and the estimator for σ^2

$$\begin{aligned}\hat{\sigma}_Q^2(N) &= \frac{1}{n-5N}\sum_{i=1}^{n}\sum_{j=1}^{N}(Y_i-\widehat{m}_{Q_j}(X_i))^2\mathbf{1}_{\{X_i\in\chi_j\}} \\ &= \frac{1}{n-5N}\sum_{i=(j-1)n/N+1}^{jn/N}\sum_{j=1}^{N}\left(y_i-\hat{\beta}_{0j}+\hat{\beta}_{1j}x_i-\hat{\beta}_{2j}x_i^2-\hat{\beta}_{3j}x_i^3-\hat{\beta}_{4j}x_i^4\right)^2 \quad (2.36)\end{aligned}$$

The resulting rule-of-thumb bandwidth selector h_{rot} is given by

$$h_{rot}=\left(\frac{35\cdot\hat{\sigma}_Q^2(N)(b-a)}{\hat{\theta}_{22}^Q(N)\cdot n}\right)^{1/5},$$

which now is completely specified and feasible due to the various pre-estimates.

2.4.2 Direct plug-in bandwidth selection

In this approach the unknown quantities in (2.35) are first replaced by nonparametric estimates. Then, for the nonparametric estimator of θ_{22} a bandwidth g is needed. An obvious candidate is the bandwidth g_{AMSE} that minimizes the AMSE (asymptotic mean squared error) of the nonparametric estimator of θ_{22}. Furthermore, a prior bandwidth λ_{AMSE} has to be determined for the nonparametric estimator of σ^2. These prior bandwidths are calculated with a parametric OLS-block-fit.

A nonparametric estimator $\hat{\theta}_{22}(g_{AMSE})$ can be defined by

$$\hat{\theta}_{22}(g)=n^{-1}\sum_{i=1}^{n}\left[\hat{m}_g^{(2)}(X_i)\right]^2, \quad (2.37)$$

where we use local polynomials of order ≥ 2. As local polynomial estimates of higher derivatives can be extremely variable near the boundaries, see Gasser, Kneip and Köhler (1991), we apply some trimming, i.e.

$$\hat{\theta}_{22}^{\alpha}(g_{AMSE})=\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}^{(2)}(X_i)\right]^2\mathbf{1}_{\{(1-\alpha)a+\alpha b<X_i<\alpha a+(1-\alpha)b\}}, \quad (2.38)$$

here the data are truncated within $100\cdot\alpha\%$ of the boundaries of support $S=[a,b]$, for some small $\alpha\in(0,1)$. Since for increasing α increases the bias, α must not be too large. In our simulations we follow the proposition $\alpha=0.05$ of Ruppert, Sheather and Wand (1995).

The prior bandwidth g_{AMSE}, i.e. the minimizer of the conditional asymptotic mean squared error of $\hat{\theta}_{22}(g)$ is given by

$$g_{AMSE}=\left[C_2(K)\frac{\sigma^2\cdot(b-a)}{|\theta_{24}|n}\right]^{1/7} \quad (2.39)$$

where the kernel dependent constant $C_2(K)$ for the quartic kernel is

$$C_2(K) = \begin{cases} \frac{8505}{13} & \text{if } \theta_{24} < 0 \\ \frac{42525}{26} & \text{if } \theta_{24} > 0 \end{cases}$$

The two unknown quantities are replaced by (block-wise) quartic parametric fits. For the prior estimation of σ^2 one uses the same as for the rule-of thumb bandwidth selector (see (2.36)). For θ_{24} we use:

$$\begin{aligned} \hat{\theta}_{24}^{Q}(\widehat{N}) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N} \widehat{m}_{Q_j}^{(2)}(X_i)\widehat{m}_{Q_j}^{(4)}(X_i)\mathbf{1}_{\{X_i\in\chi_j\}} \\ &= \frac{1}{n}\sum_{i=(j-1)n/N+1}^{jn/N}\sum_{j=1}^{N}\left(2\hat{\beta}_{2j}+6\hat{\beta}_{3j}x_i+12\hat{\beta}_{4j}x_i^2\right)\cdot 24\hat{\beta}_{4j}. \end{aligned}$$

This gives first an estimate for the g_{AMSE}, and afterwards for θ_{22}^{α}.

The nonparametric estimator for σ^2 is:

$$\hat{\sigma}^2 = \nu^{-1}\sum_{i=1}^{n}\left[Y_i - \hat{m}_{\lambda_{AMSE}}(X_i)\right]^2, \tag{2.40}$$

where $\nu = n - 2\sum_i w_{ii} + \sum_i\sum_j w_{ij}^2$ with $\{w_{ij}\}_{i,j=1}^{n}$ is the hat-matrix of $\hat{m}_{\lambda_{AMSE}}$. The prior bandwidth λ_{AMSE} is calculated as the minimizer of the conditional AMSE of $\hat{\sigma}_1^2$, see Ruppert, Sheather and Wand (1995). Hence, λ_{AMSE} is given by

$$\hat{\lambda}_{AMSE} = \left[C_3(K)\frac{\hat{\sigma}_Q^4(\widehat{N})(b-a)}{\left(\hat{\theta}_{22}^{.05}(\hat{g}_{AMSE})\right)^2 n^2}\right]^{1/9}$$

with the kernel dependent constant $C_3(K) = \frac{146735}{14339}$.

Now, the direct-plug-in bandwidth h_{dpi} is given by:

$$h_{DPI} = \left[35\frac{\hat{\sigma}^2(\hat{\lambda}_{AMSE})(b-a)}{\hat{\theta}_{22}^{.05}(\hat{g}_{AMSE})n}\right]^{1/5}.$$

2.4.3 Using smoothed bootstrap

The idea of is to apply bootstrap to estimate the MISE of $\hat{m}_h$ or some specific parameters of the regression or its derivatives. For a general description of this idea in nonparametric problems, see Hall (1990) or Härdle and Bowman (1988), though they only consider fixed designs. Cao-Abad and González-Manteiga (1993) discussed and theoretically analyzed several bootstrap methods for nonparametric kernel regression. They proposed the smooth bootstrap as an alternative to wild bootstrap because the wild bootstrap mimics the model when the design is fixed. If one refers to the random design, i.e. not the ISE or ASE but

MISE or MASE are of interest, hence the following resampling method is proposed: Draw bootstrap samples $(X_1^*, Y_1^*), (X_2^*, Y_2^*), \ldots, (X_n^*, Y_n^*)$ from the two-dimensional distribution estimate

$$\hat{F}_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i \le y\}} \int_{-\infty}^{x} \mathbf{K}_g(t - X_i)\, dt,$$

where g is a prior bandwidth asymptotically larger than h, see below. Cao-Abad and González-Manteiga (1993) state that, as the marginal density of X^* is the kernel density estimate of X given the original data and bandwidth g, and the marginal distribution of Y^* is the empirical distribution function of $\{y_i\}_{i=1}^n$, one has $E^*(Y^* \mid X^* = x) = \hat{m}_g(x)$, and a natural estimator for $Var(Y|x)$ is

$$\hat{\sigma}_g^2(x) = \frac{1}{n}\sum_{i=1}^{n} W_{gi} Y_i^2 - [\hat{m}_g(x)]^2 = Var^*(Y^* \mid X^* = x). \tag{2.41}$$

For the estimation of $\hat{\sigma}$ assuming homoscedasticity, we average (2.41) over $x = X_i^*$. Additionally, a nonparametric estimator for θ_{22} is calculated as in formula (2.37) using cubic splines on our bootstrap sample and with the same pilot bandwidth g. With an estimate of σ^2 and θ_2^2 at hand we can use formula (2.35) to calculate a smooth bootstrap bandwidth $\hat{h}_{SB}$ which is certainly still a function of the pilot bandwidth.

2.4.4 Using Wild Bootstrap

For early papers about the resampling plan of the wild bootstrap, see Cao-Abad (1991) or Härdle and Marron (1991). For its special use in bandwidth selection, see González-Manteiga, Martínez Miranda and Pérez González (2004). We will use their estimation procedure of the MSE. As we are not interested in obtaining bootstrap samples but in obtaining bootstrap estimates of the MASE, there is no need to introduce the creating of bootstrap samples. The squared bootstrap bias and the bootstrap variance can be calculated as

$$Bias_{h,g}^*(x) = \sum_{i=1}^{n} W_{hi}(x)\hat{m}_g(X_i) - \hat{m}_g(x)$$

and

$$Var_{h,g}^*(x) = \sum_{i=1}^{n} (W_{hi}(x))^2 (Y_i - \hat{m}_g(X_i))^2,$$

where g is again a pilot bandwidth that has to be chosen. For the selection of bandwidth h we are interested in the MISE or the MASE, an error criterion independent from x. For simplicity we opted for the

$$MASE(g,h) = \frac{1}{n}\sum_{i=1}^{n} MSE_{h,g}^*(X_i) \tag{2.42}$$

with $MSE_{h,g}^*(x) = \left[Bias_{h,g}^*(x)\right]^2 + Var_{h,g}^*(x)$. To get consistent estimators, for both the wild and the smooth backfitting, the pilot bandwidth g has to be larger (in

sample-size-dependent rates) than bandwidth h. Having chosen g, the MASE only depends on h so that minimizing (2.42) gives finally the optimal wild bootstrap bandwidth $\hat{h}_{WB}$. It can be easily seen, however, that the necessity of choosing a pilot (or also called prior) bandwidth, is the main disadvantage of the bootstrap methods.

2.4.5 Notes on the Asymptotic Behavior

It is clear that consistency can only be stated for the case where proper priors were used. Consequently, the rule-of-thumb estimator has no consistency properties itself, because of possible inconsistency of the there applied estimator for θ_{22}. We therefore will concentrate on results for the relative error of $\hat{h}_{DPI}$. Sheather and Jones (1991) stated for the asymptotic behavior of $\hat{h}_{DPI}$

$$\frac{\hat{h}_{DPI} - h_{MISE}}{h_{MISE}} \xrightarrow{P} D, \tag{2.43}$$

and that the method used to estimate $\hat{h}_{DPI}$, is of order $O_P(n^{-2/7})$.
Here, D is the error $\theta_{22}^{-1}\left[\frac{1}{2}\mu_4(K_{2,3})\theta_{24}G^2 + \sigma^2(b-a)||K_{2,3}||_2^2 G^{-5}\right]$ with $g = Gn^{-1/7}$ the prior bandwidth and $G > 0$ its constant. This consistency statement is based on (2.39), (2.40) with

$$\begin{aligned}
&\hat{\sigma}^2(\hat{\lambda}_{AMSE}) - \sigma^2 = O_P(n^{-1/2}) \,, \\
&\hat{\theta}_{22}(g)^{-1/5} - \theta_{22}^{-1/5} \simeq -\frac{1}{5}\theta_{22}^{-6/5}\left[\hat{\theta}_{22}(g) - \theta_{22}\right]
\end{aligned}$$

conditional on $X_1, \ldots, X_n$. Both together gives

$$\frac{\hat{h}_{DPI} - h_{MISE}}{h_{MISE}} \simeq -\frac{1}{5}\theta_{22}^{-1}\left[\hat{\theta}_{22}(g) - \theta_{22}\right]$$

leading to our (2.43), see Sheather and Jones (1991) for details. We know already from results of Fan (1992) and Ruppert and Wand (1994) that

$$h_{MISE} = h_{AMISE} + O_P(n^{-3/5})$$

so that one can conclude from (2.43) to consistency with respect to h_{AMISE}. The theoretical optimal prior bandwidth g is obtained by choosing G such that D equals zero – asymptotically not achievable, see Sheather and Jones (1991) for further discussion.

Cao-Abad and González-Manteiga (1993) studied in detail the statistical behavior of smooth bootstrap. For early consistency results of the wild bootsrap, see Cao-Abad (1991). The consistency of MSE estimation via wild bootstrap was proved in González-Manteiga, Martínez Miranda and Pérez González (2004). The optimal prior bandwidth for the both, the smoothed and the wild bootstrap is of order $n^{-2/9}$, see for example Härdle and Marron (1991). The specific expressions however, see for example Cao-Abad and

González-Manteiga (1993) or González-Manteiga, Martínez Miranda and Pérez González (2004), depend again on various unknown expressions so that we face similar problems as for h_{rot} and h_{PDI}.

2.4.6 A Mixture of methods

As already has been found by others, while some methods tend to over-smooth others undersmooth. In kernel density estimation it is even clear that the plug-in bandwidth and cross-validation bandwidth are negatively correlated. Heidenreich, Schindler and Sperlich (2010) studied the performance of bandwidths which are simple linear combinations of a plug-in plus a cross-validation bandwidth. For kernel density estimation these bandwidths turned out to perform pretty well in all of their simulation studies.

Motivated by these positive results we will also try out such mixtures of estimated bandwidths in the context of kernel regression estimation. Like Heidenreich, Schindler and Sperlich (2010) we will only consider linear mixtures of two bandwidths. In particular, we again mix a CV bandwidth or a corrected ASE -based one with a plug-in or bootstrap method based bandwidth. Depending on the weighting factor $\alpha \in (0,1)$, the mixed methods are denoted as:

$$Mix_{method1,method2}(\alpha) = \alpha \cdot \hat{h}_{method1} + (1-\alpha) \cdot \hat{h}_{method2}, \tag{2.44}$$

where $\hat{h}_\bullet$ denotes the optimal bandwidth to the respective method. We mix our bandwidth in the three following proportions, i.e. $\alpha = 1/2$, $\alpha = 1/3$ and $\alpha = 2/3$. As for all the others, we calculate the according ASE value for the resulting new bandwidths to assess the performance of the respective mix, see next Section.

2.5 Finite sample performance

Recall the MISE and MASE. Clearly, if $\int (f(x))^{-1}\,\mathrm{d}x$ is large, we expect a large integrated variance and therefore, the optimal bandwidth gives more weight on variance reduction and is therefore large. In cases of highly varying errors, i.e. a large σ^2, the same effect is observed. When the true underlying regression curve $m(\cdot)$ varies a lot, i.e. $\int (m''(x))^2\,\mathrm{d}x$ is large, a large integrated squared bias is expected so that the optimal bandwidth gives more weight on bias reduction and therefore, chooses a small bandwidth. Clearly, some selection methods will do better in estimating the bias, others in estimating the variance. The same will hold for capturing the oscillation, say $m''(\cdot)$ or the handling of sparse data areas or skewed designs. As a conclusion, a fair comparison study requires a fair amount of different designs and regression functions.

For our data generating process we first have to choose the distribution of X. Then, we have to consider which are reasonable functions for $m(x)$. Finally, we have to assume a value for the variance of the error term. We generated noisy data following the models $Y_i = 1.5 \cdot \sin(k \cdot X_i) + \sigma \cdot \varepsilon_i$ with $\varepsilon \sim \mathcal{N}(0,1)$ for different k's, different σ's and a uniform design, i.e $X_i \sim U[-1,1]$, or a standard normal design, i.e. $X_i \sim N(0,1)$. We also considered the performance of the methods where $X_i \sim 1/2 \cdot \mathcal{N}(-0.6, 1/4) + 1/2 \cdot \mathcal{N}(0.3, 1/3)$. Because the results are almost identical to the uniform distribution, we do not show the results of this design in the consideration below.

A list of all the models we used is given as:

Model	σ	Design	k	Model	σ	Design	k
1	1	uniform	6	7	0.5	uniform	4
2	1	normal	6	8	0.5	normal	4
3	0.5	uniform	6	9	1	uniform	2
4	0.5	normal	6	10	1	normal	2
5	1	uniform	4	11	0.5	uniform	2
6	1	normal	4	12	0.5	normal	2

Random numbers following a normal mixture design are an example which may easily yield a large integrated asymptotic variance. Furthermore, the data are bimodal (so that two clusters are expected) and slightly skewed. Moreover, $\int (m''(x))^2 \, \mathrm{d}x$ becomes larger as k increases so that a larger integrated squared bias is expected as k increases. The different σ's affect the integrated variance of the local linear estimator.

The aim of this section is to compare the small sample performance of all methods discussed in the previous sections. Remember the different methods: cross-validation, corrected ASE, plug-in and bootstrap. We also compare these methods with different mixtures of the classical cross-validation (CV) criterion respectively several correcting ASE methods, with the rule-of-thumb and the direct plug-in estimate (PI1 and PI2 resp.). The mixing procedure is to include one half of the optimal bandwidth $\hat{h}_{CV}$ resp. an optimal bandwidth of a corrected ASE method in different proportions with the optimal bandwidth of PI1 or PI2, then we assess the corresponding ASE value for the mixed bandwidth. The reason why this makes sense is that CV and corrected ASE methods tend to oversmooth while the PI methods tend to undersmooth the true $m(x)$.

All in all we present the following methods for estimation:

I cross-validation methods

1. CV: cross-validation
2. OSCV(L): one-sided cv (left)
3. OSCV(R): one-sided cv (right)

4. DoV: do-validation

II corrected ASE methods

5. Shib: Shibata's model selector
6. GCV: generalized cv
7. AIC: Akaikes information criterion
8. FPE: finite prediction error
9. Rice: Rice's T

III plug-in methods

10. PI1: rule-of-thumb plug-in
11. PI2: direct plug-in

IV bootstrap methods

12. SB: smooth bootstrap
13. WB: wild bootstrap

V mixtures of two methods

VI ASE: infeasible ASE

There are certainly many ways how to compare the selection methods. When having in mind that different selectors are looking at different objective functions, it is already clear that it cannot be fair to use only one criterion. Consequently, we compare the performance by different performance measures, most of them based on the averaged squared error (ASE), as this is maybe the one the practitioner is mainly interested in. More specific, the considered measures are:

m_1**:** $mean(\hat{h}_{opt})$
mean of the selected bandwidths for the different methods

m_2**:** $std(\hat{h}_{opt})$
standard deviation of the selected bandwidths

m_3**:** $mean\left[ASE(\hat{h})\right]$
classical measure where the ASE of $\hat{m}$ is calculated (and averaged over the 500 repetitions)

m_4**:** $std\left[ASE(\hat{h})\right]$
volatility of the ASE's

m_5**:** $mean(\hat{h} - h_{ASE})$
'bias' of the bandwidth selectors, where h_{ASE} is the real ASE-minimizing bandwidth

m_6**:** $mean\left[(\hat{h} - h_{ASE})^2\right]$
squared L_2 distance between the selected bandwidths and h_{ASE}

m_7**:** $mean\left[\mid \hat{h} - h_{ASE} \mid\right]$
L_1 distance between the selected bandwidths and h_{ASE}

m_8**:** $mean\left[ASE(\hat{h}) - ASE(h_{ASE})\right] = mean\left[\mid ASE(\hat{h}) - ASE(h_{ASE}) \mid\right]$
L_1 distance of the ASE's based on selected bandwidths compared to the minimal ASE

m_9: $mean\left(\left[ASE(\hat{h}) - ASE(h_{ASE})\right]^2\right)$
squared L_2 distance compared to the minimal ASE

In the following we will concentrate on the most meaningful measures, namely the bias of the bandwidths selectors (m_5), the means and standard deviations of the ASE's (m_3 and m_4), showed as box-plots, as well as the L_1-distance of the ASE's (m_8).

Without loss of generality, we used the Quartic Kernel throughout, i.e. $K(u) = \frac{15}{16}(1-u^2)^2 1_{\{|u|\leq 1\}}$. For both bootstrap procedures we tried several priors g but will present only results for the well working choice $g = 1.5 \cdot \hat{h}_{CV}$. All results are based on the calculations from 500 repetitions. In our simulation study we tried all methods for the sample sizes $n = 25$, $n = 50$, $n = 100$, and $n = 200$.

We will first compare all methods without the mixtures. In order to summarize the different methods of choosing the optimal bandwidth, we first consider the selected bandwidths and the corresponding bias for each method separately. Afterwards, we compare the methods by various measures.

Before we start with the numerical outcomes for the different methods we should briefly comment on the in practice also quite important questions of computational issues, in particular the complexity of implementation and computational costs, i.e. the time required to compute the optimal bandwidth along the considered methods. The fastest methods are the so-called corrected ASE methods. The second best in speed performance are the plug-in methods, where the rule-of-thumb plug-in is better than the direct plug-in. The fact that we only consider one-dimensional regression problems and a local linear smoother allows for an implementation such that the CV methods behave also quite good but certainly worse than the plug-in. In our implementation and for the somewhat larger sample sizes (in the end, we only consider small or moderate ones) the slowest were the bootstrap based methods, in particular the smooth bootstrap. The direct plug-in and the smooth bootstrap method turned out to be quite complex in programming. Note that in general for more complex procedures the numerical results should be better than for the other methods to legitimate the computational effort.

2.5.1 Comparison of the bias and L_1-distance for the different bandwidths (m_5, m_7)

Most of our numerical findings have been summarized in two figures: In Figure (2.7) we show the biases (m_5) and in Figure (2.8) the $L_1(h)$-distances (m_7) for all methods and models, but only for sample sizes $n = 25$ and $n = 200$.

We first summarize the behavior of CV and GCV since they behave almost identically. For the standard normal distribution (see right panel in Figure (2.7)), they are oversmoothing

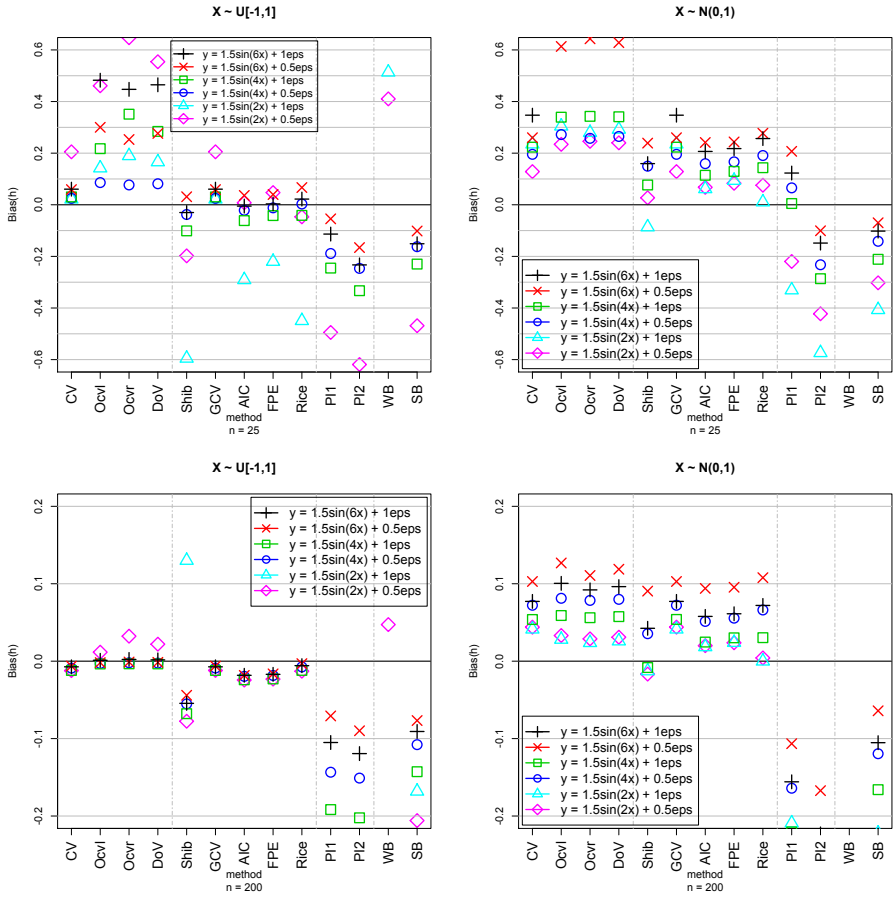

Figure 2.7: *Comparison of the bias for sample sizes $n = 25$ (above) and $n = 200$ (below)*

for all cases. For the uniform distribution the bias changes signs for increasing sample size, i.e. the bigger n the more tendency to undersmooth. Compared to all competitors, the L_1-distances are relatively small for all models, see Figure (2.8). Because of the almost identical behavior of these two methods we will only show CV in the next subsections respectively in the pictures below.

OSCV-l, OSCV-r and DoV also oversmooth for the standard normal distribution but for larger sample sizes the behavior improves considerably and compared to the competitors. Conspicuous for the normal design is that for $n = 25$ with a high frequency of the sinus function the values of m_5 and m_7 are very high. For the uniform distribution with $n = 200$ we cannot see any clear tendency to over- respectively undersmoothing, and the L_1-distance is almost zero, see also Figure (2.8). Because of the similar behavior of these three methods, and because DoV generally behaves best, we will only consider DoV in the following.

The bandwidth selection rules AIC, FPE and Rice from the second group are

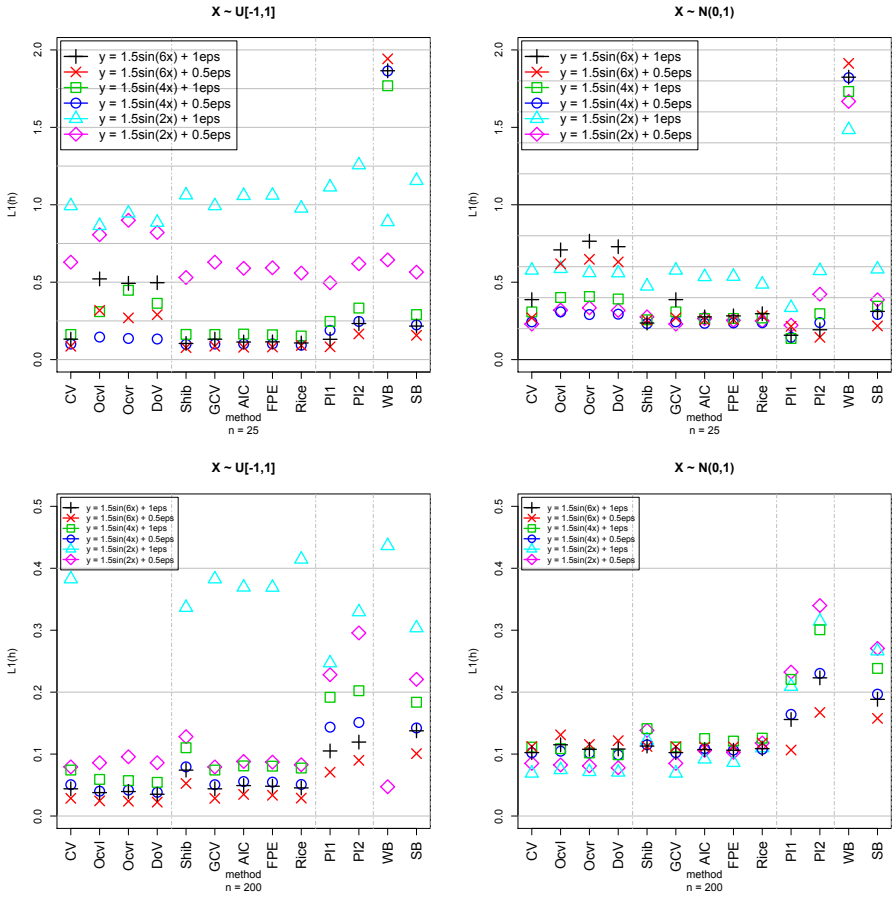

Figure 2.8: *Comparison of the L_1-distance for $n = 25$ (above) and $n = 200$ (below)*

oversmoothing for the standard normal distribution. Only for $n = 100$, $k = 2$, and $\sigma = 1$ Rice undersmooths, and has an almost zero bias (not shown in the Figure (2.7)). For the uniform design the three methods are almost always undersmoothing but in general show a good performance respective to the bias. The most noticeable for these three methods is that for $n = 25$ they behave better than CV, GCV and the one-sided CV methods, but for $n = 200$ the AIC, FPE and Rice are just as good as CV, GCV and the one-sided CV (see also Figure (2.8)). In comparison AIC, FPE and Rice seem to benefit less from increasing sample sizes, i.e. although the bias respectively the $L_1(h)$-distance is getting smaller in absolute value it is not getting smaller in the same magnitude like CV, GCV and the one-sided CV methods. In general, due to the bias, AIC, FPE and Rice show the best performance, i.e. they do not fail and are often the best. Because of the similar behavior of these three methods, and because Rice mostly behaves best, we will only consider Rice in the next sections.

The Shib selection method is almost always undersmoothing for the uniform design. For

the standard normal distribution it is oversmoothing for $n = 25$ but for the bigger samples there is no clear tendency. The main difference to the other ASE corrected methods is that Shib bandwidths are worse for the uniform design, but a little bit better for the normal design.

The plug-in methods and SB are almost always undersmoothing over all designs and sample sizes. They all undersmooth with a bias which is large in absolute value. For the standard normal design, PI1 shows a good bias behavior for the smallest sample size $n = 25$ and is best for the high frequency models. In general we can state for PI1, PI2 and SB that for $n = 25$ they are as good as all the methods from group I and group II, but for increasing sample size the value of the bias and the $L_1(h)$-distance loose compared to the other selectors. Hence, in the end, PI1, PI2 and SB seem to be worse than all the methods from the first and the second group.

The remaining method to be compared is the wild bootstrap "WB". From Figure (2.7) it can be seen that the values are often out of range except for model 11 for both sample sizes and model 9 for $n = 25$. In Figure (2.8) it can be seen that WB can only keep up with the other methods for model 9 and model 11. These two models are the smoothest of all. But WB is never the best method due to the bias and is best only for two special cases if we compare the $L_1(h)$-distances (model 9 for $n = 25$ and model 11 for $n = 200$). For the wiggly designs WB fails completely and chooses always the largest bandwidth of our bandwidth grid.

2.5.2 Comparison of L_1 and L_2-distances for the different bandwidths (m_6, m_7)

We will now summarize the performance of the selection methods according to the measures $L_1(h)$ and $L_2(h)$. In order to see the most important results, it is sufficient to concentrate on $k = 6$ and $\sigma = 1$ as all further results are almost identical to these with respect to the ordering of the considered methods (compare once again Figure (2.7) and Figure (2.8)). All in all we provide here the comparison of the selection methods along models 1, 2, 9 and 10. In Figure (2.9) we have plotted the resulting $L_1(h)$, and in Figure (2.10) the $L_2(h)$. For each of the four models we show the values for all sample sizes, i.e. for $n = 25, 50, 100, 200$.

Considering the wild bootstrap method "WB", we notice that it is only for model 9 (the smoothest) not out of the range of our plots. But even for this model we had to use a wider plotting range, because the $L_1(h)$ respectively $L_2(h)$ values turned out to be very large for basically all methods. "WB" can only compete with the other selection methods in this case, but for $n = 100$ and $n = 200$ is even here the worst of all methods. The cross validation, say "CV", method exhibits a pretty good performance for model 1; for sample

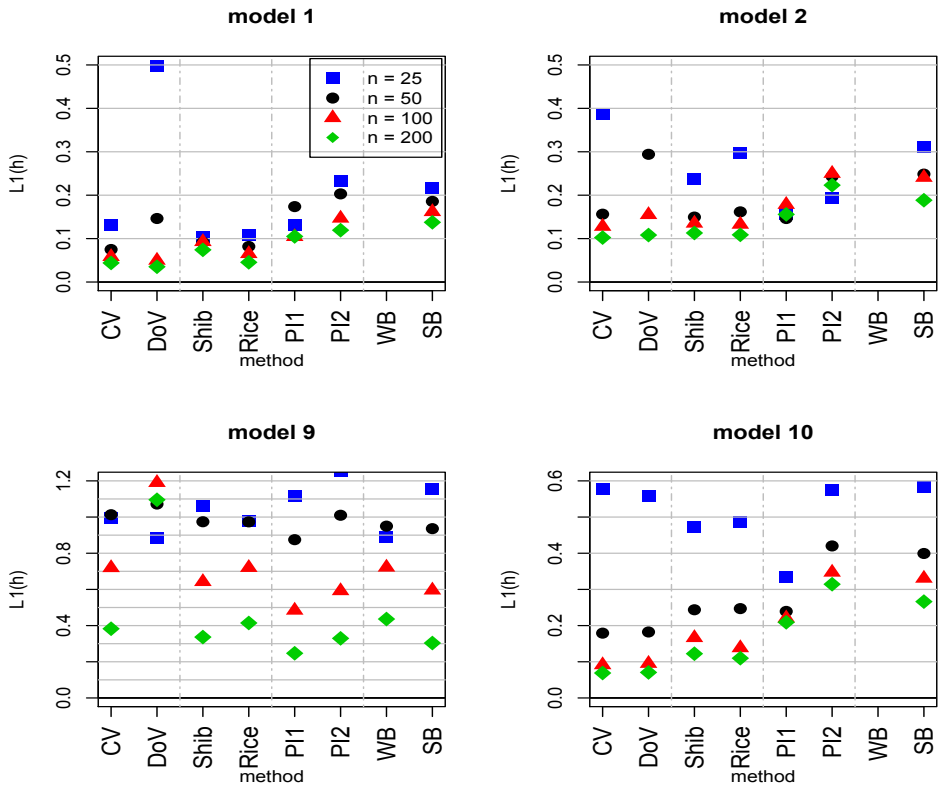

Figure 2.9: *L1(h) for each four models varying the sample size*

size $n = 50$ it is indeed the best. For model 2 and model 10 it shows only bad performances for $n = 25$ but good ones for the larger sample sizes. For model 9 it has an average behavior. This changes if we extend the cross validation idea to one-sided and do-validation. Indeed, for models 1, 2 and 10 "DoV" (and one-sided cross validation, where do-validation is based on) behaves badly only for $n = 25$, because of the resulting lack of information. It already behaves well for $n = 50$ and very well for not saying excellently for larger samples with $n = 100$ and $n = 200$. For model 9 its $L_1(h)$- respectively $L_2(h)$-values are even very good for $n = 25$. But for this very smooth model and sample sizes $n = 50$, $n = 100$ and $n = 200$ the plug-in PI1 is the best selection method. For model 10 PI1 is the best just for $n = 25$. Finally, "Shib" and "Rice" have an average behavior for all models and sample sizes, only for model 1 they are best for small samples with $n = 25$.

Summarizing we can say that the cross-validation methods need a sample size of at least 50 to perform well if we have a model that is not that smooth. For really smooth regression problems, the plug-in "PI1" does well.

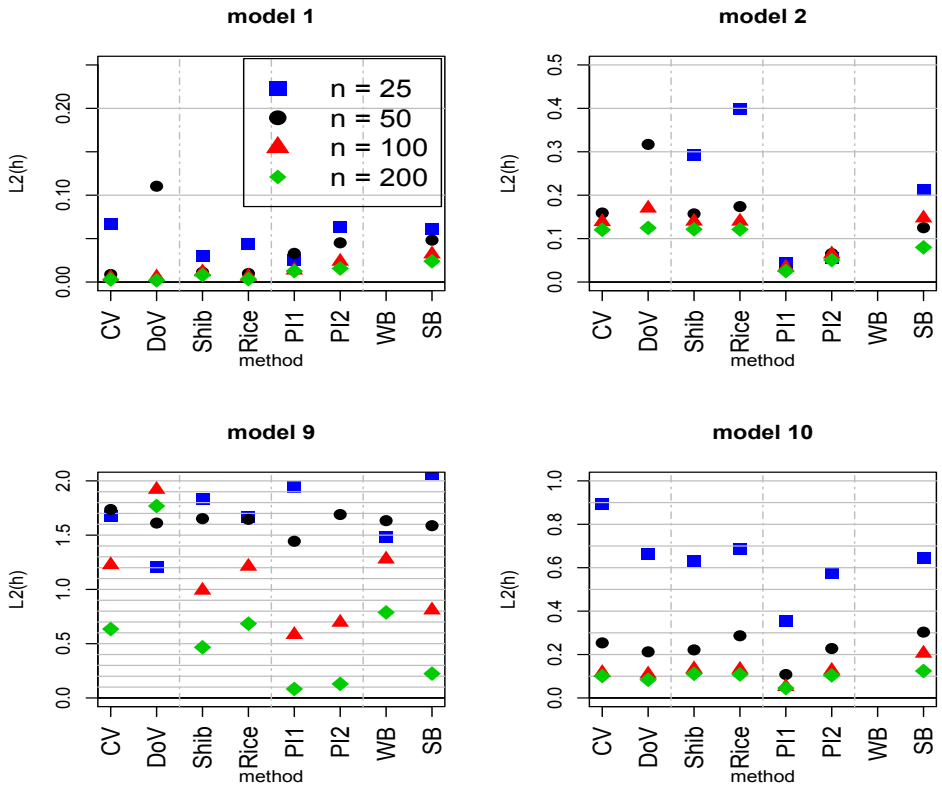

Figure 2.10: *L2(h) for each four models varying the sample size*

2.5.3 Comparison of the ASE-values (m_3, m_4)

In this subsection we summarize the results for the ASE-values of the different measures, i.e. the bandwidth that has been chosen for the respective method is inserted in the formula for the ASE. This is done because it enables us to compare rather the resulting regression performance than the bandwidths selected by the different methods. Needless to say, that the smallest ASE-value is reached with the benchmark, i.e. the true ASE optimal bandwidth. In our simulation we assumed twelve different models, i.e. we know the true value for $m(x)$ and the exact variance of the error term, what we do not in practice. For the same reasons we mentioned in the last subsection, the results for $k = 4$ and $\sigma = 0.5$ are skipped in the following. Hence, we compare only the boxplots of the selection methods along our models 1, 2, 9 and 10.

The main conclusions from the ASE-distributions can be summarized as follows. Varying the sample size, we can see from the boxplots, that for both designs, i.e. uniform design (see figure (2.11)) and standard normal design (see figure (2.12)), the means and median values for CV, DoV, Shib and Rice decrease with increasing sample sizes and decreasing frequencies. With respect to the inter quartile range (IQR henceforth) and the standard

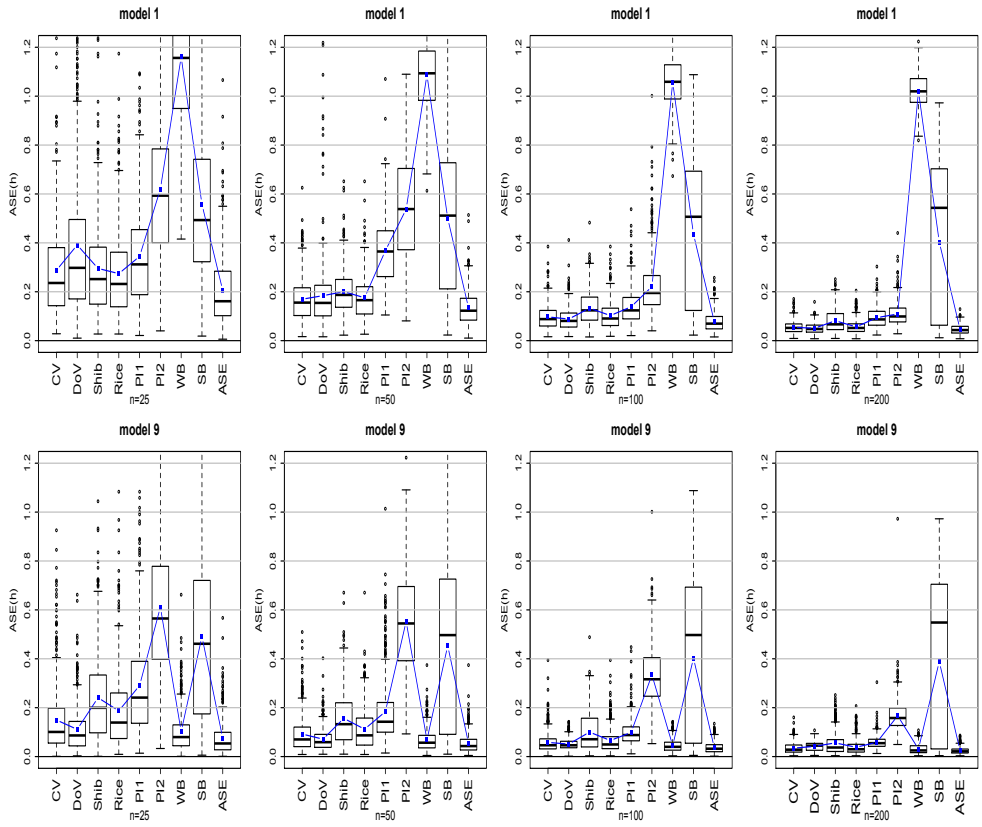

Figure 2.11: *ASE-values for $X \sim U[-1,1]$ for all sample sizes*

deviations it is almost the same with two exceptions. The first one is the IQR of DoV for model 9 and $n = 100$ is smaller than for $n = 200$, but there are more outliers for $n = 100$. The second one is Shib where the IQR increases with decreasing frequency in the uniform design for $n = 25$, $n = 50$ and $n = 100$.

For the plug-in and the bootstrap methods the results look quite messy. With respect to the IQR and the standard deviations, WB and PI1 clearly improve with increasing sample size. For PI2 it is the same for model 1, 2 and 9, but for model 10 it is the other way round. For SB the IQR and the standard deviation are getting larger with increasing sample size.

Now, we compare the methods for model 1 (see Figure (2.11), first row). DoV benefits most from increasing sample size, i.e. for $n = 25$ DoV is worst of group I, group II and PI1, but for $n = 200$ DoV is the overall best. CV and Rice behave very similar, and they are the best selectors for $n = 25$, and 2nd best for $n = 200$. Shib shows a good behavior for smaller sample sizes, but for $n = 100$ and $n = 200$ it has the largest IQR of group I and group II. In general, the plug-in methods behave worse than groups I and II, and only a little bit better than group IV.

The most noticeable of model 9 is that WB is the overall best method, there PI2 and SB

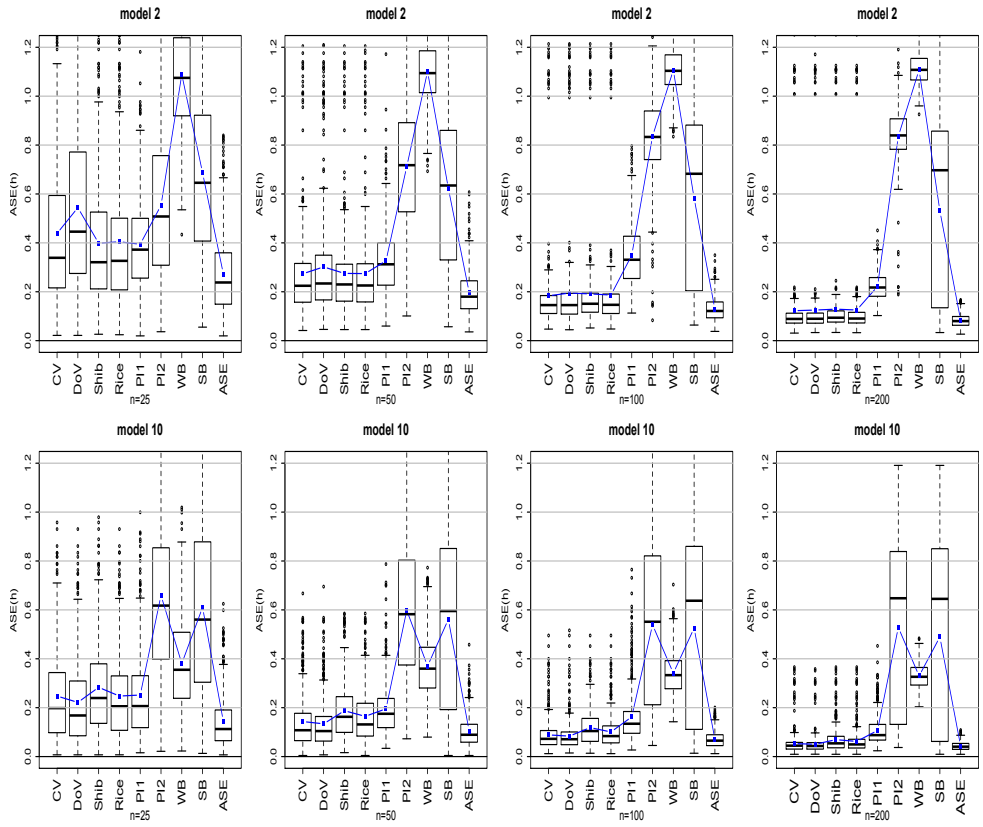

Figure 2.12: *ASE-values for $X \sim N(0,1)$ for all sample sizes*

behave worst. That is because model 9 is the smoothest model, i.e. a large bandwidth is optimal in this case. For $n = 25$ and $n = 50$ DoV is the best of I, II, and III, but for larger sample sizes CV and Rice are doing better.

The results for model 2, the most wiggly design, can be seen in figure (2.12), first row. The most interesting changes, compared to model 1, occur in the first four methods. There we have more extreme outliers the bigger the sample size is. The reason for that is that these methods have problems with outliers in the covariate X. Therefore, these outliers appear, if there is a random sample having a big proportion of observations around zero but thin tails. The behavior of the methods from group I and II is very similar, i.e. the chosen method does not have a big effect on the results. Further outcomes are similar respectively identical to model 1.

Finally, we consider the results for model 10 (see figure (2.12), second row). We state the differences to model 2 (for both $X \sim N(0,1)$) and model 9 (for both $k = 2$). In contrast to model 2, the extremity of outliers does only increase a little bit with increasing sample size which is due to the fact that the model is smoother. The difference to model 9 is that WB is not the best method for model 10. This is maybe due to the fact that model 10 is more

wiggly than model 9. But for both model 9 and model 10 selector WB does not fail completely in contrast to model 1 and model 2. For WB we can therefore state that if m is smooth enough this method can be used to estimate the bandwidth.

2.5.4 Comparison of the L_1 and L_2-distances of the ASE values (m_8, m_9)

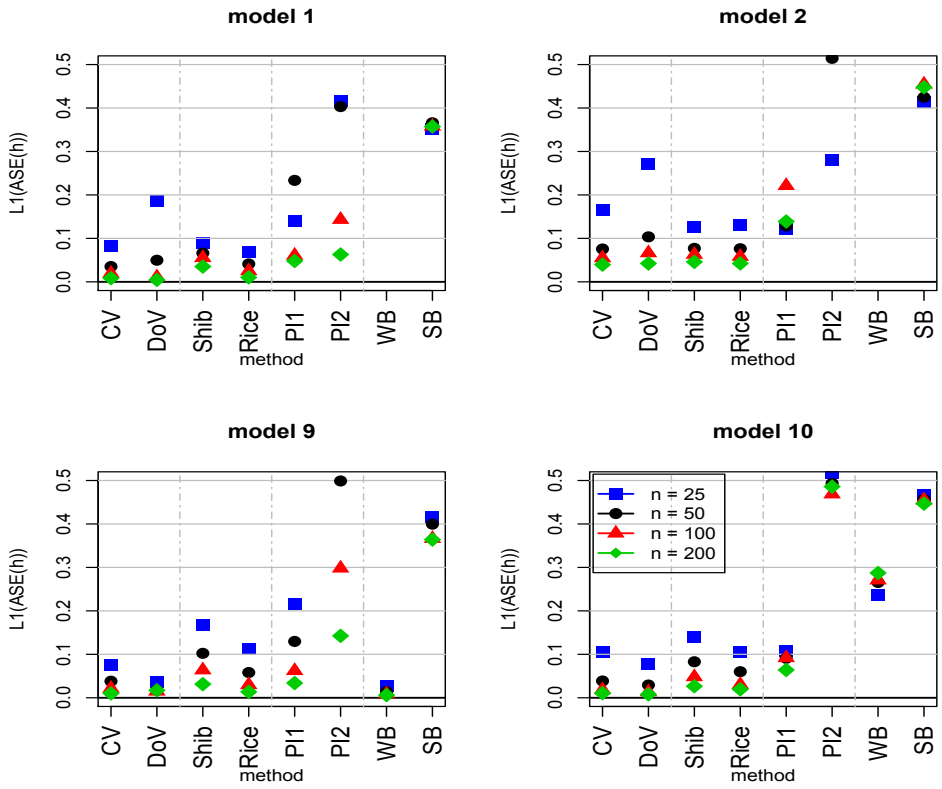

Figure 2.13: *$L_1(ASE)$ for each four models varying the sample size*

If we look at Figures (2.13) and (2.14), we can conclude that there is nothing new with respect to the comparison of the considered bandwidth selection methods. One interesting fact should be mentioned: the L_1-distances do generally not decrease with increasing sample size. In model 2 the L_1-distances increase with increasing sample size for the plug-in and bootstrap methods. In model 2 all L_1 and L_2-distances for WB are out of range. For this model PI1 is the best method for $n = 25$ but for all other sample sizes the CV and ASE-corrected methods behave better. PI2, WB and SB behave worse than the CV and ASE-corrected methods for all sample sizes.

One interesting fact for the CV and ASE-corrected methods is that there is a gap between $n = 25$ and the other sample sizes. That means, if we have a normal design respectively a

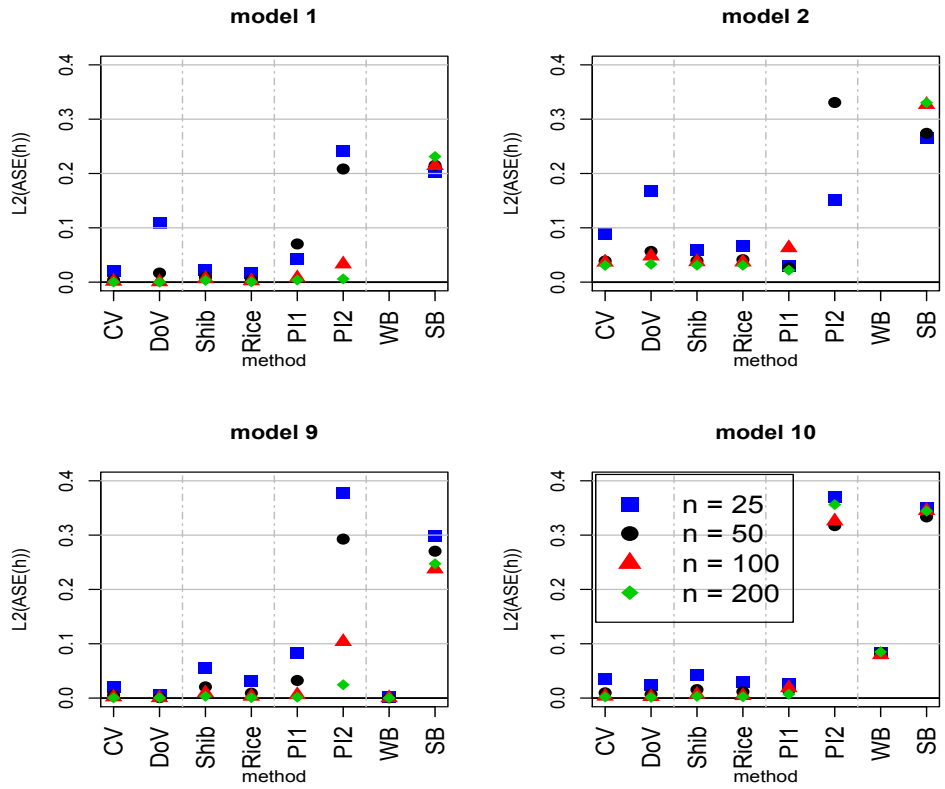

Figure 2.14: *$L_2(ASE)$ for each four models varying the sample size*

more wiggly model (see model 1) combined with an extreme small sample size, PI1 will be a good method in bandwidth estimation. Another mentionable fact is that for model 9, the smoothest model, WB is the best method when looking at the L_1 and L_2 ASE values, see Figures (2.13), (2.14). For model 10 WB is good, but not better than CV or corrected ASE based methods. That means that the decision of using WB depends more on the smoothness of m than on the smoothness of the distribution of X.

We mentioned in the beginning of Section (2.5) that PI2 and SB are more complicated to implement, and especially SB has a notable computation time. If we look at all the results we can say that PI2 and SB behave badly due to all the performance measures. Hence, there is no reason for using these two methods for bandwidth estimation for the considered models.

2.5.5 Comparison of different mixtures

Finally we tried to mix two methods in order to get better results than with only one method. We tried to mix a method that tends to oversmooth with a method that tends to

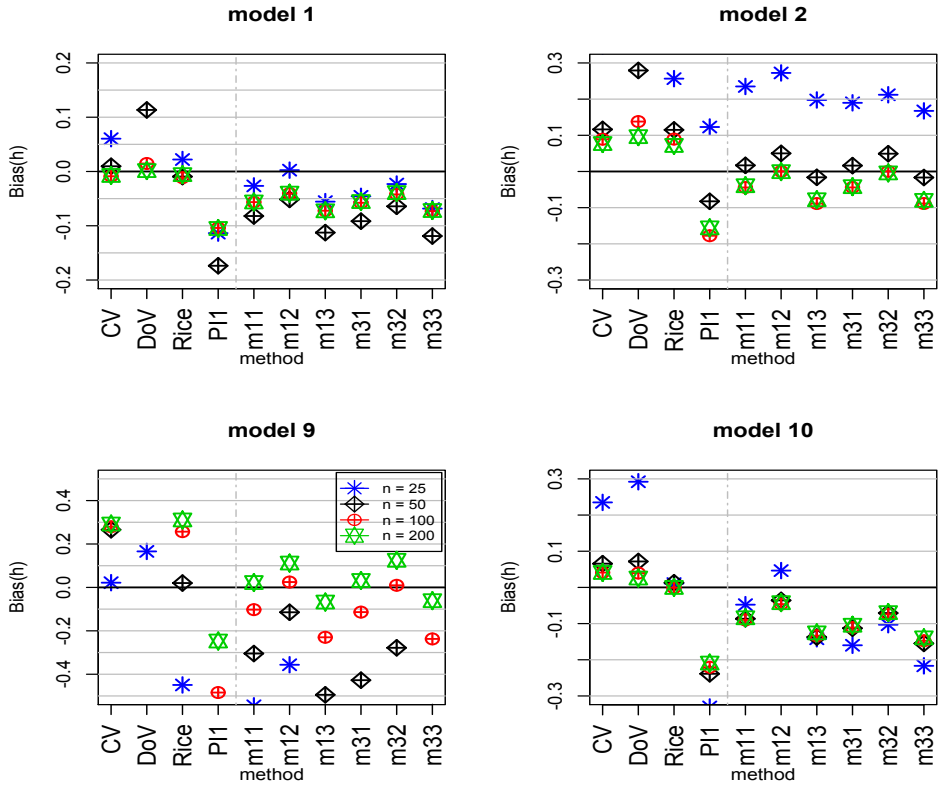

Figure 2.15: *bias(h)*

undersmooth the data. An obvious candidate is to mix the optimal bandwidth of the classical cross-validation (CV) respectively of a correcting ASE methods with one of the plug-in or a bootstrap optimal bandwidth. Recall that CV and corrected ASE methods tend to oversmooth while the PI and bootstrap methods tend to undersmooth. The mixtures will be compared with DoV which in the end is also a mixture, namely the left- and the right-sided OSCV method, respectively.

Depending on the weighting factor $\alpha \in (0,1)$, the mixed methods are denoted as in formula (2.44) by $Mix_{method1,method2}(\alpha)$. We only try to mix methods having a good performance. We also considered other mixtures, but the best results are obtained by mixing CV and Rice with PI1. Hence, the results we present here are:

1 m11: $Mix_{CV,PI1}(1/2)$ 3 m13: $Mix_{CV,PI1}(1/3)$ 5 m22: $Mix_{Rice,PI1}(2/3)$

2 m12: $Mix_{CV,PI1}(2/3)$ 4 m21: $Mix_{Rice,PI1}(1/2)$ 6 m23: $Mix_{Rice,PI1}(1/3)$

In fact, we did simulation for basically all two-folded mixtures but skip the presentation of all the other methods for the sake of brevity and because they simply behave worse.

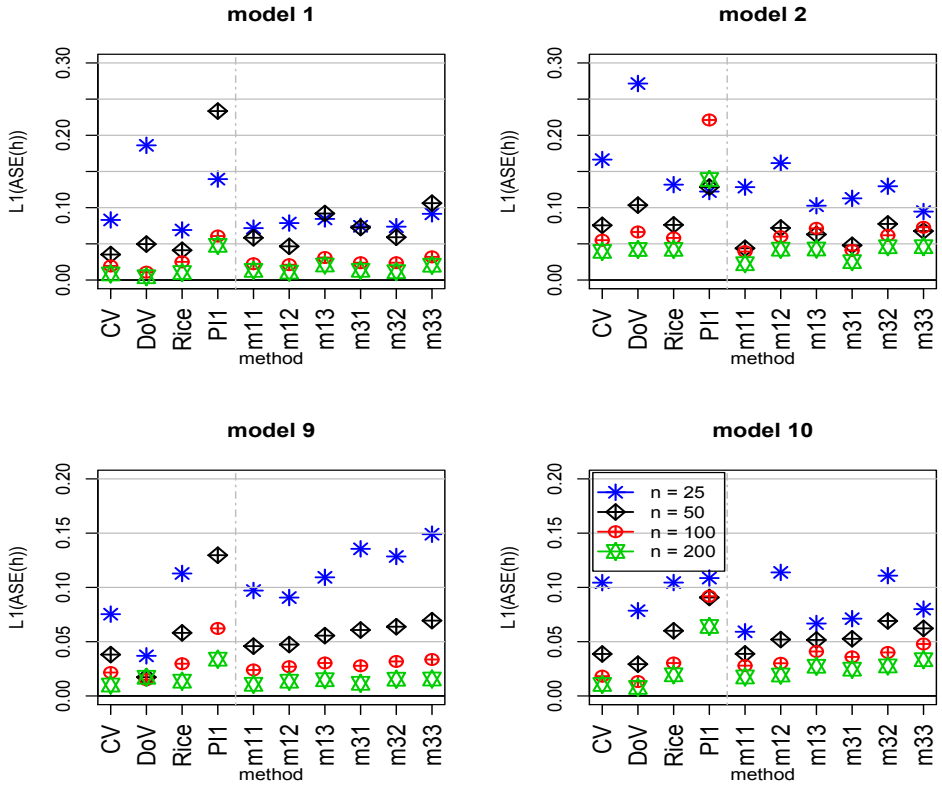

Figure 2.16: *L1(ASE)*

Specifically, we decided to show the following six different mixtures: three CV-PI1, and three Rice-PI1 mixtures.

In the Figures (2.15) and (2.16) we added DoV for obvious reasons mentioned above and because this method exhibited a pretty good performance before. The bias behavior of PI1 is almost always worst, the only exception is model 2 with a sample size of 25, where CV and DoV have the biggest bias. As already mentioned, the aim to mix methods was, to get better results than with one single method. But, we see, that the bias values of the mixtures are indeed better than for PI1 but worse than for CV or Rice. Only for model 2, the most wiggly model, we can achieve the objective of improvement. For the L1 values we get similar results, see Figure (2.16). In conclusion we can say, that the additional effort of mixing different methods seems not to be justifiable.

2.6 Conclusions

The problem of bandwidth choice is basically as old as nonparametric estimation is. While in the meantime kernel smoothing and regression has been becoming a standard tool for

explorative empirical research, and can be found in any statistical and econometric software package, the bandwidth selection can still be considered as an unsolved problem - at least for practitioners. Quite recently, Heidenreich, Schindler and Sperlich (2010) revised and compared more than thirty bandwidth selection methods for kernel density estimation. Although they could not really identify one method that performs uniformly better than all alternatives, their findings give clear guidelines at least for a certain class of densities like we typically expect and find them in social and econometric sciences.

This article is trying to offer a similar revision, comparison and guidelines for kernel regression. Though it is true that especially for large and huge data sets, today spline regression, and in particular P-spline estimation is much more common than is the use of kernel regression, the latter is still a preferred tool for many econometric methods. Moreover, it has been experienced a kind of revival in the fairway of treatment and propensity score estimation, smoothed likelihood methods and small area statistics (in the latter as a competitor to spline methods for reasons of interpretation).

To the best of our knowledge we are the first providing such a comprehensive review and comparison study for bandwidth selection methods in the kernel regression context. We have discussed, implemented and compared almost twenty selectors, completed by again almost 20 linear combinations of two seemingly negatively correlated (with respect to signs of the bandwidth bias) selectors of which the six best have been shown here. For different reasons discussed in the introduction we concentrated our study on local linear kernel estimation.

We started with a review of the idea and definition of the methods, its asymptotics, implementation and computational issues. Probably the most interesting results are summarized in the last section, i.e. Section (2.5). We could see which methods behave quite similar and found a certain ranking of methods although – like in Heidenreich, Schindler and Sperlich (2010) – no bandwidth selector performed uniformly best. Different to their study on density estimation, for regression the mixtures of methods could not really improve compared to the single use of a selector, except the so-called do-validation. This even turned out to be maybe even the best performing method though it is not always easy to implement nor computationally very fast.

For the rather small data sets considered, also the classical cross validation still performs well but should be replaced by generalized cross validation for increasing sample size. Note that for our context and estimator, CV and GCV behaved almost equivalently for the considered sample sizes. Nonetheless, already here and although we had rather wiggly as well as rather smooth functions under consideration, OSCV and especially DoV outperformed the classical CV. So it did for almost all models and sample sizes also compared to the other methods, at least when looking at the distribution of ASE, see Subsection (2.5.4). In our opinion, for the practitioner this is the most important measure.

It should be mentioned that in the reduced set of selectors, the method proposed by Rice (1984) did also a pretty fair job for the models and sample sizes considered in this article.

Chapter 3

The Africa-Dummy in Growth Regressions

The Africa-Dummy has been identified and different explanations for its appearance have been published. In this paper, the issue of the empirical identification of the Africa-Dummy is addressed. We introduce a fixed effects regression model to identify the Africa-Dummy in one regression step so that its correlations to other coefficients can be estimated. A semiparametric extension of this model checks whether the Africa-Dummy is a result of misspecification of the functional structure. Furthermore, we show that sub-Saharan African countries have a positive return to the population growth and when adding interaction effects, the Africa-Dummy is even positive. Moreover, we show that the Africa-Dummy changes dramatically over time and the punishment for sub-Saharan African countries decreases incrementally since the mid-nineties. According to the Augmented Solow Growth model, it was even insignificant since the end-nineties.

3.1 Introduction

This paper focuses on the Africa-Dummy. We use the growth model by Mankiw, Romer and Weil (1992) as a theoretic justification to run growth regressions. This model contains simplifications as groups of countries possess certain characteristics that are hard to measure and to incorporate, but represent systematic drivers for growth. For example Barro, Mankiw and Sala-i-Martin (1995) mention that international capital markets have a significant impact on growth rates, especially on the convergence of the poor countries. Another unrealistic simplification is criticized by Islam (1995). He argues that countries have fundamentally differing production functions so that comparisons between their economies are difficult. Furthermore, the endowment with resources can be infinitely substituted by capital. For example Georgescu-Roegen (1975) criticizes that this point of view is too optimistic with respect to the limitations of technological progress. Other variables that are correlated to economic growth but not incorporated in the growth model are political factors (see Collier and Gunning (1999)), diseases especially AIDS (see Were and Nafula (2003)), geographical factors and trade openness (see Sachs and Warner (1997)), ethnic diversity (see Easterly and Levine (1997)) or historical reasons such as the colonial heritage (see Price (2003)), to mention a few. Among others, these problems result in empirical weaknesses. For example Barossi-Filho, Goncalves Silva and Martins Diniz (2005) summarize that among most regressions the estimated capital share exceeds the value obtained from the national accounts and that the estimated convergence rate is usually too low. One example is the group of sub-Saharan African countries, meaning that the model by Mankiw, Romer and Weil (1992) is not able to explain the growth in sub-Saharan Africa, because its economic fundamentals incorporated in the model are not as bad as their actual performance. The result is that, if an additional variable is added, that only indicates the membership to sub-Saharan Africa, namely the Africa-Dummy, it has a significant coefficient with a negative sign. Barro (1991) for example runs a cross-sectional regression. This means that he holds an initial and a final time point fixed and calculates the growth rates in this time horizon for each country before regressing them on several explanatory variables. The result is a negative and significant Africa-Dummy. As African countries started with a lower level of income, they should converge to the income observed in regions that have similar characteristics. The presence of the Africa-Dummy shows that this is not the case. There is a lot of literature addressing this issue. For example Collier and Gunning (1999) mention that in 1975, 60% of all Africans lived in regimes that were not legally elected and democratic structures are often not achievable in the medium-term. Additionally, governments tend to implement lax monetary policies, not considering the inflationary long-run effects. They also report of high corruption, bureaucracy and a lack of public security. Another example is Were and Nafula (2003) who show how diseases and especially AIDS affect economic indicators. In order to

eliminate the Africa-Dummy, authors add variables to the growth regression. Sachs and Warner (1997) focus on the effects of trade openness and landlocked status. They conclude that a lack of liberalization and too restrictive foreign policies impair economic growth in sub-Saharan Africa. Additionally, countries without access to the sea suffer from comparative disadvantages. After controlling for these factors, the Africa-Dummy is no longer significant. Easterly and Levine (1997) point out that ethnic diversity, measured in units of spoken languages in a country, could influence the economic development in a country. They argue that a strong mixture of different racial groups causes discord about the public resources. Furthermore, diversified societies tend to civil war and lower democratization. The authors are able to explain a large share of the cross-country variation using this measure. Easterly and Levine (1997) link their result to the historical background of sub-Saharan Africa. Like Arcand, Guillaumont and Jeanneney (2000) express, the underlying problem of the continent stems from the 'carve-up' among its occupants during the 19th century. From the authors' viewpoint, this colonial heritage still causes economic drawbacks. Acemoglu, Johnson and Robinson (2001) bring up a historical explanation that is based on the origins of the colonization. Price (2003) also addresses the problem of determining the effects of colonial heritage on economic growth in sub-Saharan Africa.

Adding variables to the growth regression in order to explain the Africa-Dummy is critical. The extra variables identify unique characteristics of sub-Saharan Africa and therefore act like the Africa-Dummy. For example Levine and Renelt (1992) test the causality of different explanatory variables in growth regressions. They summarize that most of the included variables are not robust and dependent on the model. Collier and Gunning (1999) note that the addition of explanatory variables transfers the puzzle elsewhere. Furthermore, many explanatory variables that are added in growth regressions do not necessarily identify drivers for growth. Instead they are somehow correlated to what is not explained by the growth model.

The naive way in which explanatory variables are added or deleted from growth models motivates to only use the explanatory variables given by Mankiw, Romer and Weil (1992) and to accept that the Africa-Dummy is present in the data. In this situation the task is to derive statistical facts about it. First of all, we discuss how to estimate the Afrca-Dummy. Hoeffler (2002) addresses this problem and finds that the significance of the Africa-Dummy disapears when applying the System GMM. In this paper we discuss the disadvantages of the System GMM and introduce a new method, namely the Two-Groups Least-Square Dummy-Variable estimator. This estimation method has the advantages that it is able to estimate the Africa-Dummy in one regression step, that it is consistent even if the residuals are autocorrelated, that it is able to control for all fixed effects and that it does not need the assumption of equal variances of the fixed effects. Estimating the coefficients of the growth regression with the Two-Groups Least-Square Dummy-Variable estimator

identifies a negative significant Africa-Dummy. This clear punishment for sub-Saharan African economies increases if the return to investment in physical capital decreases, if the return the depreciation rate increases or if the return to school attainment increases. We check that the Africa-Dummy is not a result of misspecification of the functional structure, as it does not disappear when applying a semiparamteric extension of the Two-Groups Least-Square Dummy-Variable estimator. Furthermore, we add interaction effects and observe that sub-Saharan African countries have clearly positive returns to the depreciation rate and the Africa-Dummy is even positive and significant. Finally, we estimate the evolution of the Africa-Dummy within the period we observe data. The main result is that, when estimating exactly the regression equation that is motivated by the Augmented Solow Model, we observe that is becomes insignificant and even positive in the recent years.
The paper is structured as follows. Section (3.2) is divided into three subsections. Subsection (3.2.1) describes how the data are collected and subsection (3.2.2) describes how business cycles are removed. Many authors conduct growth regression with numerous variables and understand growth as a theory of everything. In this paper growth regressions are all justified by the augmented Solow model. It is briefly described in subsection (3.2.3). Section (3.3) deals with statistical methods to identify the Africa-Dummy. It is divided into five subsections. Subsection (3.3.1) is about the underlying statistical model and contains some notes about running the growth regressions. Subsection (3.3.2) deals with the System GMM estimator and comments on its disadvantages. Subsection (3.3.3) discusses estimators based on error components models and subsection (3.3.4) concentrates on estimating with fixed effects. The Two-Groups Least-Square Dummy-Variable estimator is introduced and identified as the best estimator to estimate country-specific dummy variables. Finally, subsection (3.3.5) gives results on identifying the Africa-Dummy and estimating the correlations of the Africa-Dummy and other coefficients.
Section (3.4) uses extensions of the Two-Groups Least-Square Dummy-Variable estimator to derive facts about the Africa-Dummy. First of all, subsection (3.4.1) relaxes the functional structure of the regression equation and checks if the Africa-Dummy is a result of a misspecification of the functional structure. Second subsection (3.4.2) estimates the interaction effects of the Africa-Dummy. Third, in subsection (3.4.3), a model is introduced that estimates one Africa-Dummy for each year in the observation period.
Section (3.5) finally concludes.

3.2 Growth Regression and the Africa-Dummy

3.2.1 Data Collection

The objective is to collect long time-series for as many countries as possible for which we can guarantee good data quality. The information sources for the empirical investigation

are the Penn World Table 6.3 (PWT), World Bank's World Development indicators and Barro and Lee (2010). Except of population growth and human capital, all data come from the PWT. It collects a broad range of macroeconomic time-series for almost all countries published by Heston, Summers and Aten (2009). The beginning of a widespread availability is 1960. Most variables are published until 2007, so that observations are obtained for 48 periods. The sample could have been increased significantly but the quality of the data for some countries is insufficient. Heston, Summers and Aten (2009) introduce a country rating system based on the number of participations in worldwide benchmark surveys, the variation of the accessible data and the quality of the statistical methods applied. This results in a grading scheme from A to D with descending order. A rating of D is regarded as too weak to be included in the sample. Therefore, only countries with a grading from A to C are incorporated in the sample. Furthermore, we only incorporate complete time-series for the relevant variables from 1960 to 2007. This also excludes countries that where separated in a sub-period, for example Germany and the countries of the Soviet Union. We excluded these countries because their incorporation would have made it necessary to unify several countries to one country or to split one country in a given period in several countries. The loss of data quality when doing this is unclear. We ended up with 81 complete time-series, one for each of the 81 countries. The time-series are 48 years long. The total sample size is therefore 3888.

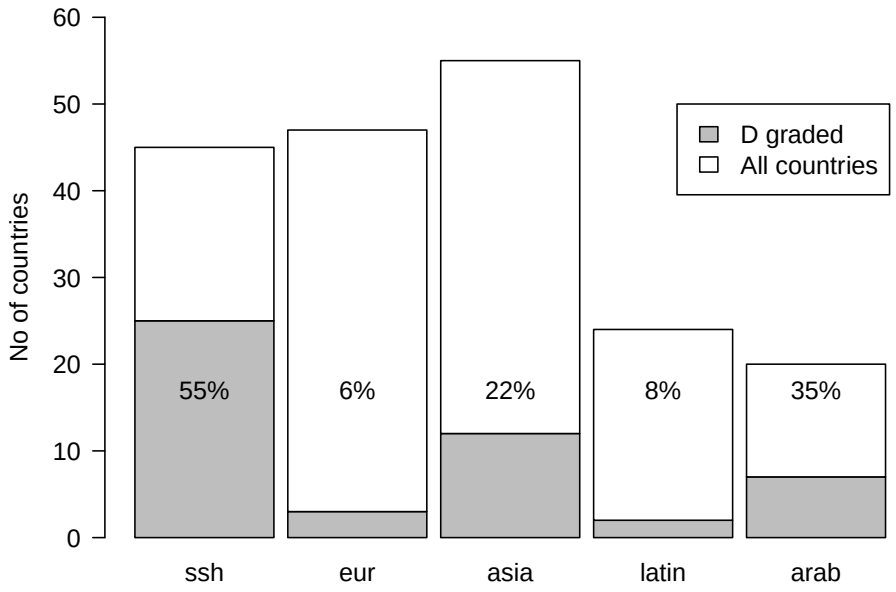

Figure 3.1: *D grading in the PWT*

The selection process of the data can cause a problem. Figure (3.1) shows that sub-Saharan African countries are much more often affected by D grading than other regions. In general, poor countries have weaker databases and are more likely to be excluded. The

question is, whether the exclusion of the D graded data causes a significant violation of the information that the whole sample would inhabit. Since we cannot reliably compare the excluded and the included data, we cannot fully answer this question. However, when deleting a poor country from the data set we make the countries of the reduced data look richer than the countries of the original sample where. This means that, if there is a bias when estimating the Africa-Dummy with the reduced sample, it is likely to have the direction that it underestimates the punishment of sub-Saharan African countries. In the same way, excluding the countries that where separated can cause a problem. The countries that are excluded as a result of this rule do not show structurally similarities. Therefore, if there is a sample selection bias resulting from this rule, we assume it to be small. Table (3.1) lists the countries included in the data set.
The preparation of the variables mainly follows Hoeffler (2002) and Caselli (2005). Because economic growth is a consequence of changes in the production function, the output of the economy is measured as the real per worker gross domestic product (GDP). This is a more precise measure of the country's potential than the per capita GDP because it answers the question how much each productive factor contributes on average to the growth in its country. Per capita figures give information about the available income for the average individual but since the participation rate in the workforce differs a lot, the per capita GDP would be a distorted indicator of the production volume of the total workforce. We denote the logarithm of the per worker GDP of country i at time t by y_{it}.
The population growth refers to the working age population which is defined in the PWT as all individuals from 15 to 64 years. We use the data for the total population and multiply them with the share of adults in working age. We denote the growth rate of the working age population of country i at time t by n_{it}. Data for depreciation rates are not available. In the literature there is accordance, as explained by Mankiw, Romer and Weil (1992), to expect the capital to wear out by 3% per year. Similarly, the advance in productivity is 2% per year for all countries. Therefore, the term $\ln(\delta + g + n_{it})$ is approximated by $\ln(0.05 + n_{it})$. We denote the logarithm of the depreciation rate of country i at year t by lnn_{it}.
The saving rate of the economy is approximated by the relative investment share of the real GDP. These data should correctly measure the savings in the case of a closed economy. We denote the logarithm of the share of country i at year t by $lnsk_{it}$.
The proxy for human capital is the educational attainment data from Barro and Lee (2010). While the PWT contains yearly data, the data from Barro and Lee (2010) are given in five years frequencies. The beginning of the observation period is 1950 and the end is 2010. In order to transfer this variable into a yearly frequency, the missing values are extrapolated by interpolation splines. When doing this, we have to be careful that we do not add an artificial parametric structure to the data. Figure (3.2) shows a graph of the educational attainments when applying natural splines. The points show the data obtained from Barro and Lee (2010) and the lines represent the natural spline functions. Figure (3.2) is

Table 3.1: *Countries*

Code	Country	Code	Country	Code	Country
ARG	Argentina	AUS	Australia	AUT	Austria
BEL	Belgium	BEN	Benin	BGD	Bangladesh
BOL	Bolivia	BRA	Brazil	BRB	Barbados
BWA	Botswana	CAN	Canada	CHE	Switzerland
CHL	Chile	CHN	China	CMR	Cameroon
COG	Congo	COL	Colombia	CRI	Costa Rica
DNK	Denmark	DOM	Dominican Republic	ECU	Ecuador
EGY	Egypt	ESP	Spain	FIN	Finland
FJI	Fiji	FRA	France	GBR	United Kingdom
GHA	Ghana	GRC	Greece	GTM	Guatemala
HKG	Hong Kong	HND	Honduras	IDN	Indonesia
IND	India	IRL	Ireland	IRN	Iran
ISL	Iceland	ISR	Israel	ITA	Italy
JAM	Jamaica	JOR	Jordan	JPN	Japan
KEN	Kenya	KOR	Korea	LKA	Sri Lanka
MEX	Mexico	MLI	Mali	MUS	Mauritius
MWI	Malawi	MYS	Malaysia	NER	Niger
NGA	Nigeria	NLD	Netherlands	NOR	Norway
NPL	Nepal	NZL	New Zealand	PAK	Pakistan
PAN	Panama	PER	Peru	PHL	Philippines
PRT	Portugal	PRY	Paraguay	ROM	Romania
RWA	Rwanda	SEN	Senegal	SGP	Singapore
SLE	Sierra Leone	SLV	El Salvador	SWE	Sweden
SYR	Syria	THA	Thailand	TTO	Trinidad Tobago
TUN	Tunisia	TUR	Turkey	TZA	Tanzania
URY	Uruguay	USA	USA	VEN	Venezuela
ZAF	South Africa	ZMB	Zambia	ZWE	Zimbabwe

representative for all countries. They all have monotonically and linearly increasing shape. Since the points do no not fluctuate a lot, we assume that the approximation error is sufficiently small. We denote the logarithm of the educational attainment data from Barro and Lee (2010) of country i and year t by $lnattain_{it}$.

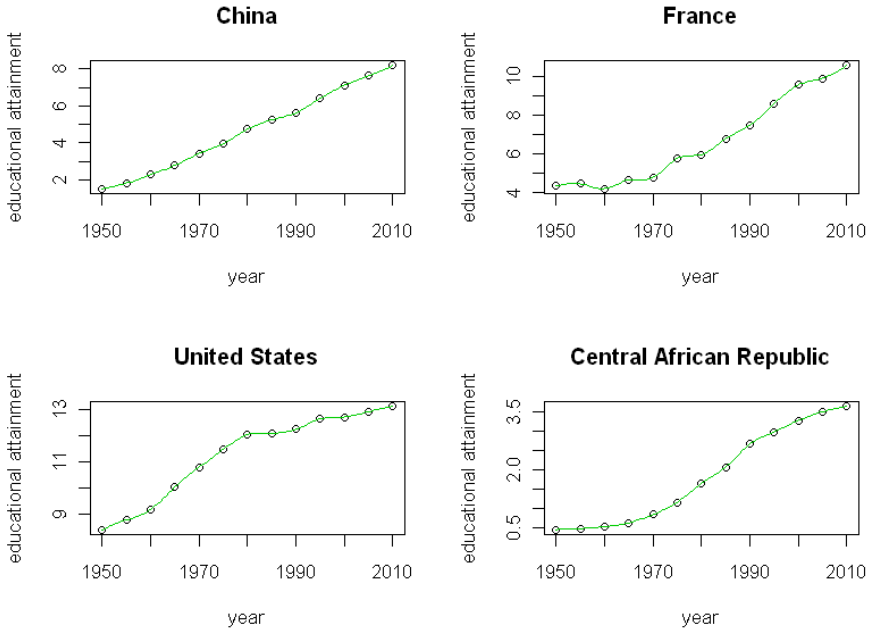

Figure 3.2: *Interpolation of schooling*

3.2.2 Smoothing

We collected four time-series, namely y_{it}, lnn_{it}, $lnsk_{it}$ and $lnattain_{it}$ for each country i. These time-series have a short term cyclical component and a trend component. The Solow model addresses long run growth but not the cyclical fluctuations. Therefore, we smooth the data. As the series have different magnitudes of short term fluctuations they have to be treated in different ways. The series lnn_{it} and $lnattain_{it}$ have only negligible short term fluctuations and are therefore not to be smoothed. The series $lnsk_{it}$ and y_{it} have severe cyclical components.
First of all, we consider the GDP per worker time-series. The easiest approach is linear smoothing. It suggests taking the arithmetical averages over several years of the GDP's per worker so that for this sub-period only the mean enters the dataset. The most common choice is the average over five years. Figure (3.3) shows the resulting growth rates when applying five years averages. It shows the time-series of four countries that serve as examples. The green points are the unsmoothed data. The horizontal black lines

demonstrate the choice of the time periods. Their heights show the reduction of these five points to one value. The blue bullets are the middle time points of each period. The sample lasts from 1960 to 2007 so that data have to be excluded in order to obtain time periods of the same length, namely five years. We excluded the values of the years 1960, 2006 and 2007. These points are labeled with a red star in figure (3.3) and their information is fully lost. Especially in case of the Philippines where the last two observations represent unusual jumps it seems not adequate to exclude this information from the sample. This is the first disadvantage of linear smoothing. Moreover, the long run growth variation within the time periods is fully lost. Another problem is the simultaneous smoothing of different time-series that interact. For example the series $lnsk_{it}$ has a different cyclical component than y_{it}. Taking five year averages smoothes these series in the same naive manner so that the interactions of the long term components of the series are distorted. This problem is especially severe when combining linear smoothed series with unsmoothed series. It is not clear which values of the unsmoothed series should represent each time period. The average however leads to over-smoothing and taking the starting values of each time period would mean to make lagged variables enter the regression.

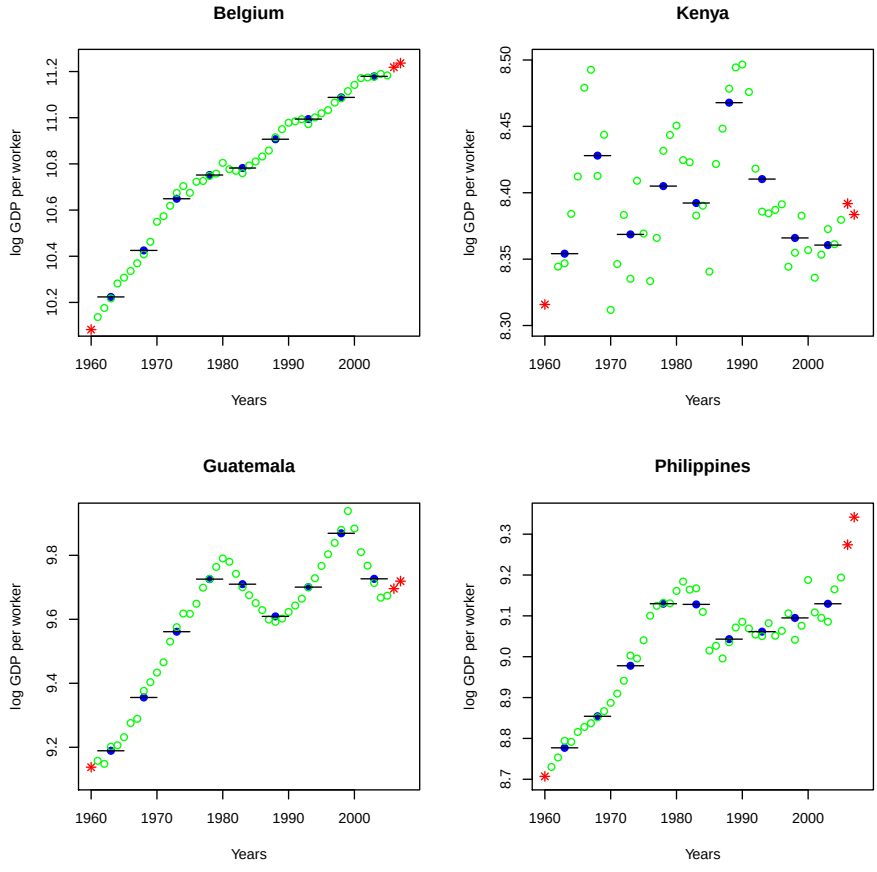

Figure 3.3: *Five years averages*

The disadvantages of linear smoothing give rise to look for another technique to remove business cycles. A prominent example is the Hodrick and Prescott filter. The HP filter decomposes a macroeconomic time-series $\tilde{\tau}_t$ in a structural trend component τ_t, which accounts for sustainable long-run growth and a cyclical component c_t. In Hodrick and Prescott (1997) it is shown how these elements can be separated. The series τ_t is obtained due to

$$\min_{\tau_t} \sum_{t=1}^{T} (\tilde{\tau}_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1} ((\tau_{t+1} - \tau_t)(\tau_t - \tau_{t-1}))^2 .$$

The first term can be interpreted as measuring the goodness-of-fit of the trend component with respect to the original series. The second part punishes for a high variation in the transformed series $\tilde{\tau}_t$. Note that minimizing the variation and maximizing the goodness of fit at the same time is a trade-off problem which is quantified by λ. The higher λ, the more variation is removed from the data. For the choice of this parameter, there are rather weak causal rationales. Hodrick and Prescott (1997) argue that $\lambda = 1600$ is a reasonable choice for quarterly data which intuitively corresponds to a value of 400 for yearly data. On the other hand, Baxter and King (1999) argue that λ should be chosen as the fourth power of a change in the frequency. In our case this corresponds to 6.25. After observing the different outputs of the smoothing with different smoothing parameters, we decided to chose $\lambda = 100$. Figure (3.4) shows the smoothed series of the yearly growth rates of the four countries Belgium, Kenya, Guatemala and Philippines. The grey points are the unsmoothed data. The smoothed data are connected with lines. It can clearly be seen that the disadvantages of linear smoothing are not shared by the HP filter.

When smoothing the series of $lnsk_{it}$ it is hard to derive the adequate smoothing parameter of one series from that of the other series. On the one hand, the series $lnsk_{it}$ have more variation than y_{it}. On the other hand the former series are of much smaller magnitude than the latter. Smoothing the two series simultaneously means that one series should not appear to be over-smoothed compared to the other. Having this in mind, we choose the smoothing parameter of $lnsk_{it}$ by visual judgment. After observing the outputs of smoothed series for different smoothing parameters we decided that $\lambda = 25$ is the appropriate parameter. The result is given in figure (3.5). The HP filter performs satisfying and is therefore selected to smooth the data.

3.2.3 The Augmented Solow Model

The neoclassical growth theory is based on the work by Solow (1956) and Swan (1956). The theory of human capital accumulation tries to account for enhancements in technology by replacing homogeneous work with education-based improvements of workers that are regarded as investments in quality. Mankiw, Romer and Weil (1992) extend the Solow model by human capital, test this 'augmented' version and observe significant

improvements in explanatory power. This model is the basis for growth regressions. We briefly describe it in what follows.
The economic agents are households and firms. Furthermore, there are three commodities and for each commodity there is a market. The commodities are output, capital and labor. When considering the corresponding markets we assume that all individuals behave rational and further information restrictions are not present. In the market for capital we think of households owning the capital $K(t)$ and lease it to the firms. The firms demand the capital $K^D(t)$. The price is $r(t)$ (real rental rate). In the market for labor the supply $\tilde{L}(t)$ comes from the households and the demand $\tilde{L}^D(t)$ comes from the firms. The price in the labor market is $w(t)$ (real wage rate). $\tilde{L}(t)$ is not a measure of headcount. It can be decomposed in a measure of working quality and a measure of the homogeneous supply per person. We decompose

$$\tilde{L}^D(t) = L(t)^{\frac{1-\alpha-\beta}{1-\alpha}} H(t)^{\frac{\beta}{1-\alpha}},$$

where $H(t)$ is the amount of human capital. In the market for output the supply consists of the total output of firms $Y(t)$ and the demand $Y^D(t)$ consists of what the households save

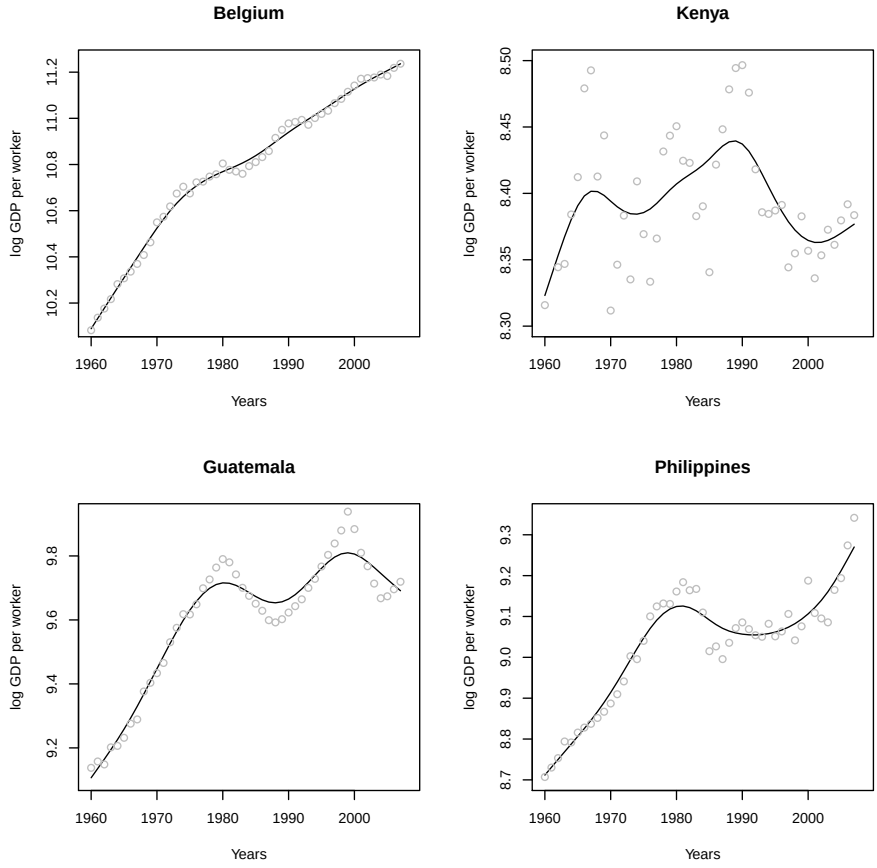

Figure 3.4: *HP Smoothing of* y_{it}

($S(t) = sY^D(t)$) and what they consume ($C(t) = (1-s)Y^D(t)$). We assume that investments equal savings ($I(t) = S(t)$). The households decide how to distribute total savings between gross investment and human capital. We assume that the summarized result of the households' decisions is that the fraction $s_K Y^D(t) = I_K(t)$ is invested in physical capital and the fraction $s_H Y^D(t) = I_H(t)$ is invested in human capital. Clearly $s_K + s_H = s$ and $I(t) = I_H(t) + I_K(t)$. The supply of output follows a production function with the input factors capital $K^D(t)$ and labor $L^D(t)$. The generated output is also influenced by the productivity $A(t)$ that characterizes the country's transformation capabilities. The improvement may consist of either of level of technology or of efficiency gains, meaning the ability to combine the input factors in an optimal way. The aggregated production function is

$$\begin{aligned} Y(t) &= K^D(t)^{\alpha} H(t)^{\beta} (L(t)A(t))^{1-\alpha-\beta} \\ &= A(t)^{1-\alpha-\beta} K^D(t)^{\alpha} \tilde{L}^D(t)^{1-\alpha}. \end{aligned} \tag{3.1}$$

The price of the output market is normalized to one, so that other prices are measured in units of the output price. Note that the fundamental difference between the input factors of

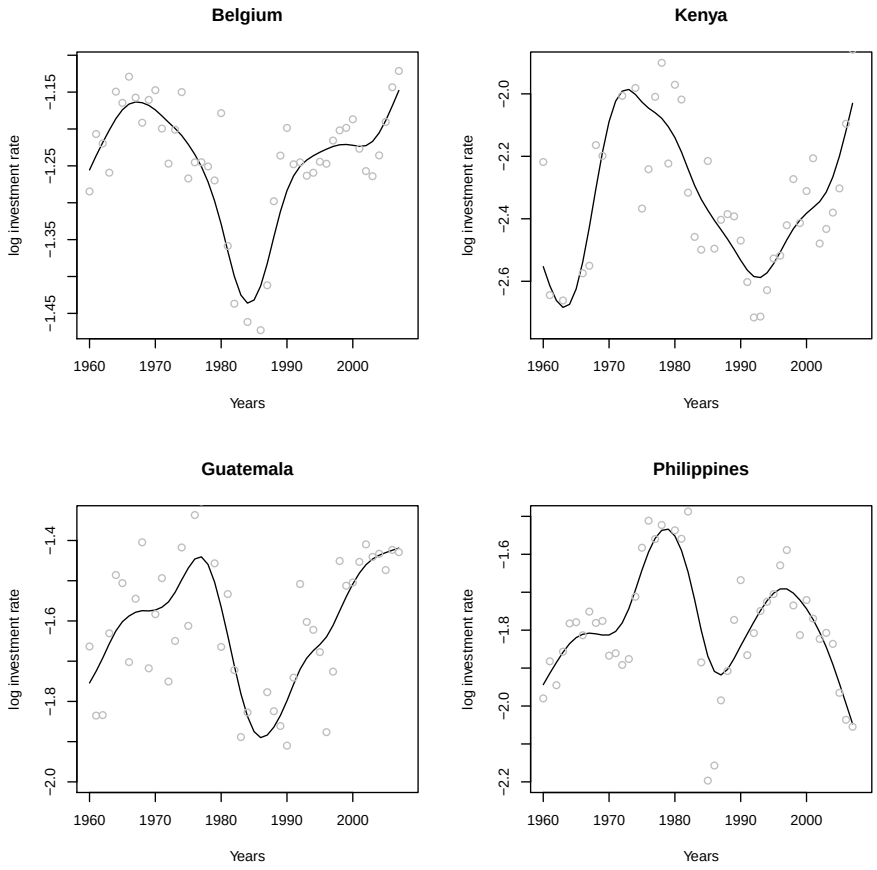

Figure 3.5: *HP Smoothing of* $lnsk_{it}$

the production function is that capital and labor are rival goods while the applied technology can spillover to any entrepreneur in the economy what means that it is a public good.

All three markets are assumed to be perfectly competitive so that the economic agents take the prices as given and in each market the appropriate price adjusts such that $L^D(t) = L(t)$ and $K^D(t) = K(t)$ and $Y^D(t) = Y(t)$. Necessarily, the two inputs capital and labor are paid at their marginal products. Therefore, it holds

$$\frac{\partial Y(t)}{\partial K(t)} = r(t) \tag{3.2}$$

and

$$\frac{\partial Y(t)}{\partial L(t)} = w(t). \tag{3.3}$$

In this setting, α is the capital intensity and β is the labor intensity in the production process. The labor force $L(t)$ and the productivity level $A(t)$ are assumed to grow exogenously at rates n and g respectively. Therefore, it holds

$$L(t) = L(0)\exp(nt) \tag{3.4}$$

and

$$A(t) = A(0)\exp(gt). \tag{3.5}$$

In any period, the investment of the prior period will be transformed into new capital minus the depreciation δ of the old capital stock. We express the production process in terms of per effective worker units. ($k(t) = K(t)/A(t)L(t)$ and $y(t) = Y(t)/A(t)L(t)$). The growth of per effective worker capital over time is

$$\dot{k}(t) = s_K y(t) - (n + g + \delta)k(t). \tag{3.6}$$

Human capital behaves like its physical counterpart. The evolution of the per effective worker unit of human capital ($h(t) = H(t)/A(t)L(t)$) is

$$\dot{h}(t) = s_H y(t) - (\delta + g + n)k(t). \tag{3.7}$$

The model is (3.1), (3.2), (3.3), (3.4), (3.5), (3.6) and (3.7). The parameters of the model are α, β, s_K, s_H, δ, n and g. Given the initial values $A(0)$, $K(0)$, $H(0)$ and $L(0)$, the model will determine the dynamic evolution of the economy. Moreover, when assuming diminishing returns to capital input ($\alpha + \beta < 1$) the model converges as time goes to infinity. The situation of convergence is called steady state. It is identified by $\dot{k} = \dot{h} = 0$. In the steady state it holds that

$$\begin{aligned} k(t) \equiv k^* &= \left(\frac{s_k^{1-\beta} s_H^{\beta}}{\delta + g + n} \right)^{1/(1-\alpha-\beta)} \\ h(t) \equiv h^* &= \left(\frac{s_k^{\alpha} s_H^{1-\alpha}}{n + g + \delta} \right)^{1/(1-\alpha-\beta)} . \end{aligned} \tag{3.8}$$

Growth outside the steady state is determined by the evolution of $k(t)$ given in (3.6) and of $h(t)$ given in (3.7). The capital stock of human and physical capital increases if the economy fosters investments (s_H or s_K increases), or if the effective depreciation $(n+g+\delta)$ decreases. The model makes quantitative predictions about the speed of convergence to steady state. Approximating around the steady state, the speed of convergence at a given time point t outside the steady state is given by

$$\frac{\partial \ln(y(t))}{\partial t} = \lambda(\ln(y^*) - \ln(y(t))), \tag{3.9}$$

with $\lambda = (n+g+\delta)(1-\alpha-\beta)$. Equation (3.9) implies that

$$\ln(y(t)) - \ln(y(0)) = (1-\exp(\lambda t))\ln(y^*) - (1-\exp(\lambda t))\ln(y(0)), \tag{3.10}$$

where $y(0)$ is the income per effective worker at some initial date. This implies that if the economy moves from the initial state 0 to the time point t halfway to steady state, it holds

$$\frac{1}{2} = \frac{\ln(y(t)) - \ln(y(0))}{\ln(y^*) - \ln(y(0))} = 1 - \exp(\lambda t)$$

which is equivalent to $t = \ln(2)/\lambda$. If for example $\lambda = 0.02$ the economy moves halfway to steady state in 34.65736 years. In general the larger λ the less time it takes the economy to move halfway to steady state.
Equation (3.10) implies that outside the steady state it holds

$$\begin{aligned}
&\ln(\frac{Y(t)}{L(t)}) = (1-\exp(-\lambda t))\ln(A(0)) + gt + \exp(-\lambda t)\ln(\frac{Y(0)}{L(0)}) \\
&(1-\exp(-\lambda t))\frac{\alpha}{1-\alpha-\beta}\ln(s_K) + (1-\exp(-\lambda t))\frac{\beta}{1-\alpha-\beta}\ln(s_H) \\
&-(1-\exp(-\lambda t))\frac{\alpha+\beta}{1-\alpha-\beta}\ln(n+g+\delta).
\end{aligned}$$

This justifies the following regression equation

$$y_{it} = \rho * y_{i(t-1)} + \beta_1 * lnn_{it} + \beta_2 * lnsk_{it} + \beta_3 * lnn_{it} + \eta_i + \nu_{it}, \tag{3.11}$$

where ν_{it} is an error with expectation zero.

3.3 Identifying the Africa-Dummy

3.3.1 Growth Regressions

Sampling Process: We denote the information of the dependent variable from some initial time point 1 up to t by $y_i^t = (y_{i1}, \ldots, y_{it})$ and the information of the exogenous variables

from some initial time point 2 up to t by $x_i^t = (x_{i2}', \ldots, x_{it}')$. We assume that $\left\{ \left(y_i^T, x_i^T \right) \ , \ i = 1, \ldots, n \right\}$ is a number of independent observations from the same probability distribution, with finite first and second order moments.
Regression Equation: We are aiming for estimating (3.11) with the Africa-Dummy. (3.11) is of the form

$$y_{it} = \rho y_{i(t-1)} + x_{it}' \beta + \eta_i + \nu_{it}. \tag{3.12}$$

The Africa-Dummy is a part of the country-specific effects

$$\eta_i = \eta_g + SSH * 1_{SSH,i} + \tilde{\eta}_i, \tag{3.13}$$

where $E(\tilde{\eta}_i) = 0$, $1_{SSH,i}$ equals 1 if country i belongs to the group of sub-Saharan African countries and 0 else and η_g is the common intercept. When plugging (3.13) in (3.12) we have

$$y_{it} = \eta_g + \rho y_{i(t-1)} + x_{it}' \beta + SSH * 1_{SSH,i} + \tilde{\eta}_i + \nu_{it}. \tag{3.14}$$

We aim to estimate the parameters ρ, β, η_g, SSH and each $\tilde{\eta}_i$.
Exogeneity: We assume

$$E(\nu_{it} | 1_{SSH,i}, y_i^{t-1}, x_i^T, \tilde{\eta}_i) = 0. \tag{3.15}$$

An implication of the assumption is that the errors ν_{it} are conditionally serially uncorrelated. Namely for $j > 0$ it holds

$$E(\nu_{it} \nu_{i(t-j)} | 1_{SSH,i}, y_i^{t-1}, x_i^T, \tilde{\eta}_i) = 0.$$

By the law of iterated expectations it also holds that

$$E(\nu_{it} \nu_{i(t-j)}) = 0.$$

Second Moments of the Errors: We assume

$$E(\nu_{it} \nu_{js}) = \begin{cases} \sigma_\nu^2, & \text{if } i = j \text{ and } s = t \\ 0, & \text{else.} \end{cases} \tag{3.16}$$

This is a very strict assumption. When looking at how to estimate equation (3.14) we discuss what happens if this assumption is violated. Furthermore, we assume that the second moments of the country-specific errors exist.

The Country-Specific Effects: We observe $\left\{ \left(y_i^T, x_i^T \right) \ , \ i = 1, \ldots, n \right\}$ but we do not observe the country-specific intercepts. The model by Mankiw, Romer and Weil (1992) indicates that the total country-specific effect η_i is determined by the growth rate of technological change g, the convergence rate λ and the initial level of technology $A(0)$. g and λ are assumed not to change between countries and over time. The initial endowment with production technology cannot be expected to be constant in all countries. Mankiw, Romer

and Weil (1992) mention several influences on $A(0)$ like resources, climate or institutions. They decompose $A(0)$ in a common component that reflects the general productivity and a component that reflects all country-specific characteristics. The assumption that $\tilde{\eta}_i$ and x_{it} are uncorrelated seems to be too strong. For example developed institutions can increase the level of human capital in the population. We assume that $\tilde{\eta}_i$ is in general correlated to every $y_{i(s-1)}$ and x_{is} for all i and s.

Vector-Matrix-Notation: First of all, we stack the time-series data of (3.12):

$$\begin{aligned}
\iota &= (1,\ldots,1)' \in \mathbb{R}^{T-1} \\
y_i &= (y_{i2},\ldots,y_{iT})' \in \mathbb{R}^{T-1} \\
y_{i(-1)} &= (y_{i1},\ldots,y_{i(T-1)})' \in \mathbb{R}^{T-1} \\
X_i &= (x_{i2},\ldots,x_{iT}) \in \mathbb{R}^{K\times(T-1)} \\
\nu_i &= (\nu_{i2},\ldots,\nu_{iT})' \in \mathbb{R}^{T-1}.
\end{aligned}$$

Equation (3.12) is

$$y_i = \rho y_{i(-1)} + X_i'\beta + \eta_i\iota + \nu_i \in \mathbb{R}^{T-1}.$$

Furthermore, we stack cross-sectional data:

$$\begin{aligned}
y &= (y_1',\ldots,y_n')' \in \mathbb{R}^{n(T-1)} \\
y_{-1} &= (y_{1(-1)}',\ldots,y_{n(-1)}')' \in \mathbb{R}^{n(T-1)} \\
X &= (X_1,\ldots,X_n)' \in \mathbb{R}^{n(T-1)\times K} \\
C &= I_n \otimes \iota \in \mathbb{R}^{n(T-1)\times n} \\
\eta &= (\eta_1,\ldots,\eta_n)' \in \mathbb{R}^{n} \\
\nu &= (\nu_1',\ldots,\nu_n') \in \mathbb{R}^{n(T-1)}.
\end{aligned}$$

Equation (3.12) is

$$y = \rho y_{-1} + X\beta + C\eta + \nu \in \mathbb{R}^{n(T-1)}. \tag{3.17}$$

(3.14) is stacked in the same way. We assume without loss of generality that the data are available in the form that exactly the first s rows belong to the group of sub-Saharan African countries. Denote

$$\begin{aligned}
\tilde{\eta} &= (\tilde{\eta}_1,\ldots,\tilde{\eta}_n)' \in \mathbb{R}^n, \\
\iota_{n(T-1)} &= (1,\ldots,1)' \in \mathbb{R}^{n(T-1)} \text{ and} \\
\iota_{n(T-1),SSH} &= (\underbrace{1,\ldots,1}_{\in\mathbb{R}^{s(T-1)}}, \underbrace{0,\ldots,0}_{\in\mathbb{R}^{(n-s)(T-1)}}) \in \mathbb{R}^{n(T-1)}.
\end{aligned}$$

(3.14) is in stacked form

$$y = \iota_{n(T-1)}\eta_g + \rho y_{-1} + X\beta + \iota_{n(T-1),SSH} * SSH + C\tilde{\eta} + \nu \in \mathbb{R}^{n(T-1)}. \tag{3.18}$$

The Bias of the Within Group Estimator: Regression equations (3.12) and (3.14) have a lagged dependent variable. Therefore, assuming exogeneity with respect to all variables, including the lagged dependent variable, will cause a bias when estimating the coefficients. This has been shown by Orcutt and Irwin (1948) and Kendall (1954) for time-series models with fixed time-series length and has been extended by Nickell (1981) for panels with fixed T (even if $n \to \infty$). In consequence, bias reduction procedures have been proposed, for example Kiviet (1995), Hahn and Kuersteiner (2002) or Phillips and Sul (2007). In the estimation methods that will be presented, the bias only occurs in the regression step where the β's are estimated. Except of the Random Effects estimator, this regression step is the same as applying the Within Group estimator. Therefore, we estimate the bias of Within Group estimator using the precise formulas as $n \to \infty$ given by Phillips and Sul (2007). Using

$$\bar{\eta}_j = y_{j\bullet} - \hat{\rho}_{WG} y_{-1\,j\bullet} - x'_{i\bullet}\hat{\beta}_{WG}$$

we can then see how mistakes in the Within Group estimation step affect the estimation of the fixed effects. Afterwards we can estimate the bias of SSH using

$$\hat{SSH} = \bar{\eta}_A - \bar{\eta}_{NA}.$$

Since the true ρ is not known, we calculate biases for different ρ's. The Within Group estimator of the coefficient of the lagged variable is biased downwards and therefore we use it as the smallest ρ to plug in. We calculate these biases for $\hat{\beta}_{WG}$ since fluctuations result in negligible small differences. The results are given in table (3.2). The biases of the fixed effects listed in this table are the maximum of all absolute values of the biases of each fixed effect. Table (3.2) shows that all biases, apart from that of the coefficient of the lagged variable, are negligible small. Calculating biases when adding more exogenous variables is not necessary since Phillips and Sul (2007) argue that the addition of exogenous variables result in smaller biases. We therefore assume in regressions using the Within Group estimator that the bias that results from the lagged variable (apart from that of the coefficient of the lagged variable itself) is negligible small.

Two-Step Regressions: There are to ways to estimate the Africa-Dummy. The two-step method first estimates (3.12) together with the country-specific effects which contain the Africa-Dummy according to decomposition (3.13). In the second step the estimated country-specific effects are used to estimate equation (3.13) and to obtain an estimator for the Africa-Dummy. This method has the disadvantage that it does not use all the available information from the correlations between the different variables of (3.14). The result is a

Table 3.2: *Biases*

True Value	Bias		
ρ	ρ (10^{-2})	lnn (10^{-4})	lnsk (10^{-4})
0.98971	-1.4298	7.1027	5.1380
0.99314	-1.3815	6.8523	3.7093
0.99657	-1.3302	6.5506	2.2294
1.00000	-1.2760	6.1961	0.7095
ρ	lnattain (10^{-4})	FE (10^{-17})	SSH (10^{-17})
0.98971	-0.6022	2.9554	4.0494
0.99314	0.5347	2.8576	3.9813
0.99657	1.7158	2.7540	3.9082
1.00000	2.9316	2.6441	3.8292

consistent estimator with a large variance because the errors made in the first regression step persist in the second regression step which itself generates an error. Moreover, this method is not able to correctly estimate the correlations between the coefficients. The other estimation method estimates (3.14) directly and does not share these disadvantages.

Lags: Running the regressions using exactly (3.14) has three drawbacks. First, the one year growth time-series shows little variation so that the coefficient of the lagged dependent variable is almost one and all other coefficients are very small. This is often called spurious regression problem. Second, we only checked that the endogeneity bias caused by the lagged dependent variable is small. Since the economy can choose its growth driving parameters as reaction of a shock, the regression is suspected to suffer from an endogeneity bias. It is natural to assume that the bias caused by the explanatory variables is much smaller than that caused by the lagged dependent variable itself, which is already negligibly small. Nevertheless, we do not know the exact correlation of explanatory variables and the error and cannot give precise formulas for the bias as done by Phillips and Sul (2007). Third, we aim for comparison of our results with that of other authors, who refer their regressions to five year time horizons taking either averaged or initial explanatory variables to represent the time horizons (see Hoeffler (2002)). Especially the first two drawbacks mentioned allow impeaching the credibility of the results obtained by the one year growth equation. Therefore, we estimate a lagged regression equation. Taking lagged variables has two drawbacks. First, we move away from the situation described by Mankiw, Romer and Weil (1992) and loose theoretic justification. Second, the model by Mankiw, Romer and Weil (1992) deals with the evolution of the differences of the logarithms of the subsequent GDP's. These can only be interpreted as growth rates if the

subsequent GDP's are close to each other, since in this case a Taylor-Expansion shows that

$$\ln(GDP_t) - \ln(GDP_{t-1}) \approx \frac{GDP_t - GDP_{t-1}}{GDP_{t-1}}.$$

Time horizons from $t-5$ to t generate larger differences between the two growth rates than time horizons from $t-1$ to t.
We obtain results with a regression with a five year lagged dependent variable and five year lagged explanatory variables. Therefore $x_{it} = (lnn_{i(t-5)}, lnsk_{i(t-5)}, lnattain_{i(t-5)})$ and the dependent variable is the five year lagged GDP per worker. We also run the regression with a one year lagged dependent variable and contemporary explanatory variables and compare the results. Note that, if we obtain similar results, the aforementioned problems of spurious regression and endogeneity for the one year lagged regression are very unlikely.

Large T problems: We discuss some well-known problems that often occur in case of large n as well as large T panels. We discuss the spurious regression problem, the unit-root problem and the cointegration problem. The spurious regression problem comes from the literature of time-series analysis. The problem is known as one yielding a nonzero β-coefficient when regressing two independent and individually integrated processes of order one on one another. Phillips and Moon (1999) provide a concept that extends the arguments about spurious regression in time-series analysis. They show that the issue of spurious regression will not arise for the panel estimates, when the cross-sectional size tends to infinity. In our case the cross-sectional size is 81 which is why we argue that we do not have the problem of a spurious regression. The Unit-Root problem is concerned with the inference of the autoregressive coefficient, when it equals one. When considering the lagged series, the autoregressive coefficient is far away from one, which is why we argue that this is not a problem in our case. There can also be a problem in case of integrated explanatory variables of order one. More precisely, if $x_{it} = x_{it-1} + \varepsilon_{it}$, Kao and Chiang (2000) show that the fixed effects estimator is biased if ε_{it} and v_{it} are correlated. We see no reason for such a correlation and therefore estimate with OLS.

3.3.2 Why we do not use System GMM

Caselli, Esquivel and Lefort (1996) applied the Difference GMM to growth regression using linear smoothed data with five year time horizons between 1960 and 1985. Bond, Hoeffler and Temple (2001) note that the Difference GMM uses weak instruments because the series of the logarithms of GDP's per capita is highly persistent and recommend the System GMM. Afterwards, many papers appear using System GMM. Roodman (2006) gives access to System GMM by implementing it in Stata. Hoeffler (2002) addresses the problem of estimating the Africa-Dummy in growth regressions and comes to the conclusion that System GMM is the preferred method. We have the impression that the System GMM is the leading method in growth regressions. As most authors use linear

smoothing instead of applying the HP filter, their time-series are shorter which leads to less instruments. The number of instruments when having time-series data with $T = 48$ is very large. This causes problems. Furthermore, Hoeffler (2002) applies a two step method for estimating the Africa-Dummy which leads to efficiency problems. Before discussing these problems we give an account of the System GMM.
First of all, we stack the time-series data of model (3.12) and write it as

$$y_i = W_i\alpha + \eta_i\iota + \nu_i$$

with $W_i = (w_{i2},\dots,w_{iT})' \in \mathbb{R}^{(T-1)\times(K+1)}$, $w_{it} = (y_{i(t-1)}, x_{it}')' \in \mathbb{R}^{K+1}$,
$\nu_i = (\nu_{i2},\dots,\nu_{iT})' \in \mathbb{R}^{T-1}$ and $\alpha = (\rho,\beta')' \in \mathbb{R}^{K+1}$. We assume the feedback assumption

$$E(\nu_{it}|x_{i2},\dots,x_{it},y_{i1},\dots,y_{i(t-1)},\eta_i) = 0. \tag{3.19}$$

This assumption was for example made by Hoeffler (2002) and is based on the idea that the economy can chose its variables as a reaction of a shock. It follows from (3.19) that for $t = 3,\dots,T$

$$E((w_{i2}',\dots,w_{i(t-1)}')'(\nu_{it} - \nu_{i(t-1)})) = 0 \in \mathbb{R}^{(K+1)(t-2)} \tag{3.20}$$

holds. These are $r_{Diff} = (K+1)(T-2)(T-1)/2$ moment conditions. Note that

$$\Delta\nu_i = D\nu_i = (\nu_{i3} - \nu_{i2},\dots,\nu_{iT} - \nu_{i(T-1)})' \in \mathbb{R}^{T-2},$$

with

$$D = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(T-2)\times(T-1)}.$$

Stacking these moment conditions gives

$$E(Z_i'D\nu_i) = 0 \in \mathbb{R}^{r_{Diff}}, \tag{3.21}$$

where

$$Z_i = \begin{pmatrix} w_{i2}' & & & & & \\ & w_{i3}' & w_{i2}' & & & \\ & & & \ddots & & \\ & & & & w_{i(T-1)}' & \dots & w_{i2}' \end{pmatrix} \in \mathbb{R}^{(T-2)\times r_{Diff}}.$$

One can derive the Difference GMM Estimator by applying the usual GMM procedure using (3.21). It was first proposed by Arellano and Bond (1991). Blundell and Bond (1998) show that the instruments of the Difference GMM estimator are weak when the autoregressive coefficient is close to one. The System GMM estimator augments the set of moments of the Difference GMM estimator by additionally assuming moment conditions

for the level equation. When doing this, Blundell and Bond (1998) observe a dramatic efficiency gain when the autoregressive coefficient is close to one. Arellano and Bond (1995) introduce the use of lagged differences as possible instruments for the equation in levels. The equation in levels is (3.12) in the form

$$y_{it} = \eta_g + w_{it}'\alpha + \tilde{\eta}_i + \nu_{it} = \tilde{w}_{it}'\tilde{\alpha} + u_{it}$$

with $\tilde{w}_{it} = (1, w_{it}')' \in \mathbb{R}^{K+2}$, $\tilde{\alpha} = (\eta_g, \alpha')' \in \mathbb{R}^{K+2}$ and $u_{it} = \tilde{\eta}_i + \nu_{it}$. Stacking time-series data gives

$$y_i = \tilde{W}_i\tilde{\alpha} + u_i \in \mathbb{R}^{T-1},$$

with $\tilde{W}_i = (\tilde{w}_{i2}, \ldots, \tilde{w}_{iT})' \in \mathbb{R}^{(T-1)\times(K+2)}$, $u_i = (u_{i2}, \ldots, u_{iT}) \in \mathbb{R}^{T-1}$. We assume that the exogenous variables can be correlated to the fixed effects but the correlations of two succeeding exogenous variables of the same country and the fixed effects have the same magnitude

$$E(\Delta x_{it}\eta_i) = 0. \tag{3.22}$$

Note that $\Delta x_{it} = x_{it} - x_{i(t-1)}$. Furthermore, we assume that the correlations of y_{i2} and η_i and that of y_{i1} and η_i are equal. This is the initial condition

$$E(\Delta y_{i2}\eta_i) = 0. \tag{3.23}$$

In this case, condition (3.23) also holds for all subsequent y_{it}. This means that two succeeding y_{it}'s have the same correlation to η_i. It will be shown in this subsection that especially in the case of growth regressions, where the autoregressive coefficient is close to one, this initial condition is unlikely to be fulfilled. However, it follows from (3.22) and (3.23) that

$$E(\eta_i(y_{i(t-1)} - y_{i(t-2)}, x_{it}' - x_{i(t-1)}')') = E(\eta_i\Delta w_{it}) \text{ for } t = 3, \ldots, T.$$

(3.22) and (3.23) imply the $(T-2)(K+1)$ moment conditions

$$E(\Delta w_{it}u_{it}) = 0 \text{ for } t = 3, \ldots, T.$$

Furthermore, it clearly holds that $E(u_{it}) = 0$. We summarize all these $r_{Lev} = (T-2)(K+2)$ moment conditions by

$$E(Z_{li}'u_i) = 0 \in \mathbb{R}^{r_{Lev}} \tag{3.24}$$

with

$$Z_{li} = \begin{pmatrix} 0 & \ldots & 0 \\ (1, \Delta w_{i3}') & & \\ & \ddots & \\ & & (1, \Delta w_{iT}') \end{pmatrix} \in \mathbb{R}^{(T-1)\times r_{Lev}}.$$

The equation in levels is $y_i = \tilde{W}_i\tilde{\alpha} + u_i \in \mathbb{R}^{T-1}$ with moment conditions $E(Z_{li}'u_i) = 0$. The equation in differences is $Dy_i = DW_i\alpha + D\nu_i = D\tilde{W}_i\tilde{\alpha} + Du_i \in \mathbb{R}^{T-2}$ with moment conditions $E(Z_i'D\nu_i) = E(Z_i'Du_i) = 0$. Stacking the levels equation on the differenced equation yields

$$y_i^\dagger = W_i^\dagger\tilde{\alpha} + u_i^\dagger \in \mathbb{R}^{2T-3},$$

with

$$y_i^\dagger = \begin{pmatrix} y_i \\ Dy_i \end{pmatrix} \in \mathbb{R}^{2T-3},\ W_i^\dagger = \begin{pmatrix} \tilde{W}_i \\ D\tilde{W}_i \end{pmatrix} \in \mathbb{R}^{(2T-3)\times(K+2)} \text{ and } u_i^\dagger = \begin{pmatrix} u_i \\ Du_i \end{pmatrix} \in \mathbb{R}^{2T-3}.$$

Summarizing all $r = r_{Lev} + r_{Diff} = (T-2)(K+2) + (K+1)(T-2)(T-1)/2$ moment conditions yields

$$E((Z_i^\dagger)'u_i^\dagger) = 0 \in \mathbb{R}^r, \tag{3.25}$$

with

$$Z_i^\dagger = \begin{pmatrix} Z_{li} & 0 \\ 0 & Z_i \end{pmatrix} \in \mathbb{R}^{(2T-3)\times r}.$$

The System GMM estimator is

$$\hat{\tilde{\alpha}}_{SysGMM} = \left[(\sum_{i=1}^n (W_i^\dagger)'Z_i^\dagger) A_n (\sum_{i=1}^n (Z_i^\dagger)'W_i^\dagger) \right]^{-1} (\sum_{i=1}^n (W_i^\dagger)'Z_i^\dagger) A_n (\sum_{i=1}^n (Z_i^\dagger)'y_i^\dagger). \tag{3.26}$$

The optimal choice of the weighting matrix A_n is the inverse of $Var((Z_i^\dagger)'u_i^\dagger)$.
Hoeffler (2002) addresses the problem of estimating the Africa-Dummy in growth regressions. She applies a two-step regression, estimating (3.12) with an intercept first and then regressing the residuals on the Africa-Dummy. This method has efficiency problems that result from the variation induced by estimating the residuals in the first step and from the GMM method in general. It is not surprising that Hoeffler (2002) observes that the negative Africa-Dummy becomes insignificant. Furthermore, correlations between the coefficients cannot be calculated with this method. Beside this, the System GMM has more problems. We first discuss the problems that result from too many instruments and motivate to reduce the number of instruments. Then, we discuss the effects of the reduction of the instruments.
One general problem of GMM is a bias that occurs when too many instruments are used (see Tauchen (1986) or Ziliak (1997)). Windmeijer (2005) observes a decreasing bias when applying the Difference GMM if the instrument count is reduced. Arellano (2003) gives analytical evidence for the bias when the number of observations and the length of the time-series go to infinity.

Furthermore, problems occur when estimating the optimal weighting matrix A_n. The number of elements to be estimated is quadratic in the number of instruments and therefore quartic in T. Moreover, the elements of the optimal matrix are fourth moments of the underlying distributions because they are second moments of the result of differenced variables times variables. Roodman (2009) notes that a common symptom for estimations of the weighting matrix is that they are singular. Therefore, the generalized inverse rather than the inverse is calculated. This can give results that are far away from the theoretical ideal. The breakdown tends to occur as the number of instruments approaches n. Therefore, n can be seen as a general benchmark for the number of instruments. We have 4554 instruments when estimating with the System GMM and n equals 81.
The Hansen J-Test (see Hansen (1982)) usually checks the validity of instruments, but as for example Bowsher (2002) observes in simulation studies, a too large number of instruments weakens the test dramatically. Roodman (2009) notes that in case of too many instruments the weights of those moments that are least well satisfied are too small. We conclude that we do not have a reliable test available that tells us how many and which instruments to choose. This problem is especially severe as the initial condition (3.23) is least likely to be fulfilled in case of highly persistent time-series as in our case. To understand this we follow the arguments of Roodman (2009). If there exists a long-run mean, it holds that

$$E(y_{it}|\eta_g, \tilde{\eta}_i, x_i^T) = E(y_{i(t+1)}|\eta_g, \tilde{\eta}_i, x_i^T),$$

which is equivalent to

$$y_{it} = \frac{x'_{i(t+1)}\beta}{1-\rho} + \frac{\eta_g}{1-\rho} + \frac{\tilde{\eta}_i}{1-\rho} \; \forall t.$$

Assuming that there exists such a long-run mean, we define the correlation of the deviations from it to $\tilde{\eta}_i$ by

$$m_{it} = E((y_{it} - (\frac{x'_{i(t+1)}\beta}{1-\rho} + \frac{\eta_g}{1-\rho} + \frac{\tilde{\eta}_i}{1-\rho}))\tilde{\eta}_i).$$

m_{it} has got interesting properties. First of all, if the initial condition (3.23) holds for example in t and therefore $E(\Delta y_{i(t-1)} u_{it}) = 0$, then this is equivalent to $m_{i(t-2)} = 0$. This is because

$$0 = E(\Delta y_{i(t-1)} u_{it}) = E(((\rho - 1) y_{i(t-2)} + x'_{i(t-1)}\beta + \eta_g + \tilde{\eta}_i)\tilde{\eta}_i)$$

is equivalent to

$$0 = E((y_{i(t-2)} - (\frac{x'_{i(t-1)}\beta}{1-\rho} + \frac{\eta_g}{1-\rho} + \frac{\tilde{\eta}_i}{1-\rho}))\tilde{\eta}_i) = m_{i(t-2)}.$$

Furthermore, if assumption (3.22) holds, it follows that $m_{it} = \rho m_{i(t-1)}$. This is because

$$\begin{aligned}
m_{it} &= E((y_{it} - (\frac{x'_{i(t+1)}\beta}{1-\rho} + \frac{\eta_g}{1-\rho} + \frac{\tilde{\eta}_i}{1-\rho}))\tilde{\eta}_i) \\
&= E((y_{it} - (\frac{x'_{it}\beta}{1-\rho} + \frac{\eta_g}{1-\rho} + \frac{\tilde{\eta}_i}{1-\rho}))\tilde{\eta}_i) \\
&= E((\rho y_{i(t-1)} - \frac{\rho}{1-\rho}x'_{it}\beta - \frac{\rho}{1-\rho}\eta_g - \frac{\rho}{1-\rho}\tilde{\eta}_i)\tilde{\eta}_i) \\
&= \rho E((y_{i(t-1)} - (\frac{x'_{it}\beta}{1-\rho} + \frac{\eta_g}{1-\rho} - \frac{\tilde{\eta}_i}{1-\rho}))\tilde{\eta}_i) \\
&= \rho m_{i(t-1)}.
\end{aligned}$$

This means that if the system has been generating numbers, such that (3.23) holds once, it also holds for all subsequent y_{it}. The initial condition for an individual is for example fulfilled if it has already achieved its long run steady state and is only fluctuating around it with respect to v_{it}. If the country is in its transition phase to its steady state, then the difference to its long-run steady state can be uncorrelated to the individual error but this is not necessarily the case. However, if $\rho < 1$ the correlations of the differences to the steady state to the individual errors decrease with speed determined by ρ. The System GMM offers the most help if ρ is close one and in which case the system is least likely to have achieved the initial condition when the observation time begins. Therefore, when the System GMM becomes especially necessary it is least likely to fulfill the underlying assumptions that allow to apply it. The Hansen J-Test does not offer help to test the validity of the moment conditions because of the large number of instruments. As the series of y is highly persistent we conclude that it is very unlikely that the initial condition holds. Roodman (2009) provides methods to reduce the instrument count. Limiting the lag-depth to one gives an instrument count which is still far too large. Another method to reduce the instrument count is collapsing. Suppose we do not assume that

$$E(Z'_i D u_{it}) = E((w'_{i2}\Delta u_{i3}, w'_{i3}\Delta u_{i4} w'_{i2}\Delta u_{i4}, \ldots, w'_{i(T-1)}\Delta u_{iT}, \ldots, w'_{i2}\Delta u_{iT})') = 0 \in \mathbb{R}^{r_{Diff}},$$

but only assume that

$$E(Z'_i D u_{it}) = E((\sum_{t=3}^{T} w'_{i(t-1)}\Delta u_{it}, \sum_{t=4}^{T} w'_{i(t-2)}\Delta u_{it}, \ldots, \sum_{t=T}^{T} w'_{i(t-(T-2))}\Delta u_{it})) = 0 \in \mathbb{R}^{(T-2)(K+1)}.$$

This means that we sum up the instruments time-wise. In the same way we can collapse the additional instruments for the System GMM estimator. Instead of assuming that $E((\Delta w_{i3}u_{i3}, \Delta w_{i4}u_{i4}, \ldots, \Delta w_{iT}u_{iT})) = 0$, we assume that $E(\sum_{t=3}^{T} \Delta w_{it}u_{it}) = 0$. The instrument count is still far too large. The only way is to collapse and to reduce the lag-depth. If we reduce the lag depth to two and collapse, we have 13 instruments. Note that reducing the number of instruments makes it possible to apply the System GMM but has large drawbacks in terms of efficiency.

Another problem of the Sytem GMM is that it needs the strict assumption that the residuals are not correlated. When this assumption is slightly violated, Least-Squares estimators are robust as they are still consistent but the System GMM suffers from a bias of unknown magnitude.
We conclude that there is a dramatic loss of efficiency due to reducing the instrument count. Additionally there is a dramatic loss of efficiency that results from the two-step method. Therefore, the significance of the Africa-Dummy is hard to determine. Furthermore, the correlations between the coefficients cannot be estimated with the two-step method. Moreover, as it is not clear to what extend the residuals are correlated, the System GMM suffers from a bias of unknown magnitude. Therefore, we do not use System GMM.

3.3.3 The Hausman-Taylor Estimator

We estimate the coefficients of equation (3.14)

$$\begin{aligned} y &= \iota_{n(T-1)}\eta_g + \rho y_{-1} + X\beta + \iota_{n(T-1),SSH} * SSH + C\tilde{\eta} + \nu \\ &= W(\eta_g, SSH, \rho, \beta')' + u, \end{aligned}$$

where $W = (\iota_{n(T-1)}, \iota_{n(T-1),SSH}, y_{-1}, X) \in \mathbb{R}^{n(T-1)\times(K+3)}$ and $u = C\tilde{\eta} + \nu \in \mathbb{R}^{T-1}$. We assume for the country-specific errors, that they are independent and that their common variance is

$$Var(\tilde{\eta}_i) = \sigma_\eta^2. \tag{3.27}$$

The Random Effects model disregards the potential correlation of η_i to the exogenous regressors. The simplest approach to estimate this model is to pool all data and then apply OLS. The pooled estimator provides consistent estimates. As the errors u_{it} are correlated, a robust choice to estimate the coefficients yields more efficient estimates. The covariance matrix of the vector $u_i = (u_{i2}, \ldots, u_{iT})$ is

$$\Sigma_{u_i} = \begin{pmatrix} \sigma_\eta^2 + \sigma_\nu^2 & \sigma_\eta^2 & \cdots & \sigma_\eta^2 \\ \vdots & \ddots & & \vdots \\ \sigma_\eta^2 & \cdots & \sigma_\eta^2 & \sigma_\eta^2 + \sigma_\nu^2 \end{pmatrix} \in \mathbb{R}^{(T-1)\times(T-1)}.$$

Therefore, the covariance matrix of the vector $u = (u_1', \ldots, u_n')' \in \mathbb{R}^{n(T-1)}$ is

$$\Sigma = \begin{pmatrix} \Sigma_{u_1} & & \\ & \ddots & \\ & & \Sigma_{u_n} \end{pmatrix} \in \mathbb{R}^{n(T-1)\times n(T-1)}.$$

Applying GLS yields an unfeasible estimator of $(\eta_g, SSH, \rho, \beta')'$, namely

$$(W'\Sigma^{-1}W)^{-1}W'\Sigma^{-1}y.$$

The solution to this is the same as regressing the quasi-demeaned y on the quasi-demeaned columns of W. If vector z is

$$z = (z_{12}, \ldots, z_{1T}, z_{22}, \ldots, z_{2T}, \ldots, z_{n2}, \ldots, z_{nT})' \in \mathbb{R}^{n(T-1)},$$

then the quasi-demeaned z is

$$\tilde{z}_{QD} = (z_{12} - \theta\bar{z}_{1\bullet}, \ldots, z_{1T} - \theta\bar{z}_{1\bullet}, z_{22} - \theta\bar{z}_{2\bullet}, \ldots, z_{2T} - \theta\bar{z}_{2\bullet}, \ldots, z_{n2} - \theta\bar{z}_{n\bullet}, \ldots, z_{nT} - \theta\bar{z}_{n\bullet})' \in \mathbb{R}^{n(T-1)},$$

where

$$\bar{z}_{i\bullet} = \frac{1}{T-1}\sum_{t=2}^{T} z_{it}$$

and

$$\theta = 1 - \sqrt{\sigma_v^2/((T-1)\sigma_\eta^2 + \sigma_v^2)}.$$

To obtain a feasible version, we estimate the variances of the error components. The pooled OLS estimator gives consistent estimates for the residuals u_{it} which we denote by $\hat{u}_{it}$ and a consistent estimator of its variance which we denote by $\hat{\sigma}_u^2$. Consistent estimators for the variances of the error components and θ are given by

$$\begin{aligned} \hat{\sigma}_{\tilde{\eta}}^2 &= \frac{1}{n(T-1)(T-2)/2 - (K+3)} \sum_{i=1}^{n}\sum_{t=1}^{T-2}\sum_{s=t+1}^{T-1} \hat{\tilde{u}}_{it}\hat{\tilde{u}}_{is}, \\ \hat{\sigma}_v^2 &= \hat{\sigma}_u^2 - \hat{\sigma}_{\tilde{\eta}}^2, \\ \hat{\theta} &= 1 - \sqrt{\hat{\sigma}_v^2/((T-1)\hat{\sigma}_\eta^2 + \hat{\sigma}_v^2)}. \end{aligned} \tag{3.28}$$

We use $\hat{\theta}$ to obtain the quasi-demeaned W and y, namely $\tilde{W}_{QD}$ and $\tilde{y}_{QD}$. The Random Effects estimator is

$$(\hat{\eta}_{g_{RE}}, \hat{SSH}_{RE}, \hat{\rho}_{RE}, \hat{\beta}_{RE}')' = (\tilde{W}_{QD}'\tilde{W}_{QD})^{-1}\tilde{W}_{QD}'\tilde{y}_{QD}. \tag{3.29}$$

As the individual effects of (3.14) are correlated to the regressors, the Random Effects estimator suffers from an endogeneity bias. Hausman and Taylor (1981) present an Instrumental Variable estimator for estimating the coefficients of (3.14). Since the Africa-Dummy already rules out systematic differences of the group of the sub-Saharan African countries, we assume

$$E(1_{SSH,i} * \tilde{\eta}_i) = 0. \tag{3.30}$$

When demeaning the regression equation, all individual variables disappear. Furthermore, we disregard the endogeneity bias induced by the lagged variable (see subsection (3.3.1)) and use (3.30). Then

$$Z = (\iota_{n(T-1)}, \iota_{n(T-1),SSH}, (I_{n(T-1)} - I_n \otimes \iota\iota')y_{-1}, (I_{n(T-1)} - I_n \otimes \iota\iota')X) \in \mathbb{R}^{n(T-1)\times(K+3)}$$

is a matrix whose columns provide instruments. Hausman and Taylor (1981) propose to use some of the explanatory variables as additional instruments. In our case we do not find

a reason for that one of the explanatory variables is uncorrelated to the individual effects. Applying the Instrumental Variable estimation method yields an estimator for $(\eta_g, SSH, \rho, \beta')'$, namely $(Z'W)^{-1}Z'y$. The solution to this is that β and ρ are estimated by the Within Group estimator and η_g and SSH are estimated by $\bar{\eta}_{NA}$ and $\bar{\eta}_A - \bar{\eta}_{NA}$ respectively, where we denote the average residual of country j by $\bar{\eta}_j = y_{j\bullet} - \hat{\rho}_{WG} y_{-1j\bullet} - x'_{i\bullet}\hat{\beta}_{WG}$, the average residual of all non sub-Saharan African countries by $\bar{\eta}_{NA} = \frac{1}{n-s}\sum_{j=s+1}^{n} \bar{\eta}_j$ and the average residual of all sub-Saharan African countries by $\bar{\eta}_A = \frac{1}{s}\sum_{j=1}^{s} \bar{\eta}_j$. Since we have error components, we apply 2SLS using (3.28). This is the Hausman-Taylor estimator.
If conditional on the regressors, individual effects can be viewed as random draws from a common population, we estimate with error components. One motivation for doing this could be that the common population characteristics are of interest. In growth regression, it is very unlikely that there is a common population. The effects of different countries are highly heterogeneous. Furthermore, the performance of individual countries is of interest. The disadvantage of Random Effects estimators is that it does not take this heterogeneity of the fixed effects into account and it is not possible to examine the performance of individual countries.

3.3.4 The Two-Groups Least-Square Dummy-Variable Estimator

The Least-Square Dummy-Variable estimator is the OLS estimator of ρ, β and of each η_i in equation (3.12)

$$\hat{\rho}_{LSDV} = \hat{\rho}_{WG},\ \hat{\beta}_{LSDV} = \hat{\beta}_{WG} \text{ and } \hat{\eta}_{LSDV,i} = \bar{\eta}_i \text{ for } i = 1, \ldots, n. \tag{3.31}$$

Since $(C'C)^{-1} = \frac{1}{T-1} I_{n(T-1)}$ the model can be identified. Equation (3.14) has $n+2$ country-specific regressors (an intercept, n country-specific errors and an Africa-Dummy). When stacking this equation and considering the country-specific regressor matrix, it has $n+2$ columns and $n(T-1)$ rows from which only n rows are different to each other. Therefore, the country-specific regressor matrix has rank n at the highest and the model cannot be identified. Therefore, applying the Least-Square Dummy-Variable estimator yields in applying a two-step regression, which has efficiency problems.
To be able to estimate (3.14) directly, we assume that the errors of the sub-Saharan African countries sum up to zero and that the errors of the non-sub-Saharan African countries sum up to zero separately

$$\sum_{i=1}^{s} \tilde{\eta}_i = 0 \text{ and } \sum_{i=s+1}^{n} \tilde{\eta}_i = 0. \tag{3.32}$$

This assumption specifies two errors precisely

$$\tilde{\eta}_s = -\tilde{\eta}_1 - \tilde{\eta}_2 - \ldots - \tilde{\eta}_{s-1} \text{ and } \tilde{\eta}_n = -\tilde{\eta}_{s+1} - \tilde{\eta}_{s+2} - \ldots - \tilde{\eta}_{n-1}.$$

Plugging (3.32) into (3.14) yields

$$y = \rho y_{-1} + X\beta + C_{SSH}\eta_{SSH} + \nu \in \mathbb{R}^{n(T-1)}, \tag{3.33}$$

with

$$\eta_{SSH} = (\eta_g, SSH, \tilde{\eta}_1, \ldots, \tilde{\eta}_{s-1}, \tilde{\eta}_{s+1}, \ldots, \tilde{\eta}_{n-1})' \in \mathbb{R}^n$$

and

$$C_{SSH} = \left(\begin{array}{cc|ccc|ccc} \iota & \iota & \iota & & & & & \\ \vdots & \vdots & & \ddots & & & & \\ \iota & \iota & & & \iota & & & \\ \iota & \iota & -\iota & \cdots & -\iota & & & \\ \hline \iota & & & & & \iota & & \\ \vdots & & & & & & \ddots & \\ \iota & & & & & & & \iota \\ \iota & & & & & -\iota & \cdots & -\iota \end{array}\right) \in \mathbb{R}^{n(T-1)\times n},$$

where the lower right box refers to the non-sub-Saharan African countries and has $n-s-1$ columns and $(n-s)(T-1)$ rows and the upper middle box refers to the sub-Saharan African countries and has $s-1$ columns and $s(T-1)$ rows. It is easy to check that

$$C'_{SSH}C_{SSH} = (T-1)\begin{pmatrix} Z_1 & & \\ & Z_2 & \\ & & Z_3 \end{pmatrix} \in \mathbb{R}^{n\times n},$$

with

$$Z_1 = \begin{pmatrix} n & s \\ s & s \end{pmatrix} \in \mathbb{R}^{2\times 2},$$

$$Z_2 = \begin{pmatrix} 2 & 1 & \ldots & 1 \\ 1 & 2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \ldots & 1 & 2 \end{pmatrix} \in \mathbb{R}^{(s-1)\times(s-1)},$$

and

$$Z_3 = \begin{pmatrix} 2 & 1 & \ldots & 1 \\ 1 & 2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \ldots & 1 & 2 \end{pmatrix} \in \mathbb{R}^{(n-s-1)\times(n-s-1)}.$$

The inverses of Z_1, Z_2 and Z_3 exist and are given by

$$Z_1^{-1} = \frac{1}{n-s}\begin{pmatrix} 1 & -1 \\ -1 & n/s \end{pmatrix} \in \mathbb{R}^{2\times 2},$$

$$Z_2^{-1} = \frac{1}{s}\begin{pmatrix} (s-1) & -1 & \dots & -1 \\ -1 & (s-1) & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ -1 & \dots & -1 & (s-1) \end{pmatrix} \in \mathbb{R}^{(s-1)\times(s-1)},$$

and

$$Z_3^{-1} = \frac{1}{n-s}\begin{pmatrix} (n-s-1) & -1 & \dots & -1 \\ -1 & (n-s-1) & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ -1 & \dots & -1 & (n-s-1) \end{pmatrix} \in \mathbb{R}^{(n-s-1)\times(n-s-1)}.$$

Therefore,

$$(C'_{SSH}C_{SSH})^{-1} = \frac{1}{T-1}\begin{pmatrix} Z_1^{-1} & & \\ & Z_2^{-1} & \\ & & Z_3^{-1} \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Note that the existence of $(C'_{SSH}C_{SSH})^{-1}$ is equivalent to that the columns of C_{SSH} are linear independent, meaning that the model can be identified. It is now easy to check that

$$M_{C_{SSH}} = I_{n(T-1)} - C_{SSH}(C'_{SSH}C_{SSH})^{-1}C'_{SSH} = I_{n(T-1)} - I_n \otimes \iota\iota' \in \mathbb{R}^{n(T-1)\times n(T-1)}.$$

Therefore, ρ and β are estimated by the Within Group estimator. Furthermore,

$$\hat{\eta}_{SSH} = (C'_{SSH}C_{SSH})^{-1}C'_{SSH}(y - \hat{\rho}_{WG}y_{-1} - X\hat{\beta}_{WG}).$$

Solving this gives the Two-Groups Least-Square Dummy-Variable estimator

$$\begin{aligned} &\hat{\rho} = \hat{\rho}_{WG},\ \hat{\beta} = \hat{\beta}_{WG},\ \hat{\eta}_g = \bar{\eta}_{NA},\ \hat{SH} = \bar{\eta}_A - \bar{\eta}_{NA}, \\ &\hat{\tilde{\eta}}_j = \bar{\eta}_j - \bar{\eta}_A \text{ for } j \in \{1,\dots,s-1\} \text{ and } \hat{\tilde{\eta}}_j = \bar{\eta}_j - \bar{\eta}_{NA} \text{ for } j \in \{s+1,\dots,n-1\}. \end{aligned} \tag{3.34}$$

With (3.34) and $-\tilde{\eta}_1 - \dots - \tilde{\eta}_{s-1} = \tilde{\eta}_s$ we have $\hat{\tilde{\eta}}_s = \bar{\eta}_s - \bar{\eta}_A$ and in the same manner $\hat{\tilde{\eta}}_n = \bar{\eta}_n - \bar{\eta}_{NA}$. The total country-specific effect of a sub-Saharan African country with index $j \in \{1,\dots,s\}$ is $\hat{\eta}_g + \hat{SH} + \hat{\tilde{\eta}}_j = \bar{\eta}_j$ and that of a non-sub-Saharan African country with index $j \in \{s+1,\dots,n\}$ is $\hat{\eta}_g + \hat{\tilde{\eta}}_j = \bar{\eta}_j$. Note that these are the country-specific effects of the Least-Square Dummy-Variable estimator.

The advantage of the Two-Groups Least-Square Dummy-Variable estimator compared with the Hausman-Taylor estimator is that it does not need the assumption of a common population. Therefore, the effects of different countries are heterogeneous. Furthermore, it allows to examine the performance of individual countries. The formulas of the Hausman-Taylor estimator for estimating the intercept, the Africa-Dummy, ρ and β are exactly the same as those of the Two-Groups Least-Square Dummy-Variable estimator but the estimators for second moments are not. The Two-Groups Least-Square Dummy-Variable estimator allows to reliably estimate the correlations of the

Africa-Dummy to other regressors. Furthermore, as it does not use the inefficient Instrumental Variable method, it is more efficient. Another example of the Least-Squares method is that it remains being consistent even if the residuals are heteroscedastic and slightly correlated.

3.3.5 Results

Tables (3.3) and (3.4) show the estimated coefficients and the standard errors. The interpretation of the five year lagged model is similar to that of the one year lagged model for all estimation methods. The coefficient of *lnn* is almost zero in the one year lagged model and at least becomes negative significant on ten percent level in the five year lagged model. It is surprising to see that the coefficient of *lnattain* is clearly negative. Figure (3.6) shows this negative correlation when multiplying the dependent variable and *lnattain* by the projection matrix that projects each vector on the orthogonal column space of that spanned by all other explanatory variables. It can clearly be seen that the negative coefficient is not a result of a misspecification of the functional structure or of influential observations. The negative coefficient was for example also identified by Islam (1995). He argues that the observed effect of human capital is either a measurement problem or relates to a misspecification of this variable by the Augmented Solow model. The indicator by Barro and Lee (2010) does not take the quality of schooling into account. It can be observed that the school attainment according to Barro and Lee (2010) incrementally increases for almost all countries but the growth rate does not. The result is a negative coefficient. Pritchett (1996) argues that this result is robust, credible and provides three possible explanations. First he argues that schooling does not necessarily create human capital, second, the returns to education fall rapidly when the demand for educated labor is stagnant and third, a large amount of human capital is used for growth hindering activities, such as a bloated bureaucracy.

Table (3.3) shows the estimated coefficients of the error components models. Random Effects suffers from an endogeneity bias and its results are slightly different than Hausman-Taylor. Table (3.4) shows the estimated coefficients of the fixed effects models. Least-Square Dummy-Variable and Two-Groups Least-Square Dummy-Variable give similar results for the time- and country-varying coefficients but Least-Square Dummy-Variable estimates a larger intercept with smaller standard errors and an equal Africa-Dummy with much larger standard errors. Hausman-Taylor has larger standard errors than Two-Groups Least-Square Dummy-Variable. The advantage of Two-Groups Least-Square Dummy-Variable can also be seen when considering correlations. Table (3.5) shows the correlations of the estimated coefficients and the estimated Africa-Dummy. Least-Square Dummy-Variable does not estimate correlations at all. Random Effects and Hausman-Taylor give similar results because they are both based on the idea of error

components. Two-Groups Least-Square Dummy-Variable gives very different results because it is based on the idea of including fixed effects as regressors. It does not need the rather strict group-wise homogeneity assumption which is why we identify it as the best estimator to calculate the correlations.
According to Two-Groups Least-Square Dummy-Variable the coefficient of the Africa-Dummy is larger, the smaller the coefficient of *lnn* and *lnattain* and the larger the coefficient of *lnsk*. Nevertheless, its correlations to the coefficient of *lnattain* and *lnn* are small. In other words, if the return to investment in physical capital increases, the punishment of belonging to sub-Saharan Africa decreases. Furthermore, if the return to the depreciation rate or the school attainment increases, the punishment of belonging to sub-Saharan Africa increases slightly.
We analyze the fixed effects estimated by the Two-Groups Least-Square Dummy-Variable estimator. The total fixed effects are $\bar{\eta}_i$. The Two-Groups Least-Square Dummy-Variable estimator is able to estimate the decomposition $\tilde{\eta}_i + \eta_g + SSH * 1_{SSH;i}$. We denote $\tilde{\eta}_i + \eta_g + SSH * 1_{SSH;i}$ by fixed effects and $\tilde{\eta}_i + \eta_g$ by corrected fixed effects. The corrected fixed effects are larger than the fixed effects in case of a sub-Saharan African country and equal for all other countries. Figure (3.7) shows boxplots of the fixed effects in the one year lagged case. We observe that the distribution of the fixed effects is slightly skewed to the left. In the one year lagged model it can be seen that adding the Africa-Dummy as a regressor results in a more symmetric distribution of the remaining parts of the fixed effects. The two outliers of the one year lagged model correspond to the sub-Saharan African country Niger and the Latin American country Nicaragua. Even though Niger is

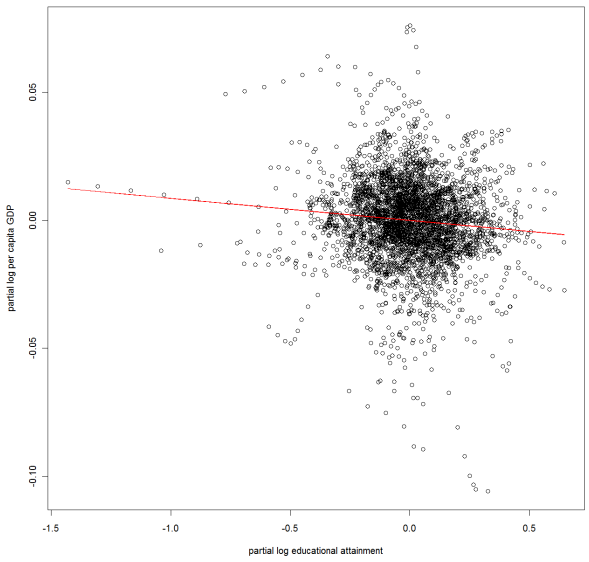

Figure 3.6: *The negative coefficient of lnattain in the growth regression.*

affected by the correction, it remains being an outlier when considering the corrected country-specific errors. Figure (3.8) shows that, when looking at the corrected fixed effects and the fixed effects of the five year lagged model, the skewness is not completely removed. The tails of the corrected fixed effect support a symmetric distribution but as the median is closer to the first quartile than to the third quartile, the distribution is slightly skewed to the left. Nevertheless, the corrected fixed effects of the five year lagged model are slightly skewed to the left. When looking at the residuals, we observe a similar

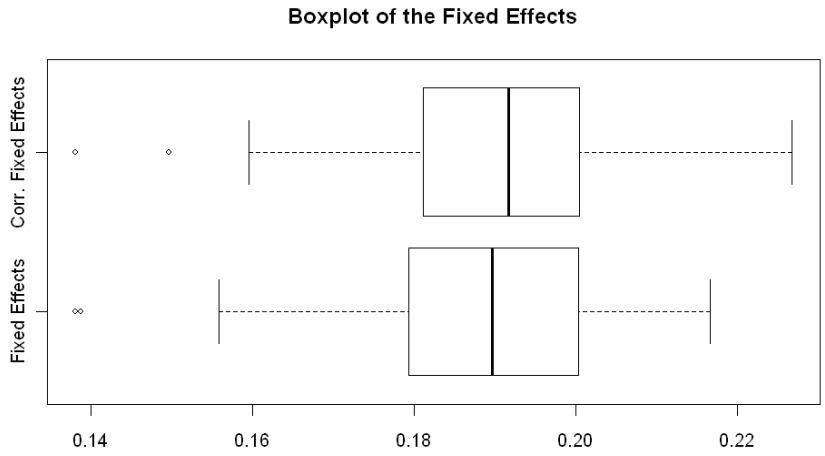

Figure 3.7: *Boxplot of the fixed effects for the one year lagged model.*

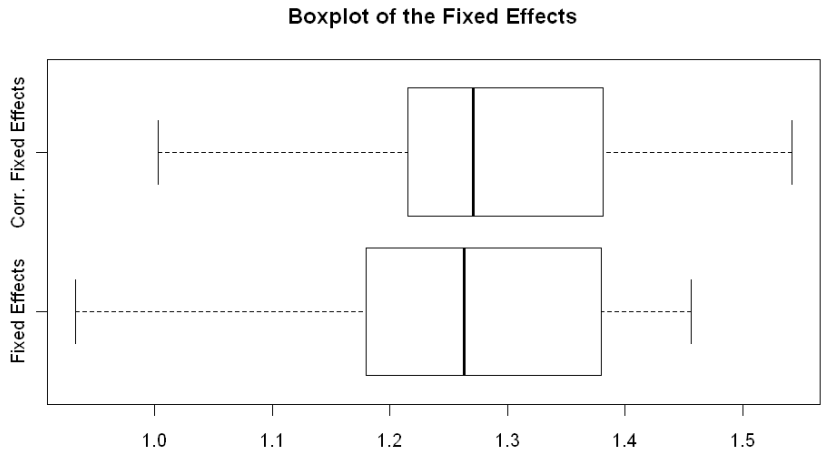

Figure 3.8: *Boxplot of the fixed effects for the five year lagged model.*

Table 3.3: *Random Effects Estimators*

	RE	RE (5)	HT	HT (5)
Intercept	0.1771***	1.2100***	0.1905***	1.2894***
	(0.0112)	(0.0633)	(0.0118)	(0.0649)
lag y	0.9898***	0.9000***	0.9897***	0.8926***
	(0.0011)	(0.0059)	(0.0011)	(0.0061)
lnn	-0.0002	-0.0282*	0.0008	-0.0240
	(0.0025)	(0.0126)	(0.0025)	(0.0127)
lnsk	0.0277***	0.0837***	0.0275***	0.0813***
	(0.0012)	(0.0062)	(0.0012)	(0.0063)
lnattain	-0.0148***	-0.0496***	-0.0150***	-0.0493***
	(0.0010)	(0.0052)	(0.0010)	(0.0053)
SSH	-0.0090	-0.1428***	-0.0109*	-0.1551***
	(0.0049)	(0.0353)	(0.0046)	(0.0301)
* $p: \leq 0.05$	**≤ 0.01	***≤ 0.001		

behavior for the one year lagged model and the five year lagged model. Its distribution is extremely heavy tailed and slightly skewed to the left. This indicates that more regressors than those given by Mankiw, Romer and Weil (1992) contribute to explaining growth. However, when adding a regressor to the growth model it is not clear whether it drives growth or is only somehow correlated to what cannot be explained by the model without that regressor.

3.4 More about the Africa-Dummy

3.4.1 Semiparametric Modeling

The growth model by Mankiw, Romer and Weil (1992) suggests the regression equation (3.14) which has a linear functional structure. We investigate if a misspecification of this functional structure is responsible for that the Africa-Dummy is negative and significant. Figure (3.9) shows a simplified example of what could happen to the Africa-Dummy when we relax the functional structure. Suppose we have only two individuals that follow two different linear regression equations with the same slope but different intercepts. In figure (3.9) we have a regressor that is uniformly distributed. The small input points between zero and a third (the red points) have the small output $y = 1 + 0.7x + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, 0.35)$. The large input points between a third and one (the blue points) have the large output $y = 1.5 + 0.7x + \varepsilon$. In figure (3.9) the red line belongs to the equation $y = 1 + 0.7x$ and the blue line belongs to the equation $y = 1.5 + 0.7x$. The green dashed line belongs to the nonlinear equation $y = 2.05 + 0.39 * \log(x)$. When only looking at figure (3.9) it is not

Table 3.4: *Fixed Effects Estimators*

	LSDV	LSDV (5)	2G LSDV	2G LSDV (5)
Intercept	0.1905***	1.2894***	0.1795***	1.1343***
	(0.0020)	(0.0134)	(0.0117)	(0.0635)
lag y	0.9897***	0.8926***	0.9897***	0.8926***
	(0.0011)	(0.0061)	(0.0011)	(0.0061)
lnn	0.0008	-0.0240	0.0008	-0.0240
	(0.0025)	(0.0127)	(0.0025)	(0.0127)
lnsk	0.0275***	0.0813***	0.0275***	0.0813***
	(0.0012)	(0.0063)	(0.0012)	(0.0063)
lnattain	-0.0150***	-0.0493***	-0.0150***	-0.0493***
	(0.0010)	(0.0053)	(0.0010)	(0.0053)
SSH	-0.0109*	-0.1551***	-0.0109***	-0.1551***
	(0.0044)	(0.0293)	(0.0017)	(0.0090)

* $p: \leq 0.05$ **≤ 0.01 ***≤ 0.001

Table 3.5: *Correlations*

	Corr lnn	Corr lnsk	Corr lnattain	Method
RE	-0.5587	-0.0129	-0.0757	Direct
RE(5)	-0.5588	-0.0487	-0.0266	Direct
HT	-0.5605	-0.0112	-0.0753	Direct
HT(5)	-0.5589	-0.0481	-0.0240	Direct
LSDV	.	.	.	Two Step
LSDV(5)	.	.	.	Two Step
2G LSDV	-0.1170	0.5641	-0.0938	Direct
2G LSDV(5)	-0.1279	0.5252	-0.0537	Direct

clear whether the data come from the two different linear processes with the same slope or from one and the same nonlinear process. This example shows that different linear models could be understood as one nonlinear model. This motivates to investigate if a significant difference in the intercepts of the growth models of the group of sub-Saharan African countries and that of all other countries disappear when relaxing the functional structure of regression equation (3.14). In other words, the task is to investigate if the input variables given in the model by Mankiw, Romer and Weil (1992) suffice in explaining sub-Saharan Africa's growth tragedy when simply relaxing the functional structure.

We use B-Splines of degree three with equidistant knots to relax the functional structure of the variables *lnn*, *lnsk* and *lnattain*. The number of knots have to be chosen in a reasonable way that takes the sample size as well as the number of regressors into account. Akaike's Information Criterion results in choosing models with too many parameters when having large samples. The Bayesian Information Criterion punishes harder for choosing a lot of

explanatory variables. Therefore, we chose the number of knots with respect to that it minimizes the Bayesian Information Criterion. More precisely, we vary the number of knots between three and ten and choose the combination that minimizes the Bayesian Information Criterion. The result for the one year lagged model is zero knots for the variables *lnn* and *lnattain* and one knot for the variable *lnsk*. The result for the five year lagged model is one knot for all variables. When running these regressions we observe that the coefficient of the lagged dependent variable increases from 0.9897 to 0.9920 in the one year lagged model and decreases from 0.8926 to 0.8911 in the five year lagged model. The intercept decreases from 0.1905 to 0.0322 in the one year lagged model and from 1.2894 to 0.8834 in the five year lagged model. The magnitude of the Africa-Dummy increases slightly from -0.0109 to -0.0113 in the one year lagged model and from -0.1551 to -0.1582 in the five year lagged model. However, in the one year lagged and five year lagged case we observe a highly significant Africa-Dummy. We conclude that the significance of the Africa-Dummy cannot be explained by a misspecification of the functional structure.

3.4.2 Interaction Effects

In this subsection we discuss how the beta coefficients of (3.14) differ for sub-Saharan African countries. We consider model (3.14) with interaction effects. Interaction effects also allow for time varying punishments of sub-Saharan African countries. The results are given in table (3.6).
First of all, we observe a positive significant interaction effect of the coefficient of *lnn*. This means that it needs to be corrected for sub-Saharan African countries such that the

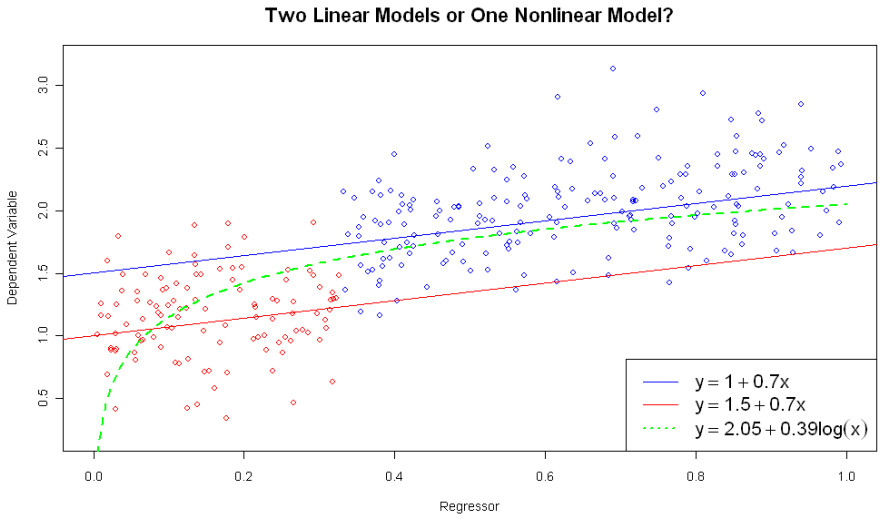

Figure 3.9: *Interpolation of schooling*

resulting coefficient is positive. For the one year lagged model the total coefficient of *lnn* is $-0.0129+0.0357=0.0228$ and for the five year lagged model $-0.0760+0.1535=0.0775$. This is counterintuitive. Figure (3.10) shows boxplots for the time-series of lnn_{it} for the sub-Saharan African countries and other countries. We observe that sub-Saharan African countries have a larger depreciation rate because of the larger rate of population growth. Furthermore, the Inter Quartile Range is smaller with more outliers. The positive coefficient of *lnn* of sub-Saharan African countries shows that the larger population growth is advantageous for the growth of sub-Saharan African countries. The difference of the total coefficient of *lnn* for sub-Saharan African countries and the coefficient for all other countries overemphasizes sub-Saharan Africa's punishment. The Africa-Dummy is positive. A low population growth rate means that there is a low birth rate or people die. For example conflicts or diseases cause high death rates but both reduce the GDP as for example war costs money or diseases cause people not to work.
Furthermore, it can be seen from the interaction effect of the estimated coefficient of the five year lagged model that the time-series of GDP per worker entails less autocorrelation than that of the other countries. This also means that less variation is explained by the GDP per worker time-series itself and indicates that other explanatory variables, such as for example those given by the model of Mankiw, Romer and Weil (1992), contribute to growth.
Moreover, in the one year lagged model, the interaction effect of *lnattain* is small and positive but significant. However, the resulting coefficient is still negative and of large magnitude.

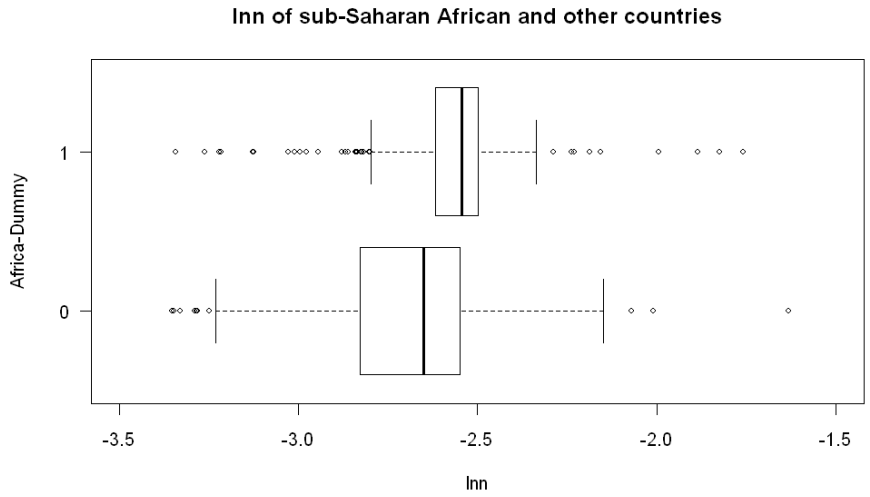

Figure 3.10: *lnn_{it} stratified by sub-Saharan African and other countries*

Table 3.6: *Estimating the coefficients of the growth regression with interaction effects*

	one year Estimate (S.E.)	five year Estimate (S.E.)
Intercept	0.1588***	1.0938***
	(0.0134)	(0.0724)
SSH	0.0646*	0.6151***
	(0.0266)	(0.1451)
lag y	0.9895***	0.8976***
	(0.0013)	(0.0070)
Int. lag y	0.0020	-0.0397**
	(0.0027)	(0.0147)
lnn	-0.0129***	-0.0760***
	(0.0031)	(0.0159)
Int. lnn	0.0357***	0.1535***
	(0.0052)	(0.0265)
lnsk	0.0268***	0.0752***
	(0.0016)	(0.0081)
Int. lnsk	0.0028	0.0145
	(0.0025)	(0.0129)
lnattain	-0.0175***	-0.0498***
	(0.0013)	(0.0070)
Int. lnattain	0.0047*	0.0017
	(0.0020)	(0.0108)
$*p: \leq 0.05$	$**\leq 0.01$	$***\leq 0.001$

3.4.3 The Development of the Africa-Dummy

In this subsection we investigate how the Africa-Dummy evolves over time. Consider the model

$$y_{it} = \eta_g + \rho y_{i(t-1)} + x_{it}'\beta + \sum_{s=2}^{T} SSH_s * d_{SSH,t}(i,s) + \tilde{\eta}_i + \nu_{it}, \tag{3.35}$$

with $t = 2,\ldots,T$ and $i = 1,\ldots,n$, where $d_{SSH,t}(i,s) = 1$ if country i belongs to sub-Saharan Africa and $s = t$ and $d_{SSH,t}(i,s) = 0$ else. We assume that this model has the same statistical properties concerning the error structure and the fixed effects as (3.14). This includes $\sum_{i=1}^{s} \tilde{\eta}_i = 0$ and $\sum_{i=s+1}^{n} \tilde{\eta}_i = 0$ to be able to identify the model. Stacking first time-series and then cross-sectional data yields

$$y = \rho y_{-1} + X\beta + (\iota_{SSH} \otimes I_{T-1})SSH + C\eta + \nu \in \mathbb{R}^{n(T-1)},$$

Table 3.7: *Coefficients with a time-varying Africa-Dummy*

	one year Estimate (S.E.)	five year Estimate (S.E.)
Intercept	0.1832***	1.2654***
	(0.0117)	(0.0636)
lag y	0.9911***	0.8964***
	(0.0011)	(0.0062)
lnn	0.0012	-0.0214
	(0.0025)	(0.0128)
lnsk	0.0277***	0.0834***
	(0.0013)	(0.0065)
lnattain	-0.0175***	-0.0510***
	(0.0012)	(0.0065)
* $p: \leq 0.05$	**≤ 0.01	***≤ 0.001

where $SSH = (SSH_2, \ldots, SSH_T)' \in \mathbb{R}^{T-1}$, $\eta = (\eta_g, \tilde{\eta}_1, \ldots, \tilde{\eta}_{s-1}, \tilde{\eta}_{s+1}, \ldots, \tilde{\eta}_{n-1})' \in \mathbb{R}^{n-1}$ and

$$C = \left(\begin{array}{c|ccc|ccc} \iota & \iota & & & & & \\ \vdots & & \ddots & & & & \\ \iota & & & \iota & & & \\ \iota & -\iota & \cdots & -\iota & & & \\ \hline \iota & & & & \iota & & \\ \vdots & & & & & \ddots & \\ \iota & & & & & & \iota \\ \iota & & & & -\iota & \cdots & -\iota \end{array}\right) \in \mathbb{R}^{n(T-1)\times(n-1)}.$$

Note that this matrix does not contain the time varying Africa-Dummies. The lower right box refers to the non sub-Saharan African countries and has $n-s-1$ columns and $(n-s)(T-1)$ rows, the upper middle box refers to the sub-Saharan African countries and has $s-1$ columns and $s(T-1)$ rows and the first column refers to the intercept. The complete dummy matrix with the Africa-Dummies is $(\iota_{SSH} \otimes I_{T-1}, C) \in \mathbb{R}^{n(T-1)\times(n+(T-1))}$ and has full column rank. In the same way we formulate the five year lagged model

$$y_{it} = \eta_g + \rho y_{i(t-5)} + x'_{i(t-5)}\beta + \sum_{s=6}^{T} SSH_s * d_{SSH,t}(i,s) + \tilde{\eta}_i + \nu_{it}.$$

The results for the estimators of the coefficients are given in table (3.7). We observe that the estimators of the coefficients of (3.35) are similar to those of (3.14). Figures (3.11) and (3.12) show that the Africa-Dummy varies a lot over time. Apart from small bumps it incrementally decreases until the beginning to mid-nineties and then increases rapidly in

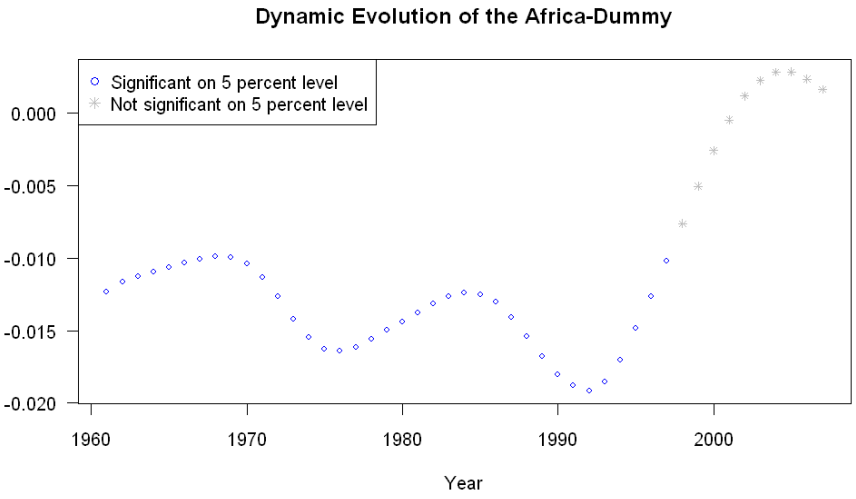

Figure 3.11: *The Evolution of the Africa-Dummy in the one year lagged model*

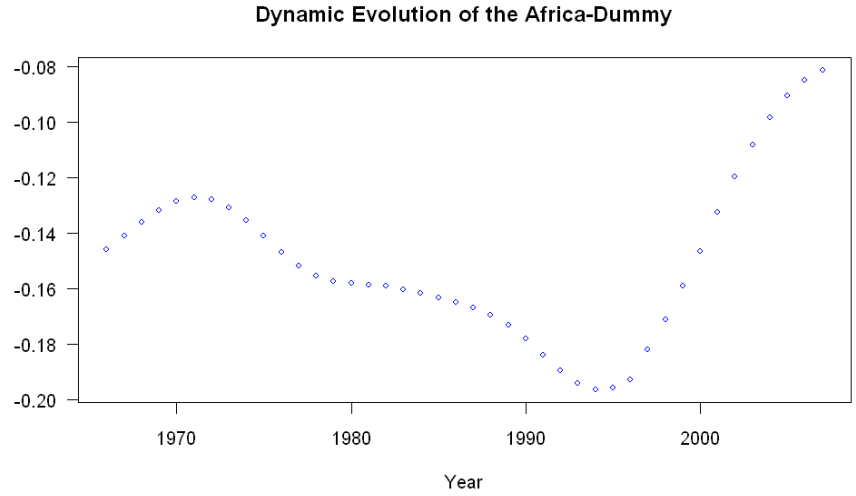

Figure 3.12: *The Evolution of the Africa-Dummy in the five year lagged model*

the recent years. When considering the one year lagged model it even becomes insignificant. Furthermore, in the one year lagged model, the two very recent Africa-Dummies are smaller than the ones before. It is not clear if this is related to a small bump or a dramatic increase of Africa's punishment. However, in the most recent years, Africa's punishment was of much smaller magnitude than before.

3.5 Conclusion

By smoothing with the Hodrick-Prescott filter, we obtain yearly time-series that represent the connection of one time-series of an economy to another. When doing this, the length of the time-series is sufficiently large, so that the endogeneity bias that results from the lagged dependent variable in growth regressions is negligibly small. Estimating the coefficients of the growth regression with the Two-Groups Least-Square Dummy-Variable estimator identifies a negative significant Africa-Dummy. This clear punishment for sub-Saharan African economies increases if the return to investment in physical capital decreases, if the return the depreciation rate increases, or if the return to school attainment increases.
The Two-Groups Least-Square Dummy-Variable estimator is also used to relax the functional structure of the growth regression equation. We observe that the significance of the Africa-Dummy does not disappear when applying a semiparametric model so that it cannot be explained by a misspecification of the functional structure.
We observe that sub-Saharan African countries have clearly positive returns to the depreciation rate. When adding interaction effects, the Africa-Dummy is even positive and significant.
Finally, an extension of the Two-Groups Least-Square Dummy-Variable estimator estimates the evolution of the Africa-Dummy within the period we observe data. It can clearly be seen that Africa-Dummy changes over time. Apart from small bumps it incrementally decreases until the beginning to mid-nineties and then increases rapidly in the recent years. When estimating exactly the regression equation that is motivated by the Augmented Solow Model, we even observe that is becomes insignificant in the recent years.

Chapter 4

A Variable-Coefficients Model for Assessing the Returns of Growth Regressions for the Poor And The Rich

Various papers demonstrate the importance of inequality, poverty and the size of the middle class for economic growth. When explaining why these measures of the income distribution are added to the growth regression, it is often mentioned that poor people behave different than rich people which translates to the economy as a whole. However, adding explanatory variables does not reflect this behavior. We formulate and apply a variable-coefficients model and show that the coefficients of the growth regressions differ a lot and this can be explained by the level of poverty, inequality and the middle class. Furthermore, we investigate how the coefficients and therefore the growth path differs for the poorer and for the richer part of the society. We argue that the differences in the coefficients impeach, on the one hand, the credibility of that the mean coefficients are informative, and, on the other hand, the credibility of the economic justification for explaining the growth path of a country with mean coefficients. Moreover, we explain that, when estimating mean coefficients, the estimation is likely to suffer from an endogeneity and from a sample selection bias.

4.1 Introduction

The literature shows that the variables inequality, poverty and the size of the middle class are important for economic growth. Usually the authors add the variables addressed in

their special question to the growth regression and observe the effects of these variables. When explaining why certain measures of the income distribution are added, it is often explained that poor people behave different than rich people and therefore, the economies as whole behave different according to their level of poverty, inequality and the size of their middle class. However, adding explanatory variables does not reflect this behavior. For example, it is hard to believe that a poor economy, in which a high number of poor people cannot provide collateral and therefore, do not have access to the credit market, has the same returns to investments in physical capital as a richer economy. This paper empirically investigates the effects of measures of the income distribution on the coefficients of the drivers for economic growth. More precisely, we consider three parameters of the income distribution, namely poverty, inequality and the size of the middle class and investigate the influence of each of these variables on the coefficients of the drivers for growth proposed by Mankiw, Romer and Weil (1992). For this purpose, we formulate and apply a variable-coefficients model. We focus on a panel data analysis using data from several countries over time.

In what follows, we report the literature concerning the relationships of growth and inequality, poverty and middle class. Afterwards, we explain, why these findings already motivate estimating growth regressions with variable coefficients.

Concerning inequality, there is a lot of literature stating that the level of inequality affects economic growth but there is no consensus about the size and direction of the effects. For example Bourguignon (2004) reports a literature review on the relationship of inequality and growth and finds that it is unclear whether inequality has positive or negative effects on growth. However, all studies state that in an economy that is subject to a lot of inequality, people behave different than in a more equal economy. There are four kinds of studies investigating the effects of inequality on growth. First, the inverted U-shaped relationship, second positive effects, third negative effects and fourth, the studies that assert that the relationship between inequality and growth depends on other parameters and is therefore not global.

First, there are studies arguing for the inverted U-shaped relationship. For example Kuznets (1955) asserts that the inequality of developing countries increases at the beginning of industrialization, before it converges to a lower level of inequality in developed countries. This idea is called the Kuznet's hypothesis. Galor and Tsiddon (1997) argue that inequality increases in the early stage of growth, because the individual's investment in human capital depends mainly on the individual's social origins. In a mature economy, investment in human capital depends less on social origins. Other studies supporting the Kuznet's hypothesis are Banerjee and Newman (1993) and Aghion and Bolton (1997). Banerjee and Duflo (2003) also argue that growth is an inverted U-shaped function of the level of inequality.

Second, the studies supporting positive effects of inequality on growth are as follows.

Kaldor (1956) argues that workers have a higher propensity of consumption compared with entrepreneurs. This implies that the saving propensity of entrepreneurs is higher than that of workers. Consequently in a more unequal economy, more investments will be made, leading to higher growth. Saint-Paul and Verdier (1993) argue that in a more unequal society, the median voter will elect a higher rate of taxation to finance public education. Consequently, human capital increases and this leads to economic growth. Galor and Tsiddon (1997) argue that in periods of technological inventions, which are likely to occur in periods of high growth, inequality increases because individuals with a high ability get more opportunities of earning money than others. Forbes (2000) asserts empirically that inequality has positive effects on growth.

Third, the following studies support the negative effect of inequality on growth. Alesina and Rodrik (1994) argue that inequality affects growth through democracy. A median voter will support high taxation under high inequality, which impedes growth. Persson and Tabelline (1990) assert empirically that equity is positively correlated with the rate of growth. Persson and Tabellini (1994) argue that in an unequal society, political decisions are more likely to produce economic policies that allocate less benefits to growth promoting activities, such as accumulation of capital and productive knowledge. Alesina and Perotti (1996) argue that inequality increases socio-political instability by fueling social discontent. High socio-political instability creates uncertainty and deters investment, and consequently growth. Rodrik (1998) argues that domestic conflicts, which can be fueled by inequality, have detrimental effects on growth. Bourguignon (1998) emphasizes the growing importance of the costs associated with investing in one's protection when violence increases. Go, Nikitin, Wang and Zou (2007) assert empirically that, as inequality increases, economic performance deteriorates. Kurita and Kurosaki (2011) use household data compiled in the Philippines and Thailand. The empirical results suggest that inequality reduces growth. There are studies asserting that the effects of inequality on growth are based on the credit market imperfection assumption. When credit market is not perfect, lenders have difficulties in distinguishing investments which are likely to fail or succeed, which is a typical situation of the asymmetric information. Lenders will demand collateral to borrowers in response. Borrowers with low level of wealth cannot afford collateral and their investment plans are likely to be wasted, leading to inefficiency and low growth. This logic is applicable not only to the efficiency loss of physical investment, but also to the efficiency loss of investment in human capital. Poor people have difficulties in establishing business and financing education for themselves and their children. They also have bad access to health care service. Piketty (1993), Galor and Zeira (1993), Banerjee and Newman (1993), Aghion and Bolton (1997) and Go, Nikitin, Wang and Zou (2007) use the imperfect credit market assumption.

Fourth, there are studies which cannot be sorted into the afore-mentioned three types of effects of inequality on growth. They assert that the relationship between inequality and

growth depends on other parameters. Benabou (1996) builds a model in which inequality is associated with high growth in the short run, but not in the long run. Xia (2010) argues that depending on the interest rate and the discount rate, inequality may have either positive or negative effects on growth. Bandyopadhyay and Tang (2011) assert that the effects of inequality may be positive or negative on growth depending on immigration shocks and redistribution policy.

There is a consensus that poverty affects growth negatively. The following reasons for poverty traps show that the people in the economies behave different and these differences can be explained by being poor or rich. The effects of poverty traps on growth are obviously negative. Moreover, any economy which excludes some segment of its population from productivity faces an efficiency loss. Poverty traps come into existence from various causes. First, investment in physical or human capital is only possible if a certain level of income is attained so that saving is possible. Second, corrupt institutions, an unequal allocation of property or customs that exclude parts of the society from production may perpetuate poverty. Therefore, the poor faces a coordination problem, which makes it unlikely that he experiences reformation. Third, high population growth among the poor impairs the growth of the poor, because the same amount of capital distributes among more people. Bowles, Durlauf and Hoff (2006) and Sachs, McArthur, Schmidt-Traub and Kruk (2004) identify several causes for poverty traps. See Kraay (2006), Lokshin and Ravallion (2004), Jalan and Ravallion (2004), Mesnard and Ravallion (2006) and McKenzie and Woodruff (2006) for empirical works supporting the existence of poverty traps.

Poverty has negative effects on growth under credit market imperfection, because poor people have difficulties in borrowing since they do not have collateral, and even if it is possible to borrow they have to pay high interest rates. This makes investment in physical capital or human capital not attractive or even impossible. Perry, Lopez and Maloney (2006) support the negative effects of poverty on growth under credit market imperfection. They show negative effects of poverty on growth empirically. Lopez and Servén (2009) argue that poverty deters investment and consequently growth and supports this empirically. Ravallion (2010) argues that financial market development influences poverty and at the same time, poverty may influence the development of the financial market. Then poverty has even larger negative effects on growth. Furthermore, low nutritional intakes of the poor deter the productivity and growth. This averse effect on productivity can also influence the offspring of the poor. See Cunha and Heckman (2007) and Lopez and Servén (2009). Azariadis (1996) argues that the poor may have higher preference on consumption than saving, because of their lower life expectancy. Therefore, poverty is associated with low saving and investment and leads to low growth.

According to the literature, the size of the middle class promotes economic growth. For example Landes (1999) and Adelman and Morris (1969) find that the large size of the

middle class is one of the reasons for industrialization and economic growth of Europe. Alesina (1994) argues that in a society with bimodal income distribution, social conflicts are more likely, which hinder growth. Birdsall, Graham and Pettinato (2000) and Sridharan (2004) support that a minimum size of the middle class is crucial when wanting to reform. Acemoglu and Zilibotti (1997) and Doepke and Zilibotti (2005) assert that a large middle class promotes entrepreneurship. As the middle class demands high quality goods it promotes growth for example in the model of Murphy, Schleifer and Vishny (1989). Easterly (2001) supports the positive effect of the size of the middle class empirically. This shows that the economy as a whole behaves more efficient, when the size of the middle class is large.

The literature review shows that poverty, inequality and the size of the middle class affect economic growth. The effects of poverty on growth are considered to be negative, whereas there is no consensus on the effects of inequality on growth. The size of the middle class is regarded as growth promoting. But if the economies behave so different according to these variables, the question arises, how informative the mean returns to the drivers of economic growth are. The literature review shows for example that there is no reason to believe that poor countries have the same return to investments in physical capital than rich countries. Imagine a growth regression of the form

$$growth = \beta * (growth\ driver) + error$$

and imagine that the sample is clearly divided into poor and rich countries. In this situation, it is very likely to hold that $\beta_{poor} \neq \beta_{rich} \neq \beta_{mean}$. The mean coefficient only reflects a theoretical situation that might not be fulfilled in any of the country groups. Furthermore, the deviations from the mean coefficient are highly suspicious to move simultaneously with the dependent variable, which implies an endogeneity problem. If for example poor countries have a smaller return to the growth driver than the rich countries, this difference is very likely to move simultaneously with the growth performance, as there must be some reason for that the poor countries are poor and that the rich countries are rich. Furthermore, there are problems when putting the model to data. Poor countries have systematically weaker databases and therefore, the estimation of β_{mean} is highly suspicious to suffer from a sample selection bias. This is not the case if we separate the two coefficients β_{poor} and β_{rich} from the beginning.

In real growth regressions, we have many different growth drivers and a large set of countries that cannot clearly be separated in the two distinct groups *poor* and *rich*. However, the aforementioned problems are the same. The two-groups-example motivates to estimate the growth regression with a variable-coefficients model in which a "continuous transition" from poor to rich is possible. This transition is explained by the country's individual levels of poverty, inequality and middleclass in each year.

In this situation, there is another reason for estimating with variable coefficients. Many authors add several variables to the growth regression and understand growth as a theory of

everything. This however only shows that the added variables are somehow correlated to what cannot be explained by the growth regression without these variables. It is not clear whether these extra variables really identify drivers for growth. When only adding a lot of variables to the growth regression, we loose economic justification. For this reason, we aim to stay close to the growth model by Mankiw, Romer and Weil (1992), for which we have economic justification. Estimating with variable coefficients allows staying close to the underlying model, as the set of growth drivers is not extended.
This text is divided into four sections. Section (4.2) is itself divided into three subsections. Subsection (4.2.1) gives an account of the augmented Solow growth model, deals with collecting reasonable measures of the explanatory variables from the data and explains, how growth regressions with these data are conducted. Subsection (4.2.2) deals with the different estimation methods of growth regressions. We explain why random effects models and the System GMM are unfavorable methods and explain the advantages of estimating with fixed effects. When doing this, the endogeneity bias is identified to be negligible small. This discussion allows formulating the desired variable-coefficients model in subsection (4.2.3).
The results of estimation are given in section (4.3). It is divided into two subsections. Subsection (4.3.1) gives the results of the variable-coefficients model applied to mean growth per worker. It is shown that the growth driving coefficients differ dramatically and this can, to a large extend, be explained by the amount of poverty and inequality and the size of the middle class. But this motivates another question which is addressed in subsection (4.3.2). If measures of the income distribution affect the growth behavior of the economy, then how do the poorer and the richer part of the economy grow. We show that the coefficients of the growth drivers of the growth of the richer part of the economy differs greatly from that poorer part. These differences naturally affect the measures of the income distribution, which in turn affect the growth path of the GDP per worker.
Section (4.4) finally concludes.

4.2 Statistical Modelling and Data Collection

4.2.1 The Model, the Data and Growth Regressions

In the last 20 years, many papers about growth empirics appeared. Usually, the authors add some basic explanatory variables that have been identified as drivers for growth and extend this set of variables by a special variable that is concerned with the topic the author wants to investigate. Therefore, numerous variables have been added to growth regressions with the consequence that growth can be seen as a theory of everything. It can be criticized that the extra variables are only identified to be somehow correlated to what cannot be explained by the basic explanatory variables, but they cannot be identified to be real drivers

for growth. We use the growth model from Mankiw, Romer and Weil (1992) to justify the growth regressions. This model explains the growth of the GDP per worker between an initial time-point 0 and a final time-point t

$$\begin{aligned}\ln(\frac{Y(t)}{L(t)}) &= (1-\exp(-\lambda t))\ln(A(0)) + gt + \exp(-\lambda t)\ln(\frac{Y(0)}{L(0)}) \\ &(1-\exp(-\lambda t))\frac{\alpha}{1-\alpha-\beta}\ln(s_K) + (1-\exp(-\lambda t))\frac{\beta}{1-\alpha-\beta}\ln(s_H) \\ &-(1-\exp(-\lambda t))\frac{\alpha+\beta}{1-\alpha-\beta}\ln(n+g+\delta).\end{aligned} \tag{4.1}$$

$L(t)$ is the labor force at time t and grows with rate n. $Y(t)$ is the GDP at time t. It is assumed that the fraction s_K of the GDP is invested in physical capital and the fraction s_H is invested in human capital. $A(t)$ is the productivity that characterizes the country's transformation capabilities. It grows with rate g. δ is the depreciation rate of capital. Therefore, the total depreciation rate is $n+g+\delta$. λ is the convergence rate. See Mankiw, Romer and Weil (1992) for a more detailed description of these notations. This equation motivates to collect time-series for every country and each time-point. We understand every time-point in the time-series as an initial time-point to explain the GDP per worker for the next time-point according to (4.1).

When estimating the parameters of this growth equation, the objective is to collect long time-series for as many countries as possible for which we can guarantee good data quality. The data to be collected are the per worker GDP's, the depreciation rates, the investments in physical capital and a series consisting of a proxy for human capital. Koehler, Sperlich and Vortmeyer (2011) collect and smooth data for a wide range of countries using the information sources Penn World Table 6.3 (PWT) published by Heston, Summers and Aten (2009), World Bank's World Development indicators and Barro and Lee (2010). We describe their data set in what follows. The observations are obtained yearly from 1960 to 2007 for 81 countries. Koehler, Sperlich and Vortmeyer (2011) collect and smooth GDP's per worker because this addresses the question how much each productive worker contributes on average to the growth of his country, which is closer to the model by Mankiw, Romer and Weil (1992) than the per capita values. As these series follow business cycles, Koehler, Sperlich and Vortmeyer (2011) smooth the data using the Hodrick-Prescott filter (see Hodrick and Prescott (1997)). The logarithm of the per worker GDP of country i at time t is denoted by y_{it}. The depreciation rate, according to the model by Mankiw, Romer and Weil (1992), is the sum of the depreciation rate of capital, the growth rate of productivity and the population growth. Koehler, Sperlich and Vortmeyer (2011) approximate the sum of the depreciation rate of capital and the growth rate of productivity by 5 % per year for all countries and collect data for the population growth of the countries for each year. The series of the logarithm of the depreciation rate of country i at year t is denoted by lnn_{it}. The saving rate of the economy is approximated by the relative investment share of the real GDP. As the series follow business cycles which are

not addressed by the model by Mankiw, Romer and Weil (1992), Koehler, Sperlich and Vortmeyer (2011) smooth the data. The logarithm of these series according to country i at year t is denoted by $lnsk_{it}$. The basis for the proxy for human capital is the educational attainment data from Barro and Lee (2010). In order to transfer this variable into a yearly frequency, Koehler, Sperlich and Vortmeyer (2011) extrapolate the missing values by interpolation splines. We denote the logarithm of the yearly educational attainment data of country i and year t by $lnattain_{it}$.

The data from Koehler, Sperlich and Vortmeyer (2011) are subject to a selection process. Heston, Summers and Aten (2009) introduce a country rating system based on the number of participations in worldwide benchmark surveys, the variation of the accessible data and the quality of the statistical methods applied. This results in a grading scheme from A to D with descending order. A rating of D is regarded as too weak to be included in the sample. Therefore, only countries with a grading from A to C are incorporated in the sample. Furthermore, only complete time-series are incorporated for the relevant variables. This also excludes countries that were separated in a sub-period, for example Germany and the countries of the Soviet Union, as their incorporation would have made it necessary to unify several countries to one country or to split one country in a given period in several countries. The loss of data quality when doing this is unclear. The selection process of excluding countries that are D-graded can cause a problem. Poor countries have weaker databases and are more likely to be excluded. Therefore, the estimation of the mean coefficients of equation (4.1) would be highly suspicious to suffer from a sample selection bias. However, the varying coefficients model is less prone to suffer from this sample selection bias, as the level of poverty, inequality and middle class is controlled for. In the same way, excluding the countries that were separated can cause a problem. But the countries that are excluded as a result of this rule do not show structurally similarities. Therefore, if there is a sample selection bias resulting from this rule, we assume it to be very small.

For the purpose of collecting data for measuring poverty, inequality and the size of the middle class, we use the income distribution data from Sala-i-Martin (2006). This ensures that the different measures from the income distribution are calculated by one and the same source of information. The data from Sala-i-Martin (2006) consist of complete time-series from the year 1970 to the year 2000 of some points, at which the world distribution of income is evaluated, for a large range of countries. The data contain two parts of information. The first part is the GDP per capita which is available by the PWT. The second part is the dispersion around the mean. To measure the dispersion, the first task is to estimate the quintile income shares. When doing this, Sala-i-Martin (2006) uses the microeconomic income surveys reported by Deininger and Squire (1999) and updated by the United Nations University's World Institute for Development Research. Obviously, survey data are not available for every country and every year. For countries with good

data quality, meaning that the GDP per capita is available and income surveys are reported for various years, the quintiles are estimated from the survey data in the years in which they are available and missing values are estimated using a linear time-trend forecast. For a country for which the GDP per capita is available, but there is only one survey available, the quintiles for the years in which no survey is available are estimated using the information concerning the time-trends from neighboring countries for which various surveys are available. Two countries are defined as neighbors, if they are in the same region. The regions are those given by World Bank, namely East Asia and Pacific, Eastern Europe and Central Asia, Latin America and the Caribbean, Middle East and North Africa (MENA), South Asia, Sub-Saharan Africa, High-Income Non-OECD and High-Income OECD. Having collected the quintiles for the countries for which at least one survey is available, the quintiles of a country for which the GDP per capita but no survey is available can be conducted by using the averaged information from the neighboring countries for which at least one survey is available. Countries for which no GDP data are available are excluded from the sample. Having collected these income shares, a kernel density estimator is used to approximate the underlying density. When doing this, the choice of the smoothing parameter is important (see Heidenreich, Schindler and Sperlich (2010)). Sala-i-Martin (2006) decided to use the same appropriate bandwidth for all countries and all years. Afterwards, the estimated density functions are evaluated at a hundred points. The result is the data set from Sala-i-Martin (2006) consisting of the hundred points of the estimated income distribution for 138 countries for every year, starting in the year 1970 end ending in the year 2000.

For measuring poverty, we use the fraction of the total population with income less than one dollar per day. Following Sala-i-Martin (2006), one dollar refers to the price level of 1996. Sala-i-Martin (2006) reports a hundred evaluated points of the distribution of income. From this, we plot the income distribution for each country and each year. Figure (4.1) demonstrates this. The x-axis is the income level and the y-axis is the population. We use natural splines to interpolate the points. The fraction of the total population with income less than one dollar per day is the area A divided by the total area $A+B$ in figure (4.1). We denote this fraction of country i at year t by $pov1d_{it}$.

For measuring inequality, we use the well known Gini coefficient. We checked that the regression results do not change when using the Theil index instead of the Gini index. First of all, we construct the Lorenz curve for each country and year. Figure (4.2) demonstrates this. The x-axis is the cumulative share of people from lowest to highest incomes and the y-axis is the cumulative share of income. The cumulative share of people from lowest to highest incomes is constructed by adding up the population associated with the points given by Sala-i-Martin (2006). The cumulative share of income is constructed by adding up the corresponding incomes. We interpolate the points using interpolation splines. The Gini coefficient is the area A divided by the area A+B in figure (4.2). We denote the Gini

coefficient of country i at year t by $gini_{it}$.
We use a relative definition for the middle class, as different countries with very different levels of incomes are compared. Easterly (2001) proposes to use the income share controlled by the middle three quintiles of the income distribution. This means that the middle class is the share of the total income, that the middle sixty per cent of the population earn. This can be easily constructed from the Lorenz curve. Let g_{it} be the Lorenz curve of country i at year t and let pop_{it} be the total cumulative population. In this situation, the relative middle class of country i at year t is

$$middleclass_{it} = \frac{g_{it}(0.8pop_{it}) - g_{it}(0.2pop_{it})}{g_{it}(pop_{it})}.$$

Furthermore, we are interested in collecting the average incomes of the richest and the poorest twenty per cent of a country. Usually, we use per worker GDP to calculate growth. The number of workers is attained by multiplying the total population by the share of working age population. The average per worker GDP for the richest and the poorest twenty per cent is difficult to calculate from Sala-i-Martin (2006)'s data, as these data correspond to per capita values. The effects of multiplying the share of working age population by the average per capita GDP for the richest and the poorest twenty per cent

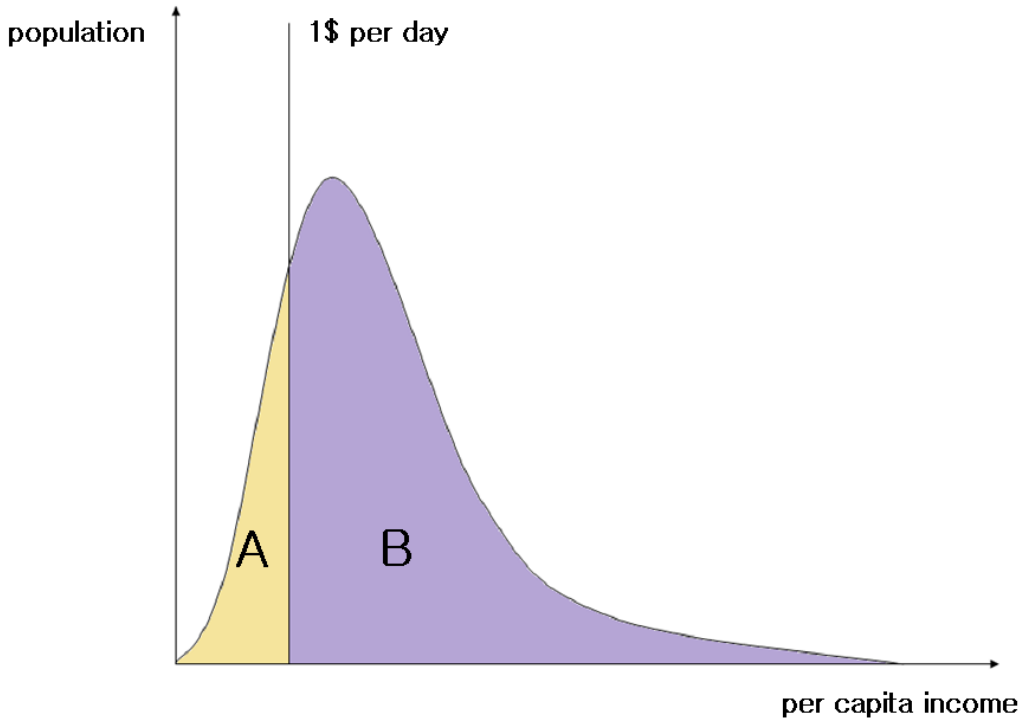

Figure 4.1: *A sketch of the income distribution*

would be distorting, since the share of working age population is likely to differ greatly between the poorest and the richest individuals of the country. Therefore, we use per capita income to calculate the growth of the poorest and richest twenty per cent of the country. A comparison of growth regressions for the total population, when using per capita GDP instead of per worker GDP, shows that the quantitative differences are negligible small. We construct the data from the Lorenz curve. The average GDP per capita of the poorest twenty per cent of country i at year t is

$$\frac{g_{it}(0.2pop_{it})}{0.2pop_{it}}.$$

This measure was introduced by Ravallion and Chen (2003) for measuring pro-poor growth. In the same way, the average GDP per capita of the richest twenty per cent of country i at year t is

$$\frac{g_{it}(pop_{it}) - g_{it}(0.8pop_{it})}{0.2pop_{it}}.$$

The reference year of the Penn World Table 6.3 is 2005, whereas the reference year of Sala-i-Martin (2006)'s data is 1996. As both data sets are combined, we multiply these

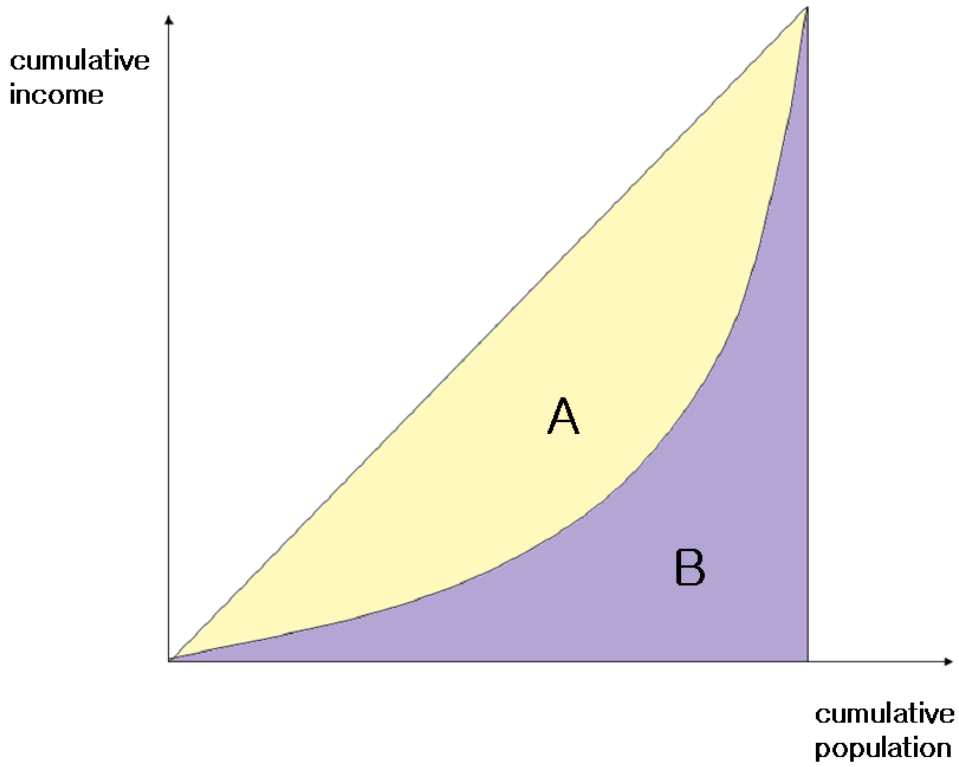

Figure 4.2: *A sketch of the Lorenz curve. The Lorenz curve is the line between the segments A and B.*

fractions by $80.31577/100$, which is the Consumer Price Index of 1996 divided by that of 2005 from the World Bank. The series are subject of business cycles. Koehler, Sperlich and Vortmeyer (2011) smooth the logarithm of the average GDP series using the filter by Hodrick and Prescott (1997). They use the smoothing parameter 100. As we combine our data with the data collected by Koehler, Sperlich and Vortmeyer (2011), we use the same smoothing parameter for smoothing the price adjusted logarithm of the average GDP per capita of the poorest twenty per cent of country i at year t and that of the richest twenty per cent. We denote these variables by $lny_{rich,it}$ and $lny_{poor,it}$ respectively. When combining the data by Koehler, Sperlich and Vortmeyer (2011) and those mentioned above, we end up with complete time-series starting in 1970 and ending in 2000 for the 81 countries that the data by Koehler, Sperlich and Vortmeyer (2011) contain.
Having collected the complete time-series for a large range of countries, equation (4.1) justifies the following regression equation

$$y_{it} = \rho * y_{i(t-l)} + x'_{it}\beta + \eta_i + \nu_{it}, \tag{4.2}$$

where ν_{it} is an error with expectation zero, l is some lag depth and the vector x_{it} consists of the explanatory variables given by the model by Mankiw, Romer and Weil (1992). This regression equation was for example used by Islam (1995). However, the assumption that the β-coefficients are independent from the state of development of a country seems to be strong. In this paper, we investigate how the β-coefficients change, when they are dependent of measures of the income distribution of every country and every year. More precisely, we assume

$$y_{it} = \rho y_{i(t-l)} + x'_{it}\beta_{it} + \eta_i + \nu_{it},$$

where the β_{it}'s are explained by a linear model with an intercept, poverty, inequality, the middle class and their squared values.

4.2.2 Methods To Estimate Growth Regressions

In this subsection, we discuss the methods to estimate growth regressions with random variable coefficients. It suffices to deal with estimating the regression equation of the form (4.2), as the random-coefficients model consists of two regression equations that can be written as one regression equation that has the form of a linear regression equation without random coefficients. Furthermore, it suffices dealing with the case that the lag depth equals one, as higher lag depths can be incorporated very easily. We denote $y_i^t = (y_{i1}, \ldots, y_{it})$ and $x_i^t = (x'_{i2}, \ldots, x'_{it})$ and assume that

$$\left\{ \left(y_i^T, x_i^T\right) \ , \ i = 1, \ldots, n \right\}$$

is a number of independent observations from the same probability distribution with finite first and second order moments. The errors ν_{it} are assumed to have zero means and finite

second moments. We discuss three different ways to estimate growth regressions. First, the Difference and the System GMM, second, the models based on the idea of error components and third, the idea of estimating each fixed effect individually.
The widely known methods to estimate growth regressions based on the GMM procedure are the Difference GMM and the System GMM. An application of the Difference GMM to growth regression is given by Caselli, Esquivel and Lefort (1996) and of the System GMM is given by Hoeffler (2002). These methods do not assume strict exogeneity. Instead they only assume that at a given time-point, present and future residuals are uncorrelated to present explanatory variables. Concerning growth regressions, this is based on the idea that the economies are able to choose their drivers for growth as a reaction of a shock. This assumption implies that lagged explanatory variables are valid instruments for the regression equation in differences, in which the country-specific effects disappear. Applying the usual GMM procedure on the equation in differences yields the Difference GMM estimator. This was first proposed by Arellano and Bond (1991). Bond, Hoeffler and Temple (2001) note that the Difference GMM uses weak instruments because the series of the logarithms of GDP's per capita is highly persistent and recommend the System GMM. In general, Blundell and Bond (1998) show that the instruments of the Difference GMM estimator are weak when the autoregressive coefficient is close to one. The System GMM estimator augments the set of moments of the Difference GMM estimator by additionally assuming moment conditions for the level equation. The additional assumption of the System GMM estimator is that lagged explanatory variables are allowed to be correlated to the fixed effects, but the correlation of two succeeding time-points and the fixed effects is of the same magnitude. This implies that differenced explanatory variables are instruments for the equation in levels. It suffices to assume this for all differenced x-variables and only the difference of the first two lagged dependent variables, since it can be shown that in this case, all subsequent differenced lagged dependent variables are also valid instruments. This is what Blundell and Bond (1998) call the initial condition. When additionally assuming these instruments for the equation in levels, Blundell and Bond (1998) observe a dramatic efficiency gain when the autoregressive coefficient is close to one. Roodman (2006) gives access to the System GMM by implementing it in Stata. There is no doubt, that the System GMM is a very popular method for estimating growth regressions. We now argue why we think that the System GMM not adequate for our purposes.
First of all, problems occur because of the large number of instruments. This results in a bias for which Arellano (2003) gives analytical evidence when the number of observations and the length of the time-series go to infinity. Windmeijer (2005) observes a decreasing bias when applying the Difference GMM if the instrument count is reduced. This bias is a general problem of GMM as shown by Tauchen (1986) or Ziliak (1997). Apart from the bias, problems occur when estimating the optimal weighting matrix in the GMM procedure, as the number of elements to be estimated is quadratic in the number of

instruments and therefore quartic in T. Moreover, the elements of the optimal weighting matrix are fourth moments of the underlying distributions, because they are second moments of the result of differenced variables times variables. Roodman (2009) notes that a common symptom for estimations of the weighting matrix is that they are singular. Therefore, the generalized inverse rather than the inverse is calculated. This can give results that are far away from the theoretical ideal. This breakdown tends to occur as the number of instruments approaches n. Therefore, n can be seen as a general benchmark for the number of instruments. In our case, the number of instruments is much larger than n. The Hansen J-Test (see Hansen (1982)) usually checks the validity of instruments, but as for example Bowsher (2002) observes in simulation studies, a too large number of instruments weakens the test dramatically. Roodman (2009) notes that in case of too many instruments the weights of those moments that are least well satisfied are too small. We conclude that we do not have a reliable test available that tells us how many and which instruments to choose. Because of the large number of instruments when the length of the time-series is large, Roodman (2009) provides methods to reduce the instrument count. This is limiting the lag-depth and collapsing. Collapsing means that instead of assuming that all lagged explanatory variables are individual instruments for the equation in differences, it is only assumed that the time-wise sum of the lagged variables is a valid instrument and in the same way the additional instruments for the equation in levels are time-wise summed up to one instrument. Applying one of these methods does not help to reduce the instrument count to a level where the problems mentioned above are not relevant. Applying both methods can reduce the instrument count dramatically, but reducing the number of instruments makes it possible to apply the System GMM but has large drawbacks in terms of efficiency.

Another problem of the System GMM is that it needs the very strict assumption that the residuals are not correlated. When this assumption is slightly violated, System GMM suffers from a bias of unknown magnitude. In growth regressions it is generally very unlikely that the residuals are perfectly uncorrelated because growth can be a theory of everything and therefore there are always drivers for growth which are not incorporated in the growth regression.

Moreover, the initial condition is a very strict assumption and unlikely to be fulfilled in growth regressions. Roodman (2009) shows that the validity of the initial condition is equivalent to that the correlation of the deviations from the long-run mean of the series y_{it} and the fixed effect is zero. On the other hand, if the initial condition is not true when the observation period begins, this correlation converts to zero. However, the speed of convergence is dramatically slow if the autoregressive coefficient is close to one and this is exactly the reason for using the System GMM rather than the Difference GMM. The Hansen J-Test does not offer help to test the validity of the moment conditions, unless we reduce the instrument count by limiting the lag depth and collapsing, which results in

dramatic efficiency problems.
Because of either the efficiency problems when reducing the instrument count or the technical problems when not reducing the instrument count, the bias of unknown magnitude when the residuals are not perfectly homoscedastic and uncorrelated and the lack of clarity of the validity of the assumptions that allow to apply the System GMM, we do not use it.
In this paragraph, we give an account of Random Effects estimators. They assume that the country-specific effects are independently drawn and come from similar distributions. Random Effects estimators aim to derive facts about the process that generates the country-specific effects. It is assumed that the explanatory variables and the country-specific effects are uncorrelated, that the country-specific effects are independent, that their variances exist and that these variances are equal. Furthermore, we assume strict exogeneity even though we have a lagged variable. We will argue in this subsection that the bias resulting from this assumption is negligible small. We decompose the country-specific effects in the sum of a general intercept and a country-specific error. This results in a regression equation with an error structure that consists of two components, the violations of the fixed effect to its mean and the residual of the original regression equation. In this situation, the simplest approach is to pool all data and then apply OLS. The pooled estimator provides consistent estimates. The errors are correlated, as they consist of two components from which one is time-constant. Therefore, a robust choice to estimate the coefficients yields more efficient estimates. The disadvantage is that the assumption that the explanatory variables and the country-specific errors are uncorrelated is very unlikely to be fulfilled in growth regressions. The model by Mankiw, Romer and Weil (1992) indicates that the total country-specific effect is determined by the growth rate of technological change, the convergence rate and the initial level of technology. Whereas the growth rate of the technological change and the convergence rate can assumed to be constant across countries and over time, this is not true for the initial endowment with production technology. Mankiw, Romer and Weil (1992) mention several influences on initial endowment with production technology, like resources, climate or institutions. These influences are correlated to the explanatory variables. For example developed institutions can increase their level of human capital in the population. Therefore, the assumption that the country-specific errors and the explanatory variables are uncorrelated seems to be too strong. This results in an endogeneity bias of unknown magnitude. Another problem of Random Effects estimators is the idea of viewing country-specific effects as random draws from a common population. For example, there is no reason to assume that the country-specific effects have one and the same equal variance. Furthermore, when viewing country-specific effects as random draws from a common population, is not possible to examine the performance of individual countries, as this is simply not incorporated in the model. We conclude that the disadvantages of Random

Effects estimators are the unrealistic assumptions of uncorrelation of the country-specific effects and the explanatory variables and that the country-specific effects are random draws from a common population. Because of these disadvantages, we do not use Random Effects estimators.
The disadvantages of Random Effects estimators motivate to estimate each country-specific effect as an individual parameter. The Fixed Effects estimator applies the Least-Squares technique to estimate all coefficients of the growth regression equation, including each country-specific effect as a single regressor. We assume strict exogeneity. As the regression equation contains a lagged dependent variable, this assumption causes a bias. This autoregressive bias has been shown by Orcutt and Irwin (1948) and Kendall (1954) for time-series models with fixed time-series length and the results have been extended by Nickell (1981) for panels with fixed T (but $n \to \infty$). In consequence, bias reduction procedures have been proposed, for example Kiviet (1995), Hahn and Kuersteiner (2002) or Phillips and Sul (2007). Phillips and Sul (2007) give precise formulas for the bias of the Within Group estimator as $n \to \infty$. It suffices to look at the bias of the Within Group estimator, because applying the Fixed Effects estimator of ρ and β is the same as applying the Within Group estimator. Afterwards, we see how mistakes in the Within Group estimation step affect the estimation of the fixed effects. Since the true ρ is not known, we calculate biases for different potential values of ρ. The Within Group estimator of the coefficient of the lagged variable is biased downwards and therefore we use it as the smallest ρ to plug in. We calculate these biases for the Within Group estimator of the β-coefficient, namely $\hat{\beta}_{WG}$, since fluctuations result in negligible small differences. The results are given in table (4.1). Note that the length of the time-series equals 30. The biases of the fixed effects listed in this table are the maximum of all absolute values of the biases of each fixed effect. Table (4.1) shows that all biases, apart from that of the coefficient of the lagged variable, are negligible small. Calculating biases when adding more exogenous variables is not necessary since Phillips and Sul (2007) argue that the addition of exogenous variables result in smaller biases. Note that, we only checked that the endogeneity bias caused by the lagged dependent variable is small. Since the economy can choose its growth driving parameters as reaction of a shock, the total endogeneity bias is not known. It is natural to assume that the bias caused by the explanatory variables is much smaller than that caused by the lagged dependent variable itself, which is already negligibly small. We therefore assume in the regressions using the Within Group estimator that the bias that results from the lagged variable (apart from that of the coefficient of the lagged variable itself) is negligible small.

We conclude that the GMM estimators and Random Effects estimators have serious disadvantages. Above all, they suffer from a bias of unknown magnitude. The bias when assuming strict exogeneity and estimating with fixed effects can be shown to be negligible small. Moreover, estimating with Least-Squares yields consistent estimates, even if the

Rho	Bias Rho	Bias lnn	Bias lnsk	Bias lnattain	Max Bias FE
0.9896	$-2.569*10^{-2}$	$-1.049*10^{-5}$	$2.805*10^{-4}$	$-1.116*10^{-4}$	$5.016*10^{-17}$
0.9922	$-2.524*10^{-2}$	$-1.060*10^{-5}$	$1.847*10^{-4}$	$-0.183*10^{-4}$	$4.923*10^{-17}$
0.9948	$-2.477*10^{-2}$	$-1.101*10^{-5}$	$0.876*10^{-4}$	$0.772*10^{-4}$	$4.829*10^{-17}$
0.9974	$-2.429*10^{-2}$	$-1.173*10^{-5}$	$-0.105*10^{-4}$	$1.745*10^{-4}$	$4.731*10^{-17}$
1	$-2.380*10^{-2}$	$-1.278*10^{-5}$	$-1.094*10^{-4}$	$2.737*10^{-4}$	$4.632*10^{-17}$

Table 4.1: *The Nickell Bias with* $T = 30$.

residuals are correlated. Furthermore, when estimating each country-specific effect individually, we do not need the assumption that the country-specific effects come from a common population. Therefore, we estimate each country-specific effect individually and apply Least-Squares.
Before we do this, we discuss some well-known problems that often occur in case of large n and T panels. We discuss the spurious regression problem, the unit-root problem and the cointegration problem. The spurious regression problem comes from the literature of time-series analysis. The problem is known as one yielding a nonzero β-coefficient when regressing two independent and individually integrated processes of order one on one another. Phillips and Moon (1999) provide a concept that extends the arguments about spurious regression in time-series analysis. They show that the issue of spurious regression will not arise for the panel estimates, when the cross-sectional size tends to infinity. In our case the cross-sectional size is 81 which is why we argue that we do not have the problem of a spurious regression. The Unit-Root problem is concerned with the inference of the autoregressive coefficient, when it equals one. When considering the lagged series, the autoregressive coefficient is far away from one, which is why we argue that this is not a problem in our case. There can also be a problem in case of integrated explanatory variables of order one. More precisely, if $x_{it} = x_{it-1} + \varepsilon_{it}$, Kao and Chiang (2000) show that the fixed effects estimator is biased if ε_{it} and v_{it} are correlated. We see no reason for such a correlation and therefore estimate with OLS.

4.2.3 The Variable-Coefficients Model

In this subsection, we introduce a variable-coefficients model, in which the coefficients are dependent on an intercept, poverty, inequality and the size of the middle class. The development of varying-coefficients models with applications in econometrics started in the seventies. For example Wachter (1970) motivates to estimate wage equations with a linear regression equation with variable coefficients. Singh, Nagar, Choudhry and Baldev (1976) introduce a variable-coefficients model which allows incorporating time-trends in the regression that explains the coefficients. A more generalized version of this model, in

which some of the coefficients are functions of other exogenous variables, was introduced by Amemiya (1978). Amemiya (1978) also applies this model to panel data. The following model was indicated by Amemiya (1978) but not explicitly derived. We deal with the case that the lag depth is one as further lag depths can be incorporated easily. Regression equation (4.2) allows for variable intercepts across countries. Other parameters of the model are constant. There is no reason to assume that the β-coefficients are constant over time and individual countries and therefore, we assume that

$$y_{it} = \rho y_{i(t-1)} + x'_{it}\beta_{it} + \eta_i + \nu_{it}. \tag{4.3}$$

Incorporating one β-coefficient for every unit would mean to estimate $n(T-1)k+n+1$ parameters with only $n(T-1)$ data. This does obviously not have a chance to work. We assume that each β_{itk} fulfills the following linear relationship

$$\beta_{itk} = \tilde{z}'_{it}\gamma_k + a_{itk},$$

where $k \in \{1,2,3\}$. $\tilde{z}_{it}$ consists of an intercept, poverty, inequality, the size of the middle class and their squared values. When stacking the β-coefficients we have

$$\begin{pmatrix} \beta_{1it} \\ \beta_{2it} \\ \beta_{3it} \end{pmatrix} = \begin{pmatrix} \tilde{z}'_{it} & 0 & 0 \\ 0 & \tilde{z}'_{it} & 0 \\ 0 & 0 & \tilde{z}'_{it} \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} + \begin{pmatrix} a_{it1} \\ a_{it2} \\ a_{it3} \end{pmatrix} \in \mathbb{R}^3. \tag{4.4}$$

We denote

$$\begin{pmatrix} \tilde{z}'_{it} & 0 & 0 \\ 0 & \tilde{z}'_{it} & 0 \\ 0 & 0 & \tilde{z}'_{it} \end{pmatrix} = Z_{it} \in \mathbb{R}^{3\times M}, \quad \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} = \gamma \in \mathbb{R}^M \text{ and } \begin{pmatrix} a_{it1} \\ a_{it2} \\ a_{it3} \end{pmatrix} = a_{it} \in \mathbb{R}^3.$$

Moreover, we assume that

$$E(a_{it}a'_{js}) = \begin{cases} \Lambda, & \text{if } i = j \text{ and } s = t \text{ and} \\ 0, & \text{else.} \end{cases} \tag{4.5}$$

This means that the correlations of the β's do not change over time and across individuals

$$\begin{aligned} \Lambda &= \begin{pmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} \\ \lambda_{21} & \lambda_{22} & \lambda_{23} \\ \lambda_{31} & \lambda_{32} & \lambda_{33} \end{pmatrix} \\ &= \begin{pmatrix} Var(\beta_{it1}) & Cov(\beta_{it1},\beta_{it2}) & Cov(\beta_{it1},\beta_{it3}) \\ Cov(\beta_{it2},\beta_{it1}) & Var(\beta_{it2}) & Cov(\beta_{it2},\beta_{it3}) \\ Cov(\beta_{it3},\beta_{it1}) & Cov(\beta_{it3},\beta_{it2}) & Var(\beta_{it3}) \end{pmatrix} \text{ for all } i \text{ and } t. \end{aligned}$$

The residuals of (4.3) are uncorrelated and homoscedastic

$$E(\nu_{it}\nu_{js}) = \begin{cases} \sigma^2, & \text{if } i = j \text{ and } s = t \text{ and} \\ 0, & \text{else.} \end{cases} \tag{4.6}$$

In this situation, when stacking the time-series data of (4.3)

$$y_i = \rho y_{i(-1)} + \eta_i \iota_{T-1} + X_i \beta_i + \nu_i \in \mathbb{R}^{T-1},$$

with

$$\begin{aligned}
\iota_{T-1} &= (1,\ldots,1)' \in \mathbb{R}^{T-1},\\
y_{i(-1)} &= (y_{i1},\ldots,y_{i(T-1)})' \in \mathbb{R}^{T-1},\\
X_i &= \begin{pmatrix} x_{i2}' & & & \\ & x_{i2}' & & \\ & & \ddots & \\ & & & x_{iT}' \end{pmatrix} \in \mathbb{R}^{(T-1)\times 3(T-1)},\\
\beta_i &= (\beta_{i2}',\ldots,\beta_{iT}')' \in \mathbb{R}^{3(T-1)} \text{ and}\\
\nu_i &= (\nu_{i2},\ldots,\nu_{iT})' \in \mathbb{R}^{T-1}.
\end{aligned}$$

Furthermore, we stack the time-series data of (4.4)

$$\beta_i = Z_i \gamma + a_i \in \mathbb{R}^{3(T-1)},$$

with

$$\begin{aligned}
Z_i &= (Z_{i2}',\ldots,Z_{iT}')' \in \mathbb{R}^{3(T-1)\times M} \text{ and}\\
a_i &= (a_{i2}',\ldots,a_{iT}')' \in \mathbb{R}^{3(T-1)}.
\end{aligned}$$

Furthermore, it follows from (4.5)

$$E(a_i a_i') = I_{T-1} \otimes \Lambda \in \mathbb{R}^{3(T-1)\times 3(T-1)}$$

and from (4.6)

$$E(\nu_i \nu_i') = \sigma^2 I_{T-1}.$$

After stacking-time series data of (4.3), we additionally stack cross-sectional data

$$y = \rho y_{-1} + C\eta + X\beta + \nu, \tag{4.7}$$

with

$$\begin{aligned}
y_{-1} &= (y_{1(-1)}',\ldots,y_{n(-1)}')' \in \mathbb{R}^{n(T-1)},\\
C &= I_{T-1} \otimes \iota \in \mathbb{R}^{n(T-1)\times n},\\
\eta &= (\eta_1,\ldots,\eta_n)' \in \mathbb{R}^n,\\
X &= \begin{pmatrix} X_2 & & & \\ & X_3 & & \\ & & \ddots & \\ & & & X_n \end{pmatrix} \in \mathbb{R}^{n(T-1)\times 3n(T-1)},\\
\beta &= (\beta_1',\ldots,\beta_n')' \in \mathbb{R}^{3n(T-1)} \text{ and}\\
\nu &= (\nu_1',\ldots,\nu_n')' \in \mathbb{R}^{n(T-1)}.
\end{aligned}$$

Furthermore, after stacking time-series data of (4.4), we additionally stack cross-sectional data

$$\beta = Z\gamma + a \in \mathbb{R}^{3n(T-1)}, \tag{4.8}$$

with

$$\begin{aligned} Z &= (Z_1', \ldots, Z_n')' \in \mathbb{R}^{3n(T-1)\times M} \text{ and} \\ a &= (a_1', \ldots, a_n')' \in \mathbb{R}^{3n(T-1)}. \end{aligned}$$

We plug (4.8) into (4.7)

$$\begin{aligned} y &= \rho y_{-1} + C\eta + X\beta + \nu \\ &= \rho y_{-1} + C\eta + \\ &\begin{pmatrix} X_2 & & & \\ & X_3 & & \\ & & \ddots & \\ & & & X_n \end{pmatrix} \left[\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} \gamma + \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \right] + \nu \\ &= \rho y_{-1} + C\eta + \begin{pmatrix} X_1 Z_1 \\ X_2 Z_2 \\ \vdots \\ X_n Z_n \end{pmatrix} \gamma + \begin{pmatrix} X_1 a_1 \\ X_2 a_2 \\ \vdots \\ X_n a_n \end{pmatrix} + \nu \\ &= \rho y_{-1} + C\eta + W\gamma + u \in \mathbb{R}^{n(T-1)}, \end{aligned}$$

with

$$\begin{aligned} W &= (Z_1' X_1', \ldots, Z_n' X_n')' \in \mathbb{R}^{n(T-1)\times M} \text{ and} \\ u &= (a_1' X_1', a_2' X_2', \ldots, a_n' X_n')' + \nu \in \mathbb{R}^{n(T-1)}. \end{aligned}$$

The regression equation

$$y = \rho y_{-1} + C\eta + W\gamma + u \in \mathbb{R}^{n(T-1)} \tag{4.9}$$

has $n + M + 1$ parameters. If the matrix

$$(y_{-1}, C, W) \in \mathbb{R}^{n(T-1)\times(n+M+1)}$$

has full column rank, the model can be identified. The following calculation shows that this model has uncorrelated but heteroscedastic errors

$$
\begin{aligned}
E\left[uu'\right] &= E\left[\left(\begin{pmatrix} X_1a_1 \\ X_2a_2 \\ \vdots \\ X_na_n \end{pmatrix} + \nu\right)\left(\begin{pmatrix} X_1a_1 \\ X_2a_2 \\ \vdots \\ X_na_n \end{pmatrix} + \nu\right)'\right] \\
&= \begin{pmatrix} X_1E(a_1a_1')X_1' & & & \\ & X_2E(a_2a_2')X_2' & & \\ & & \ddots & \\ & & & X_nE(a_na_n')X_n' \end{pmatrix} + \sigma^2 I_{n(T-1)} \\
&= \begin{pmatrix} x_{12}'\Lambda x_{12} & & & & & & \\ & \ddots & & & & & \\ & & x_{1T}'\Lambda x_{1T} & & & & \\ & & & \ddots & & & \\ & & & & x_{n2}'\Lambda x_{n2} & & \\ & & & & & \ddots & \\ & & & & & & x_{nT}'\Lambda x_{nT} \end{pmatrix} + \sigma^2 I_{n(T-1)} \\
&\in \mathbb{R}^{n(T-1)\times n(T-1)}.
\end{aligned}
$$

We conclude that when we estimate (4.3) using (4.4), with the assumptions (4.5) and (4.6), it suffices to estimate equation (4.9). We now discuss how to estimate the coefficients of (4.9). Applying OLS gives a consistent estimator, but as we have heteroscedasticity, Generalized-Least-Squares yields a more efficient estimator. Amemiya (1978) proposed to estimate the β's first, using

$$
y = \rho y_{(-1)} + C\eta + \begin{pmatrix} X_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}\beta_1 + \begin{pmatrix} 0 \\ X_2 \\ \vdots \\ 0 \end{pmatrix}\beta_2 + \ldots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ X_n \end{pmatrix}\beta_n + \nu \in \mathbb{R}^{n(T-1)}
$$

and then estimate γ, Λ and σ^2 using the estimated β's. This would mean to estimate $3n(T-1)+n+1$ parameters with only $n(T-1)$ data. This does obviously not have a chance to work. One possibility to obtain a feasible estimator is to make use of the linear structure of $E(uu')$ and formulate an auxiliary regression in the following way. We first apply OLS to (4.9) and obtain consistent estimators for ρ, η and γ. Then, we extract the residuals from this regression and regress the squared residuals on the variables given in the linear structure of $E(uu')$ and estimate σ^2, λ_{11}, λ_{22}, λ_{33}, λ_{12}, λ_{13} and λ_{23}. The reciprocal fitted values of this regression can be used as weights to estimate the coefficients of (4.9). The idea is to iterate this, until the estimated coefficients ρ, η and γ do not change

from one step to another up to a small but specified error. This method has the problem that, when having estimated the residuals of (4.9) in one step, one has to find the λ's such that the matrix Λ has the characteristics of a covariance matrix, such that it is symmetric and positive definite, to obtain weights for the next step. Applying OLS on the auxiliary regression does not guarantee this and results in negative weights. Estimating the Cholesky decomposition of the matrix Λ is not possible because of the resulting multicollinearity. Incorporating the symmetry condition is easy and it is possible to formulate inequality conditions for the λ's, such that Λ fulfills some of the characteristics of a covariance matrix. We could for example apply the method of Goldfarb and Idnani (1982) and Goldfarb and Idnani (1983) to force the diagonal elements of Λ to be positive. We note that the resulting optimization procedure that calculates the λ's itself has errors and we are not sure if the result of iterated least-squares combined with the iterated solution of the optimization procedure in every step converges to the desired result.
Therefore, we estimate the coefficients of (4.9) as follows:

(1) We estimate the coefficients of (4.9) in the first step using OLS,

(2) we extract the residuals,

(3) we estimate the coefficients of (4.9) again using least-squares with the reciprocal squared residuals as weights.

Extracting the residuals from this regression (repeat step (2)) gives weights for the next regression (repeat step (3)) and so on. We iterate this procedure, until the sum of squared differences of the coefficients from one step to the next is smaller than 0.005. This ensures that the average squared difference from one step to another is approximately 0.00005.

4.3 Results

4.3.1 The Effects on Economic Growth

In this subsection, we investigate the effects of poverty, inequality and the middle class on the coefficients of the growth equation of the GDP per worker. Running the regressions using a one year lagged dependent variable and contemporaneous explanatory variables has three drawbacks. First, the one year growth time-series shows little variation so that the coefficient of the lagged dependent variable is almost one and all other coefficients are very small. This can lead to a spurious regression problem. Second, we only checked that the endogeneity bias caused by the lagged dependent variable is small. Since the economy can choose its growth driving parameters as reaction of a shock, the regression is suspected to suffer from an endogeneity bias. It is natural to assume that the bias caused by the

explanatory variables is much smaller than that caused by the lagged dependent variable itself, which is already negligibly small. Nevertheless, we do not know the exact correlation of explanatory variables and the error and cannot give precise formulas for the bias as done by Phillips and Sul (2007). Third, we aim for comparison of our results with that of other authors, who refer their regressions to time horizons, taking either averaged or initial explanatory variables to represent the time horizons. Especially the first two drawbacks mentioned allow impeaching the credibility of the results obtained by the one year growth equation. Therefore, we estimate a lagged regression equation with $x_{it} = (lnn_{i(t-3)}, lnsk_{i(t-3)}, lnattain_{i(t-3)}) \in \mathbb{R}^3$ and the dependent variable is of the lag depth three. Furthermore,

$$\tilde{z}_{it} = (1, pov_{i(t-3)}, pov^2_{i(t-3)}, gini_{i(t-3)}, gini^2_{i(t-3)}, middleclass_{i(t-3)}, middleclass^2_{i(t-3)})' \in \mathbb{R}^7,$$

which implies $M = 21$. In this case the time-series covers 31 years, namely from the year 1973 to the year 2003. Furthermore, the cross-sectional size is $n = 81$. When displaying the estimated coefficients, we report the level of significance of the coefficient $(.)$ by $(.)^{***}$ if the p-value is almost zero, $(.)^{**}$ if the p-value is smaller than 0.01, $(.)^{*}$ if the p-value is smaller than 0.05 and $(.)$ if the p-value is larger than 0.1. The estimated autoregressive coefficient is 0.9318^{***}. The regression that explains β_1, which is the coefficient of *lnn* is

$$\begin{aligned}\beta_{1it} = &-2.1952^{***} + 1.3705^{***} pov_{i(t-3)} - 3.4198^{***} pov^2_{i(t-3)} + 2.9553 gini^{***}_{i(t-3)} \\ &- 0.8132^{***} gini^2_{i(t-3)} + 0.8917^{***} middleclass_{i(t-3)} + 2.9860^{***} middleclass^2_{i(t-3)}.\end{aligned}$$

A graphical illustration of this is given in figure (4.4). The plots show the evolution of β_1 if the variables *pov* (upper left), *gini* (upper right) or *middleclass* (bottom left) change. For the variables that are constant in each plot, we plug in the corresponding averaged variables. The boxplots (bottom right) show the estimated coefficients stratified for different country groups. When choosing the country groups, the task is to identify similar countries and combine them to groups such that each group contains a minimum number of representatives. We choose the groups Asia, Latin (which is identical to the Latin America and Caribbean group defined by the World Bank), sub-Saharan Africa (SSA), High Income (which consists of the High Income OECD and the High Income Non-OECD group as defined by the World Bank and is denoted by HI) and the group of other countries (which consists of the Middle East and North Africa group and the Eastern Europe group as defined by the World Bank). The latter group mentioned is formed because the results of their representatives are similar. The group of Asia consists of 11 countries, the Latin group of 19, the SSA group of 17, the HI group of 27 and the group of other countries has 7 representatives. Figure (4.3) demonstrates graphically how the mean poverty rates (upper left), the mean Gini rates (upper right) and the mean middle class rates (bottom middle) of each of the groups evolve over time. Note that these figures differ from those obtained by Sala-i-Martin (2006) because he addresses the total world distribution of income whereas

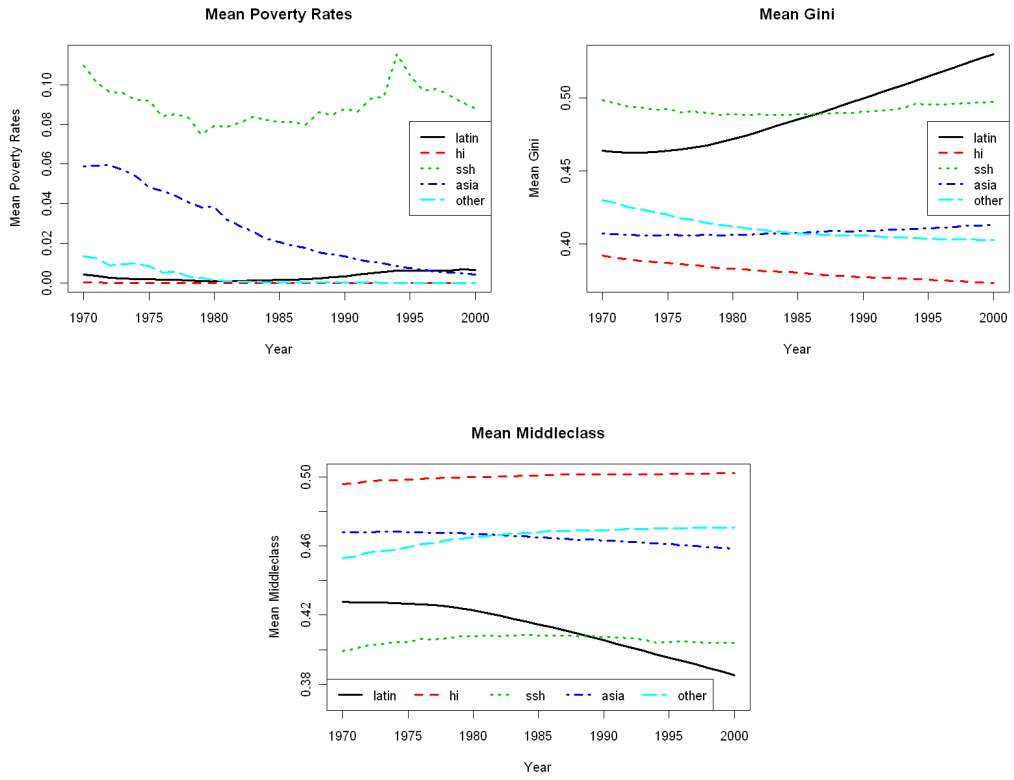

Figure 4.3: *The evolution of poverty, inequality and the middle class stratified for the groups of countries.*

we deal with country averages. This means that poor but populous countries like China or India are dramatically underweighted. Furthermore, Sala-i-Martin (2006)'s data set contains more countries. We observe that sub-Saharan African countries have the highest poverty rates and the poverty rates of Asian countries decrease incrementally. All other regions have much smaller poverty rates. While sub-Saharan African countries have high poverty rates and Gini coefficients, Latin American countries only suffer from an extreme and incrementally increasing inequality but not so much from extreme poverty. The fraction earned by the middle class of Latin American countries incrementally decreases. It is not surprising to see that the HI countries have the lowest inequality rates and the largest rates of middle class. Figure (4.4) shows that differences in inequality and the income earned by the middle class have a much larger impact on β_1 than differences in the poverty rate. The relationship of *gini* and *middleclass* to β_1 is very similar, namely almost linear and increasing. This forces β_1 in two opposing directions, as countries with large inequality usually have a small middle class and vice versa. Nevertheless, the fraction earned by the middle class does not fully determine the rate of inequality. Therefore, the increasing relationship of inequality and middle class to β_1 motivates again to differentiate

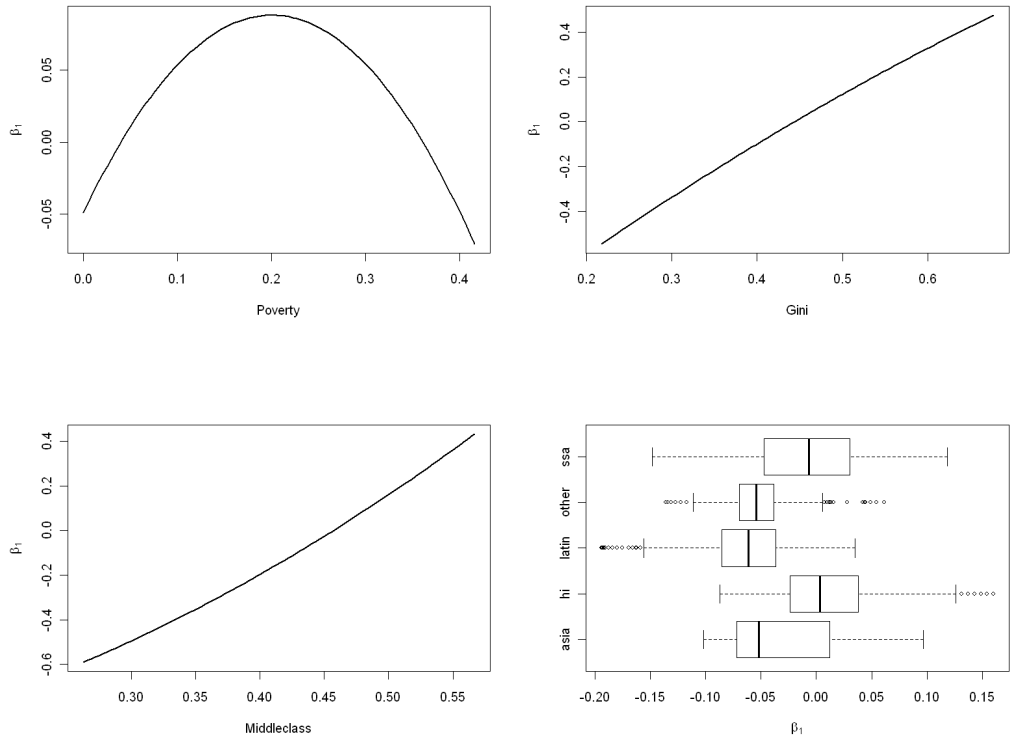

Figure 4.4: *The effects of poverty, inequality and the middle class on β_1 and the β_1's stratified for the groups of countries.*

the two variables. The poverty rate has a much smaller impact. We observe an inverted U-shaped relationship. The returns of *lnn* are theoretically the largest, when inequality is large even though the middle class earns a high fraction of the total income and a serious fraction of the total population (approximately 20 %) earns below the poverty line. The boxplots (bottom right) of figure (4.4) show how this transfers to the countries. The returns to *lnn* are especially large for sub-Saharan African and HI countries. According to the group of sub-Saharan African countries this result coincides with that of Koehler, Sperlich and Vortmeyer (2011). We also show that the β_1-coefficients of sub-Saharan African countries have larger variation than other countries. The coefficients of the groups Latin and Other are smaller on average and have less variation. It is interesting to see that not only for HI and sub-Saharan African but also for other countries the returns to *lnn* can be positive.

The estimated regression equation that explains the coefficient of *lnsk* is

$$\begin{aligned}\beta_{2it} = &-0.0519^{**} - 1.2799^{***} pov_{i(t-3)} + 3.7563^{***} pov^2_{i(t-3)} - 3.8873^{***} gini_{i(t-3)} \\ &+ 3.9307^{***} gini^2_{i(t-3)} + 5.6845^{***} middleclass_{i(t-3)} - 7.3372^{***} middleclass^2_{i(t-3)}.\end{aligned}$$

This is plotted in figure (4.5). It can be observed that poverty, inequality and the middle

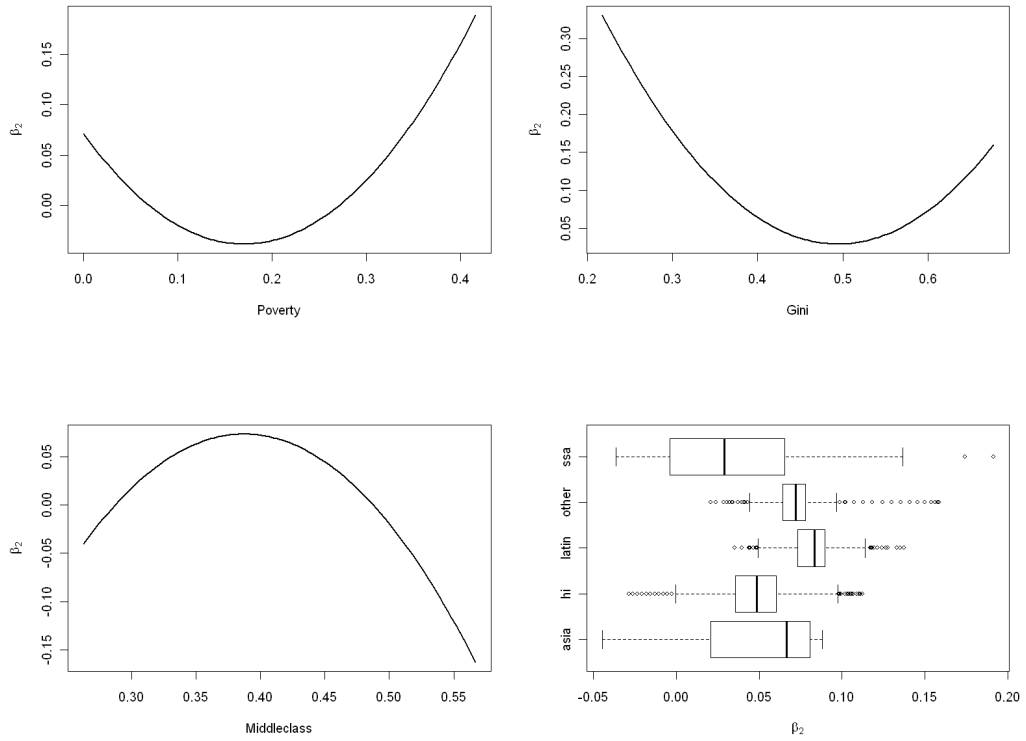

Figure 4.5: *The effects of poverty, inequality and the middle class on β_2 and the β_2's stratified for the groups of countries.*

class all have a remarkable impact on the coefficient of investment in physical capital. We observe a U-shaped relationship for the variables *pov* and *gini* and an inverted U-shaped relationship for the variable *middleclass*. Therefore, the returns to investments are the highest, when poverty and inequality are either extraordinary low or large and the middle class earns between 40 % and 45 % of the total income. Again, high inequality usually goes hand in hand with a low share earned by the middle class and vice versa. The boxplots of figure (4.5) show that sub-Saharan Africa which is characterized by large inequality, small middle class and large poverty has the smallest returns to physical capital on average. The coefficients are also subject to large variation. The coefficients of Asia show that the underlying distribution is skewed and that the coefficients are also small on average. The other groups have smaller variation and have larger coefficients on average. The largest returns to physical capital are observed for Latin American countries. These countries are characterized by small poverty rates but extreme inequality und a small middle class. The group of sub-Saharan African countries shows that the returns to investments in physical capital are small when extreme poverty goes hand in hand with extreme inequality.

The estimated regression equation for the coefficient of *lnattain* is given by

$$\begin{aligned}\beta_{3it} = &-2.8512^{***} + 0.6369^{***} pov_{i(t-3)} - 0.4078^{***} pov^2_{i(t-3)} + 0.5740^{***} gini_{i(t-3)} \\ &+ 2.2057^{***} gini^2_{i(t-3)} + 5.7917^{***} middleclass_{i(t-3)} - 2.3057^{***} middleclass^2_{i(t-3)}.\end{aligned}$$

First of all, the majority of the estimated coefficients of *lnattain* are negative. This is counterintuitive and indicates that years of schooling are not alone responsible for measuring human capital. For example, the quality of schooling is not incorporated in this variable. When estimating the growth regression equation without a variable-coefficients model, Islam (1995) and Koehler, Sperlich and Vortmeyer (2011) also observe a negative coefficient. Islam (1995) argues that the observed effect of human capital is either a measurement problem or relates to a misspecification of this variable by the Augmented Solow model. Koehler, Sperlich and Vortmeyer (2011) note that while school attainment incrementally increases for almost all countries, the growth rate fluctuates a lot and the final result is a negative coefficient. Pritchett (1996) argues that this result is robust, credible and provides three possible explanations. First he argues that schooling does not necessarily create human capital, second, the returns to education fall rapidly when the demand for educated labor is stagnant and third, a large amount of human capital is used for growth hindering activities, such as a bloated bureaucracy. Because of this counterintuitive result we are careful when interpreting the coefficients. Moreover, the boxplots in figure (4.6) show that the coefficients are not fundamentally different across the country groups. However, figure (4.6) shows that high inequality, high poverty and a large output of the variable *middleclass* cause a high return to school attainment. Differences in poverty have a much smaller impact on the coefficient than inequality and the share earned by the middle class.

The results clearly show that the poverty rate, the fraction of income earned by the middle class and inequality force the coefficients to differ a lot. This basically shows two things about estimating the usual growth regressions without varying coefficients, where the means of the overall coefficients are addressed. First of all, if the economies behave dramatically different according to these variables, the question arises how informative the mean of these coefficients actually is. The countries behave different and none of the observed country groups have coefficients like the overall mean. Therefore, the mean coefficients reflect a theoretical situation that is not fulfilled by any of the country groups. Second, when only estimating the mean coefficients of all countries, the regression is highly suspected to suffer from an endogeneity problem. The individual deviations from the means of the coefficients are very likely to move simultaneously with the level of growth. For example the existence of a negative and significant Africa-Dummy (see Koehler, Sperlich and Vortmeyer (2011)) shows that the poorest region with an extraordinary large inequality and small middle class has systematically smaller growth rates than all other countries. This indicates that the deviations from the means of the coefficients of the growth equation of this group of countries, which is summarized in the

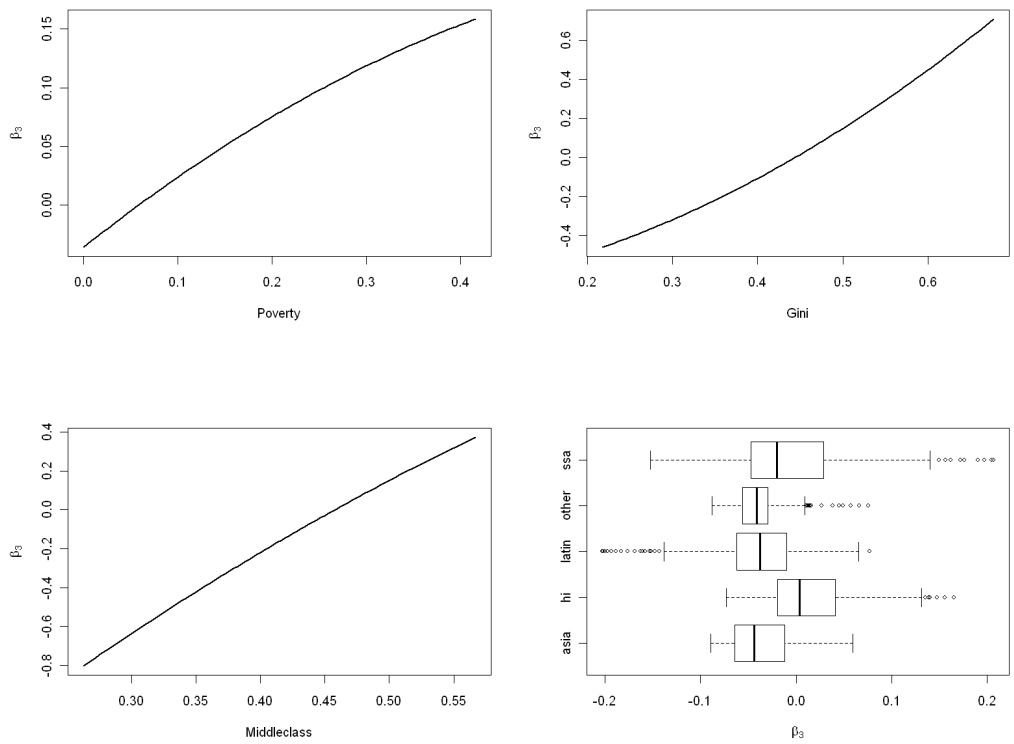

Figure 4.6: *The effects of poverty, inequality and the middle class on β_3 and the β_3's stratified for the groups of countries.*

coefficient of the Africa-Dummy, imply significantly lower growth.

4.3.2 The Effects on the Economic Growth of the Poor and the Rich

In subsection (4.3.1) we observed that measures of the income distribution affect the path of economic growth. According to measures of inequality, poverty and the share earned by the middle class, the economies behave in their specific way that might be totally different to the growth path of the theoretical mean of all countries. This motivates to investigate how the income distribution evolves. We investigate the growth path of the upper and the lower twenty per cent of the society. Differences in the growth path of the poor and the rich naturally affect measures of the income distribution which in turn affect growth as seen in subsection (4.3.1). As discussed in subsection (4.2.1) we collected data for the upper and lower twenty per cent of the society. These series consist of GDP's per capita, whereas the data used in (4.3.1) are GDP's per worker which is preferred when working with the model by Mankiw, Romer and Weil (1992). We do not have data for the GDP's per worker of the upper and lower twenty per cent and it would be distorting when transforming the per capita values into per worker values, as the information of being a potential worker is suspected to be highly correlated with the income. Therefore, we do not compare the results obtained by estimating the growth path of GDP per worker with those given in this subsection. Instead, we investigate how differences in the measures of the income distribution affect the behavior of the poor and that of the rich and detect similarities and differences. Therefore, we apply the model derived in subsection (4.2.3) with dependent variables $lny_{rich,it}$ for the investigation of the poorer twenty per cent and $lny_{poor,it}$ for the investigation of the richer twenty per cent.
The autoregressive coefficient for the poor is 0.9704 and that for the rich is 0.9366. This demonstrates that the series of the dependent variable is more persistent for the poor than for the rich.
The evolution of the coefficient of *lnn* of the poorer 20 per cent is

$$\begin{aligned}\beta_{1it,\,poor} = &-0.1806^{***} + 1.9801^{***} pov_{i(t-3)} - 5.0660^{***} pov^2_{i(t-3)} + 6.4187^{***} gini_{i(t-3)} \\ &- 5.5026^{***} gini^2_{i(t-3)} - 9.6406^{***} middleclass_{i(t-3)} + 13.4882^{***} middleclass^2_{i(t-3)}\end{aligned}$$

and that of the richer twenty per cent is given by

$$\begin{aligned}\beta_{1it,\,rich} = &-1.9793^{***} + 1.3614^{***} pov_{i(t-3)} - 3.2350^{***} pov^2_{i(t-3)} + 1.8338^{***} gini_{i(t-3)} \\ &- 0.1519^{***} gini^2_{i(t-3)} 2.4768^{***} middleclass_{i(t-3)} + 0.3237^{***} middleclass^2_{i(t-3)}.\end{aligned}$$

This is graphically demonstrated in figure (4.7). As in subsection (4.3.1), averaged variables are used for the variables held constant in each plot. First of all, it has to be noted that the main drivers for the coefficient of *lnn* are inequality and the share of income of the middle class. Differences in the poverty rate result in smaller differences of the β_1's. The

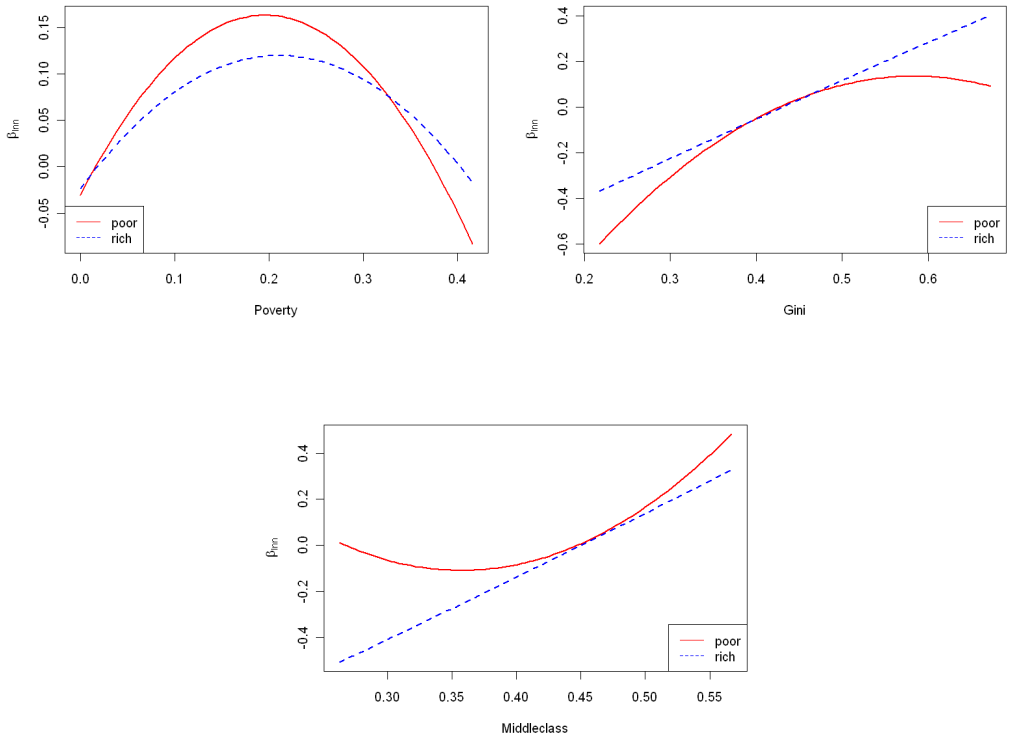

Figure 4.7: *The effects of poverty, inequality and the middle class on β_{lnn} of the poorest and richest twenty per cent*

function β_1 dependent on the rate of poverty follows an inverted U-shaped relationship for the poor and the rich. The relationship of β_1 to inequality and the share earned by the middle class for the rich is almost linear and increasing. Again, as large inequality usually goes hand in hand with a small share earned by the middle class, the increasing relationships for both inequality and the share earned by the middle class forces the coefficient in two opposite directions. Furthermore, it is interesting to observe that in case of large inequality the coefficient for the poor is smaller than that for the rich whereas in case of a small share of income earned by the middle class we observe a larger coefficient. The case of very small inequality and a large share earned by the middle class also forces the coefficient of *lnn* in opposing directions. Small inequality results in a smaller β_1 for the poor than for the rich and a large share earned by the middle class results in a larger β_1. The evolution of the coefficient of *lnsk* for the poorer 20 per cent of the population is

$$\begin{aligned}\beta_{2it} = &-3.9012^{***} - 1.3569^{***} pov_{i(t-3)} + 3.2247^{***} pov^2_{i(t-3)} - 9.5966^{***} gini_{i(t-3)} \\ &+ 12.1551^{***} gini^2_{i(t-3)} + 25.0670^{***} middleclass_{i(t-3)} - 26.9079^{***} middleclass^2_{i(t-3)}\end{aligned}$$

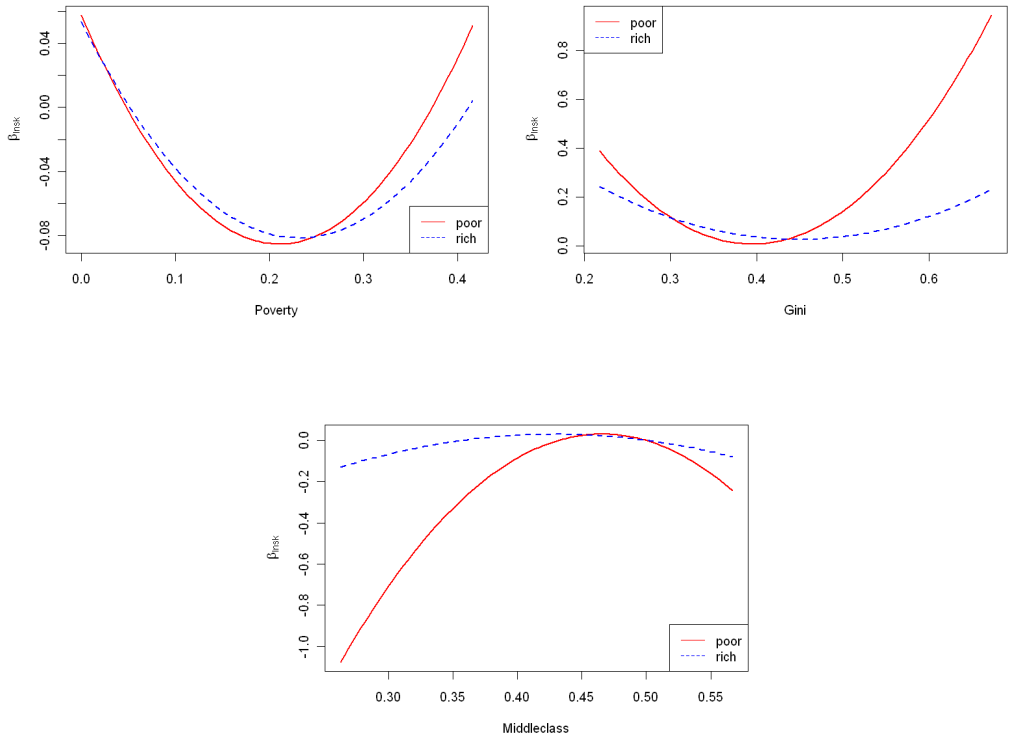

Figure 4.8: *The effects of poverty, inequality and the middle class on β_{lnsk} of the poorest and richest twenty per cent*

and for the richer twenty per cent is

$$\begin{aligned}\beta_{2it} = &-0.1896^{***} - 1.1640^{***} pov_{i(t-3)} + 2.5135^{***} pov^2_{i(t-3)} - 3.6437^{***} gini_{i(t-3)} \\ &+ 4.0661^{***} gini^2_{i(t-3)} + 4.9471^{***} middleclass_{i(t-3)} - 5.7624^{***} middleclass^2_{i(t-3)}.\end{aligned}$$

This is graphically demonstrated in figure (4.8). Differences in poverty result in smaller differences of β_2 than differences in inequality or middle class. It can clearly be seen that the rich profit more from investments in physical capital if they do not share with the middle class. On the other hand, in case of very large inequality, the return to investment for the poor is much larger than that for the rich. The evolution of the coefficient of *lnattain* for the poorer 20 per cent of the population is

$$\begin{aligned}\beta_{3it} = &-3.0506^{***} + 1.3524^{***} pov_{i(t-3)} - 3.9498^{***} pov^2_{i(t-3)} + 0.6753^{***} gini_{i(t-3)} \\ &- 2.1525^{***} gini^2_{i(t-3)} 6.1614^{***} middleclass_{i(t-3)} - 2.6327^{***} middleclass^2_{i(t-3)}\end{aligned}$$

and for the richer twenty per cent is given by

$$\begin{aligned}\beta_{3it} = &-2.8011^{***} + 0.4760^{***} pov_{i(t-3)} - 1.8265^{***} pov^2_{i(t-3)} - 1.3797^{***} gini_{i(t-3)} \\ &+ 4.2849^{***} gini^2_{i(t-3)} + 7.9466^{***} middleclass_{i(t-3)} - 4.9629^{***} middleclass^2_{i(t-3)}.\end{aligned}$$

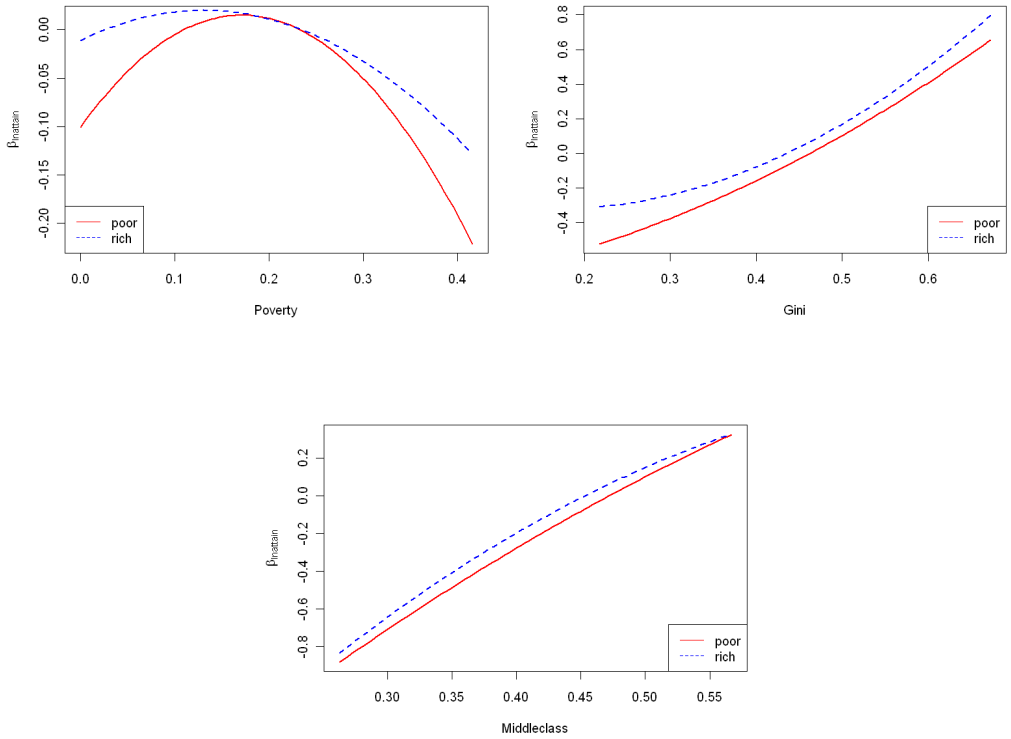

Figure 4.9: *The effects of poverty, inequality and the middle class on $\beta_{lnattain}$ of the poorest and richest twenty per cent*

This is graphically demonstrated in figure (4.9). As observed when investigating the growth of the per worker GDP, we observe that the coefficients are likely to be negative. Because of this counterintuitive result, we have to be careful when interpreting the results. What we do observe is that differences in poverty result in much smaller differences of β_3 than differences in inequality and the share earned by the middle class. The impact of inequality and middle class on β_3 is similar for the poor and the rich. Investments in human capital are slightly better for the rich in case an extremely small or extremely large poverty rate.

The significance of the coefficients of the regressions for the poorer and the richer twenty per cent shows that the measures of the income distribution affect the growth path of the poorer twenty per cent in a different way than that of the richer twenty per cent. Note that differences in the growth path of the poorer and the richer twenty per cent partially determine measures of the income distribution. Furthermore, as shown in subsection (4.3.1), measures of the income distribution affect the behavior of the economies which is demonstrated by the differences in the returns of the growth drivers.

4.4 Conclusion

Collecting long time-series of yearly frequency for a large range of countries allows applying a variable-coefficients model, in which differences to the mean coefficient are explained by poverty, inequality and the share of income earned by the middle class. The results show that the coefficients are highly different and this can be explained by the level of poverty, inequality and the share of income earned by the middle class.
There are several reasons for estimating with variable coefficients. First, adding measures of the income distribution to the set of explanatory variables of the growth regression alone, does not model the reason for adding these variables, namely that the poor behave different than the rich. Furthermore, we loose economic justification when adding a lot of variables to the growth regression. Second, the mean of the coefficients is not an informative parameter of the growth equation because of the dramatically different outputs of the different subgroups of the coefficients. Third, the differences of the coefficients to their means are highly suspicious to move simultaneously with growth, which indicates an endogeneity problem. Fourth, as poor countries have weaker databases and are therefore more likely to be excluded from the data, the difference in the coefficients indicates a sample selection bias when estimating mean coefficients.
Outstanding results are that sub-Saharan African countries have highly varying and large returns to population growth and highly varying and small returns to physical capital. Latin American countries experience highly negative returns to population growth and large positive returns to physical capital. The high income countries also have large returns to population growth and positive returns to physical capital. All country groups experience negative returns to school attainment, which indicates once again that the variable does not take important information, such as quality of schooling into account. When expressing the coefficients as functions of poverty, inequality and the share of income earned by the middle class, we observe that poverty has much smaller effects on the coefficients than inequality and the share earned by the middle class. Large inequality usually goes hand in hand with a small share earned by the middle class and vice versa. The fact that this tends to move the coefficients in opposite directions demonstrates the importance of incorporating both variables.
We also investigate the growth path of the poorer and the richer twenty per cent of the society. First of all, we observe that the returns of the growth regression are highly dependent on poverty, inequality and the share earned by the middle class. Furthermore, the returns of the two subgroups of the total population are impacted in different ways. Outstanding results are that in case of extremely high and extremely small inequality, the return to population growth is smaller for the poor than for the rich, whereas in case of an extremely small share earned by the middle class or an extremely large share earned by the middle class, it is larger. Furthermore, in case of extremely large inequality the return to

investment in physical capital is larger for the poor and in case of an extremely small share earned by the middle class it is smaller. This shows again that small inequality and a large share earned by the middle class or large inequality and a small share earned by the middle class tend to force the coefficients in different directions. The differences of the returns dependent on poverty for the poor and the rich are small. The aforementioned differences of the poor and the rich naturally affect the parameters of the income distribution, which in turn affect the growth path of the GDP per worker. This undermines again the importance of considering the income distribution when modeling the growth path.

Chapter 5

Conclusion

The thesis consists of three papers, whose main conclusions are given separately in the end of each chapter. The task of this chapter is to give a critical examination of the findings of the thesis and to reflect the author's personal view.

The first paper is concerned with bandwidth selection in nonparametric kernel regression. Given the need of automatic data-driven bandwidth selectors for applied statistics, several bandwidth selection methods have been introduced in kernel regression. They differ quite a bit, and although there already exist more selection methods than for any other regression smoother we can still see coming up new ones. In fact, the discussion about estimating the optimal bandwidth has been going on for the last four decades. One could say that this discussion is a never-ending story. The rising number of methods makes the practioner experience a rising complexity. Nowadays, the practioner not really faces to the problem of choosing the optimal bandwidth for the data set at hand; instead, the problem is rather to choose the method that chooses the bandwidth. This gives the need for an exhaustive report as presented in chapter (2). To the best of our knowledge, we are the first providing such a comprehensive review and comparison study for bandwidth selection methods in the kernel regression context. By this means, this essay not only contributes to the ongoing discussion but also provides a basis for further discussion about optimal bandwidth choice. However, an optimal choice that works best for all data sets we looked at could not be given. Practitioners with some experience in nonparametric applications choose the bandwidth using implemented standard routines or just by eye judgment. This method is quite promising and in most cases adequate, especially for explorative statistics.

The second and third papers are concerned with growth regressions. The eye-catching thing about growth regressions is that they are always conducted in a different way. First, there is no consensus about which explanatory variables really drive growth. The range of topics with their corresponding explanatory variables is so large, that growth can be seen as a theory of everything. In an extreme case, the unique output of a large number of explanatory variables only indentifies individual countries and therefore behaves like a

dummy variable itself. Thereby, it is not clear whether these variables really drive growth. This demands an economic model that regressions can be based on. We always justify our regressions by the well known augmented Solow Model. Second, the range of statistical methods used to estimate the coefficients of growth regressions is extremely large. Many authors artificially shorten their time-series so that being technically able to apply the highly popular and well-known GMM methods. Needless to say, that this is highly critical. Some authors estimate with fixed effects, others apply a random effects estimator. All these methods produce different results, giving a lot of space for economic interpretation. This gives the need for an extensive discussion about the estimation method. Chapter (3) as well as (4) entail such a discussion about which estimator works best for the corresponding data set.

Chapter (3) is concerned with the Africa-Dummy in growth regressions. On the one hand, a prominent stylized fact about economic growth is that when comparing two otherwise similar countries, the one with the lower initial mean income will tend to see the higher rate of growth. On the other hand, the growth performance of sub-Saharan African countries is significantly worse compared to that of all other countries. The task of chapter (3) is not to add more explanatory variables to the growth regression until the sub-group of sub-Saharan African countries is fully identified and the Africa-Dummy disappears. Instead, we find it necessary to derive statistical facts about sub-Saharan Africa's growth punishment. We develop a statistical method, that is able to identify the Africa-Dummy and that can moreover be extended to derive empirical facts about it. Thereby, we show how the Africa-Dummy interacts with the other explanatory variables, to what extend the parametric linear structure of growth regressions is responsible for its appearance and how it evolves over time.

While the Africa-Dummy is a correction of the intercept having to be made for the special sub-group of sub-Saharan African countries, chapter (4) deals with differences in the other coefficients. Without concentrating on the Africa-Dummy alone, we develop and apply a variable-coefficients model where the coefficients are explained by the country's individual level of poverty, inequality and the share earned by its middle class. Thereby, we show that the country's coefficients differ a lot. The model allows dealing with widely criticized drawbacks of growth regressions. First, it rules out the problem of a possible sample selection bias, resulting from the fact that poor countries have weaker data bases and are therefore more likely to be excluded from the data set. Second, it is more informative than a model estimating the mean coefficients. Third, it deals with the problem that differences to the mean coefficients move simultaneously with the dependent variable indicating an endogeneity problem. Fourth, it gives an example of how extra variables can be added to the growth regression without extending the set of explanatory variables proposed by the underlying growth model that justifies the regression.

Bibliography

ACEMOGLU, D. AND JOHNSON, S. AND ROBINSON, J. A. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *The American Economic Review* **91**(5) 1369–1401.

ACEMOGLU, D. AND ZILIBOTTI, F. (1997). Was Prometheus Unbound by Chance? Risk, Diversification, and Growth. *Journal of Political Economy* **105**(4) 709–751.

ADELMAN, I. AND MORRIS, C. T. (1969). Society, Politics, and Economic Development: A Quantitative Approach. *The Economic Journal* **79**(313) 160–163.

AGHION, P. AND BOLTON, P. (1997). A Theory of Trickle-Down Growth and Development. *Review of Economic Studies* **64**(2) 151–172.

AKAIKE, H. (1970). Statistical Predictor Information. *Annals of the Institute of Statistical Mathematics* **22** 203–217.

AKAIKE, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* **19**(6) 716–723.

ALESINA, A. AND RODRIK, D. (1994). Distributive Politics and Economic Growth. *Quarterly Journal of Economics* **109**(2) 465–490.

ALESINA, A. (1994). Voting for Reform: Democracy, Political Liberalization, and Economic Adjustment: Democracy, Liberalization and Economic Adjustment (Chapter 2). *Oxford University Press*

ALESINA, A. AND PEROTTI, R. (1996). Income Distribution, Political Instability, and Investment. *European Economic Review* **40**(6) 1203–1228.

AMEMIYA, T. (1978). A Note on a Random Coefficients Model. *International Economic Review* **19**(3) 793–796.

ARCAND, J. L. AND GUILLAUMONT, P. AND JEANNENEY, S. G. (2000). How to make a Tragedy: On the Alleged Effect of Ethnicity on Growth. *Journal of International Development* **12**(7) 925–938.

ARELLANO, M.(2003) Modelling Optimal Instrumental Variables for Dynamic Panel Data Models. CEMFI, Madrid.

ARELLANO, M. AND BOND, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies* **58**(2) 277–297.

ARELLANO, M. AND BOVER, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics* **69** 29–51.

AZARIADIS, C. (1996). The Economics of Poverty Traps. Part One: Complete Markets. *Journal of Economic Growth* **1** 449–486.

BANDYOPADHYAY, D, AND TANG, X. (2011). Understanding the Economic Dynamics Behind Growth and Inequality Relationships. *Journal of Macroeconomics* **33** 14–32.

BANERJEE, A. AND DUFLO, E. (2003). Inequality and Growth: What Can the Data Say? *Journal of Economic Growth* **8**(3) 267-299.

BANERJEE, A. V. AND NEWMAN, A. F. (1993). Occupational Choice and the Process of Development. *Journal of Political Economy* **101**(2) 274–298.

BAROSSI-FILHO, M. AND GONCALVES SILVA, R. AND MARTINS DINIZ, E. (2005). The Empirics of the Solow Growth Model: Long-Term Evidence. *Journal of Applied Economics* **8**(1) 31–51.

BARRO, R. J. (1991). Economic Growth in a Cross Section of Countries. *The Quarterly Journal of Economics* **106**(2) 407–443.

BARRO, R. J. AND LEE, J. W. (2010). A New Data Set pf Educational Attainment in the World, 1950 − 2010. National Bureau of Economic Research, Working Paper 15902.

BARRO, R. J. AND MANKIW, N. G. AND SALA-I-MARTIN, X. (1995). Capital Mobility in Neoclassical Models of Growth. *American Economic Review* **85**(1) 103–15.

BAXTER, M. AND KING, R. G. (1999). Measuring Business Cycles: Approximate Band-Pass Filters for Economic time-series. *The Review of Economics and Statistics* **81**(4) 575–593.

BENABOU, R. (1996). Heterogeneity, Stratification, and Growth: Macroeconomic Implications of Community Structure and School Finance. *American Economic Review* **86**(3) 584-609.

BIRDSALL, N. AND GRAHAM, C. AND PETTINATO, S. (2000). Stuck in the Tunnel: Is Globalization Muddling the Middle Class? *Center on Social and Economic Dynamics Working Paper 14*.

BLUNDELL, R. AND BOND, S. (1998). Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics* **87**(1) 115–143.

BOND, S. AND HOEFFLER, A. AND TEMPLE, J.(2001). GMM Estimation of Empirical Growth Models. Discussion Paper No 01/525

BOURGUIGNON, F. (1998). Équité Et Croissance Économique: une nouvelle analyse? *Revue Française D'économie* **13**(3) 25–84.

BOURGUIGNON, F. (2004). The Poverty-Growth-Inequality Triangle. *Indian Council for Research on International Economic Relations, New Delhi Working Papers*

BOWSHER, C.G. (2002). On testing Overidentifying Restrictions in Dynamic Panel Data Models. *Economics Letters* **77** 211–220.

BOWLES, S. AND DURLAUF, S. N. AND HOFF, K. (2006). Poverty Traps. *Princeton University Press*

CAO-ABAD, R. (1991). Rate of Convergence for the Wild Bootstrap in Nonparametric Regression. *The Annals of Statistics* **19**(4) 2226–2231.

CAO-ABAD, R. AND GONZÁLEZ-MANTEIGA, W. (1993). Bootstrap Methods in Regression Smoothing. *Nonparametric Statistics* **2**(4) 379–388.

CASELLI, F. AND ESQUIVEL, G. AND LEFORT, F. (1996). Reopening the Convergence Debate: A New Look at Cross-Country Growth Empirics. *Journal of Economic Growth* **1**(3) 363–389.

CASELLI, F. (2005). Accounting for Cross-Country Income Differences. *Handbook of Economic Growth*, Chapter 9, 679-741.

CHENG, M.-J. AND FAN, J. AND MARRON, J.S. (1997). On automatic boundary corrections. *The Annals of Statistics* **25**(4) 1691–1708.

CHIU, S.-T. (1990). On the Asymptotic Distributions of Bandwith Estimates. *The Annals of Statistics* **18** 1696–1711.

CLARK, R. M. (1977). Non-Parametric Estimation of a Smooth Regression Function. *Journal of the Royal Statistical Society, Series B* **39**(1) 107–113.

COLLIER, P. AND GUNNING, J. W. (1999). Explaining African Economic Performance. *Journal of Economic Literature* **37**(1) 64–111.

CRAVEN, P. AND WAHBA, G. (1979). Smoothing Noisy Data With Spline Functions. *Numerische Mathematik* **31** 377–403.

CUNHA, F. AND HECKMAN, J. (2007). The Technology of Skill Formation. *American Economic Review* **97**(2) 31-47.

DOEPKE, M. AND ZILIBOTTI, F. (2005). Social Class and the Spirit of Capitalism. *Journal of The European Economic Association* **3**(2-3) 516-524.

EASTERLY, W. (2001). The Middle Class Consensus and Economic Development. *Journal of Economic Growth* **6**(4) 317–335.

EASTERLY, W. AND LEVINE, R. (1997). Africa's Growth Tragedy: Policies and Ethnic Divisions. *The Quarterly Journal of Economics* **112**(4) 1203–50.

DEININGER, K. AND SQUIRE, L. (1996). A New Data Set Measuring Income Inequality. *The World Bank Economic Review* **10**(3) 565–591.

FAN, J. (1992). Design-Adaptive Nonparametric Regression. *Journal of American Statistical Association* **87**(420) 998–1004.

FAN, J. AND GIJBELS, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics* **20**(4) 2008–2036.

FORBES, K. J. (2000). A Reassessment of the Relationship between Inequality and Growth. *The American Economic Review* **90**(4) 869–887.

GALOR, O. AND ZEIRA, J. (1993). Income Distribution and Macroeconomics. *The Review of Economic Studies* **60**(1) 35–52.

GALOR, O. AND TSIDDON, D. (1997). The Distribution of Human Capital and Economic Growth. *Journal of Economic Growth* **2**(1) 93–124.

GALOR, O. AND TSIDDON, D. (1997). Technological Progress, Mobility, and Economic Growth. *American Economic Review* **87**(3) 363–382.

GASSER, T. AND KNEIP, A. AND KÖHLER, W. (1991). A Fast and Flexible Method for Automatic Smoothing. *Journal of the American Statistical Association* **86** 643–652.

GASSER, T., MÜLLER H.G. (1979). Kernel Estimation of Regression Functions. *Lecture Notes in Mathematics 757, eds. T. Gasser and M. Rosenblatt, Heidelberg: Springer-Verlag* **87**(420) 998–1004.

GEORGESCU-ROEGEN, N. (1975). Dynamic Models and Economic Growth. *World Development* **11-12**(3) 765–783.

GO, D. AND NIKITIN, D. AND WANG, X. AND ZOU, H. (2007). Poverty and Inequality in Sub-Saharan Africa: Literature Survey and Empirical Assessment. *Annals Of Economics And Finance* **8**(2) 251–304.

GOLDFARB, D. AND IDNANI, A. (1982). Dual and Primal-Dual Methods for Solving Strictly Convex Quadratic Programs. *Numerical Analysis* 226–239.

GOLDFARB, D. AND IDNANI, A. (1983). A Numerically Stable Dual Method For Solving Strictly Convex Quadratic Programs. *Mathematical Programming* **27**(1) 1–33.

GONZÁLEZ-MANTEIGA, W., MARTÍNEZ MIRANDA, M.D. AND PÉREZ GONZÁLEZ, A. (2004). The choice of smoothing parameter in nonparametric regression through Wild Bootstrap. *Computational Statistics & Data Analysis* **47** 487–515.

HAHN, J. AND KUERSTEINER, G. (2002). Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects When Both n and T Are Large. *Econometrica* **70**(4) 1639–1657.

HALL, P. (1990). Using the bootstrap to estimate mean square error and select smoothing parameters in nonparametric problems. *Journal of Multivariate Analysis* **32** 177–203.

HALL P., MARRON J.S. AND PARK B.U. (1992). Smoothed cross-validation. *Probabability Theory and Related Fields* **92** 1–20.

HANSEN, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* **50**(4) 1029–1054.

HÄRDLE, W. (1992). Applied Nonparametric Regression. Cambridge University Press.

HÄRDLE, W. AND BOWMAN, A.W. (1988). Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands. *Journal of the American Statistical Association* **83**(401) 102–110.

HÄRDLE, W., HALL, P., MARRON, J.S. (1988). How far are automatically chosen Smoothing Parameters from their Optimum. *Journal of American Statistical Association* **83**(401) 86–95.

HÄRDLE, W., MARRON, J.S. (1985). Optimal Bandwidth Selection in Nonparametric Regression Function Estimation. *The Annals of Statistics* **13**(4) 1465–1481.

HÄRDLE, W., MARRON, J.S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics* **19**(2) 778–796.

HART, J. D. AND YI, S. (1998). One-Sided Cross Validation, *Journal of American Statistical Association* **93**(442) 620–631.

HAUSMAN, J. A. AND TAYLOR, W. E. (1981). Panel Data and Unobservable Individual Effects. *Econometrica* **49**(6) 1377–1398.

HEIDENREICH, N.B. AND SCHINDLER, A. AND SPERLICH, S.(2010). Bandwidth Selection Methods for Kernel Density Estimation - A Review of Performance, SSRN-id1726428.

HESTON, A. AND SUMMERS, R. AND ATEN, B. (2009). Penn World Table Version 6.3. *Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania*.

HODRICK, R. J. AND PRESCOTT, E. C. (1997). Postwar U.S. Business Cycles: An Empirical Investigation. *Journal of Money, Credit and Banking* **29**(1) 1–16.

HOEFFLER, A.E. (2002). The augmented Solow model and the African growth debate. *Oxford Bulletin of Economics and Statistics* **64**(2) 135–158.

HURVICH, C. M., SIMONOFF, J.S. AND TSAI C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *Journal of the Royal Statistical Society, Series B* **60**(2) 271–293.

ISLAM, N. (1995). Growth Empirics: A Panel Data Approach. *The Quarterly Journal of Economics* **110**(4) 1127–70.

JALAN, J. AND RAVALLION, M. (2004). Insurance Against Poverty (Chapter 5). *Oxford University Press*.

KALDOR, N. (1956). Alternative Theories of Distribution. *The Review of Economic Studies* **23**(2) 83–100.

KAO, C. AND CHIANG, M. H. (2000). On The Estimation And Inference Of A Cointegrated Regression In Panel Data. *Advances in Econometrics* **15** 179–222.

KENDALL, M.G. (1954). Note on the Bias in the Estimation of Autocorrelations. *Biometrika* **41** 403–404.

KIVIET, J.F. (1995). On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models. *Journal of Econometrics* **68** 53–78.

MAX KÖHLER, M. AND SPERLICH, S. AND VORTMEYER, J.(2011). The Africa-Dummy in Growth Regressions. Courant Research Centre "Poverty, Equity and Growth in Developing and Transition Countries: Statistical Methods and Empirical Analysis", Discussion Paper No 94.

KRAAY, A. (2006). When is Growth Pro-Poor? Evidence from a Panel of Countries. *Journal of Development Economics* **80** 198–227.

KURITA, K. AND KUROSAKI, T. (2011). The Dynamics of Growth, Poverty and Inequality: A Panel Analysis of Regional Data from Thailand and the Philippines. *Asian Economic Journal* **25**(1) 3–33.

KUZNETS, S (1955). Economic Growth and Income Inequality. *The American Economic Review* **45**(1) 1–28.

LANDES, D. (1999). The Wealth and Poverty of Nations: Why Some Are So Rich and Some So Poor. *Norton* **8**(3).

LEVINE, R. AND RENELT, D. (1992). A Sensitivity Analysis of Cross-Country Growth Regressions. *The American Economic Review* **82**(4) 942–63.

LOKSHIN, M. AND RAVALLION, M. (2004). Household Income Dynamics in Two Transition Economies. *Studies in Nonlinear Dynamics and Econometrics* **8**(3).

LOPEZ, H. AND SERVÉN, L. (2009). Too Poor to Grow. *Policy Research Working Paper 5012*.

MALLOWS, C.L. (1973). Some Comments on C_p. *Technometrics* **15**(4) 661–675.

MAMMEN, M., MARTINEZ-MIRANDA, M.D., NIELSEN, J.P. AND SPERLICH, S. (2011). Do-validation for Kernel Density Estimation. *Journal of the American Statistical Association* **106**(494) 651–660.

MANKIW, N. G. AND ROMER, D. AND WEIL, D. N. (1992). A Contribution to the Empirics of Economic Growth. *The Quarterly Journal of Economics* **107**(2) 407–437.

MARRON, J.S. (1986). Will the Art of Smoothing ever become a Science. Function Estimates, Contemporary Mathematics 59, Providence, RI: American Mathematical Society, 169–178.

MCKENZIE, D.J. AND WOODRUFF, C. (2006). Do Entry Costs Provide an Empirical Basis for Poverty Traps? Evidence from Mexican Microenterprises. *Economic Development and Cultural Change* **55**(1) 3–42.

MESNARD, A. AND RAVALLION, M. (2006). The Wealth Effect on New Business Startups in a Developing Economy. *Economica* **73** 367–392.

MURPHY, K. M. AND SCHLEIFER, A. AND VISHNY, R. W. (1989). Industrialization and the Big Push. *Journal of Political Economy* **97**(5) 1003-1026.

NADARAYA, E.A. (1964). On Estimating Regression. *Theory of Probability and its Application* **9** 141–142.

NICKELL, S. (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica* **49**(6) 1417–26.

ORCUTT, G. H. AND IRWIN, J. O. (1948). A Study of the Autoregressive Nature of the time series Used for Tinbergen's Model of the Economic System of the United States. *Journal of Royal Statistical Society* **10**(1) 1–53.

PARK, B.U. AND MARRON, J.S. (1990). Comparison of Data-Driven Bandwidth Selectors. *Journal of the American Statistical Association* **85**(409) 66–72.

PIKETTY, T. (1993). Imperfect Capital Markets and the Persistence of Initial Wealth Inequalities. *London School of Economics Suntory Toyota Centre for Economics and Related Disciplines Working Paper* **TE**(92) 255.

PERSSON, T. AND TABELLINE, G. (1990). Politico-Economic Equilibrium Growth: Theory and Evidence. *mimeo*

PERSSON, T. AND TABELLINI, G. (1994). Is Inequality Harmful for Growth? *American Economic Review* **84**(3) 600–621.

PERRY, G.E. AND LOPEZ, J.H. AND MALONEY, W.F. (2006). Poverty Reduction and Growth: Virtuous and Vicious Circles. *The World Bank*.

PHILLIPS, P. C. B. AND SUL, D. (2007). Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence. *Journal of Econometrics* **137** 162–188.

PHILLIPS, P. C. B. AND MOON, H. (1999). Linear Regression Limit Theory for Nonstationary Panel Data. *Econometrica* **67** 1057–1111.

PRICE, G. N. (2003). Economic Growth in a Cross-section of Nonindustrial Countries: Does Colonial Heritage Matter for Africa? *Review of Development Economics* **7**(3) 478–495.

PRIESTLEY, M. B. AND CHAO, M.T. (1972). Non-parametric function fitting. *Journal of the Royal Statistical Society, Series B* **34**(3) 385–392.

PRITCHETT, L. (1996). Where Has All the Education Gone. The World Bank, Policy Research Working Paper 1581.

RAVALLION, M. (2010). Why Dont We See Poverty Convergence? *Development Research Group, World Bank, Working Paper* **4974**.

RAVALLION, M. AND CHEN, S. (20003). Measuring Pro-Poor Growth *Economic Letters* **78**(1) 93–99.

RICE, J. (1984). Bandwith Choice for Nonparametric Regression. *The Annals of Statistics* **12**(4) 1215–1230.

RODRIK, D. (1998). Where Did All the Growth Go? External Shocks, Social Conflict and Growth Collapses. *Centre for Economic Policy Research Discussion Paper* **1789**.

ROODMAN, D.(2006). How to Do xtabond2: An Introduction to Difference and System GMM in Stata. Center for Global Development, Working Paper Number 103.

ROODMAN, D. (2009). A Note on the Theme of Too Many Instruments. *Oxford Bulletin of Economics and Statistics* **71**(1) 135–158.

RUPPERT, D., SHEATHER, S.J. AND WAND, M.P. (1995). An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of the American Statistical Association* **90**(432) 1257–1270.

RUPPERT, D. AND WAND, M.P. (1994). Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics* **22**(3) 1346–1370.

SACHS, J. D. AND MCARTHUR, J. W. AND SCHMIDT-TRAUB, G. AND KRUK, M. (2004). Ending Africa's Poverty Trap. *Brookings Papers on Economic Activity* **2004**(1) 117–216.

SACHS, J. D. AND WARNER, A. M. (1997). Sources of Slow Growth in African Economies. *Journal of African Economies* **6**(3) 335–76.

SAINT-PAUL, G. AND VERDIER, T. (1993). Education, Democracy, and Growth. *Journal of Development Economics* **42**(2) 399–407.

SALA-I-MARTIN, X. (2006). The World Distribution of Income: Falling Poverty and Convergence, Period. *Quarterly Journal of Economics* **121**(2) 351–397.

SHEATHER, S.J. AND JONES, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B* **53**(3) 683–690.

SHIBATA, R. (1981). An Optimal Selection of Regression Variables. *Biometrika* **68**(1) 45–54.

SINGH, B. AND NAGAR, A. L. AND CHOUDHRY, N. K. AND BALDEV,R. (1993). On the Estimation of Structural Change: A Generalization of the Random Coefficients Regression Model. *International Economic Review* **17**(2) 340–361.

SOLOW, R. M. (1956). A Contribution to the Theory of Economic Growth. *The Quarterly Journal of Economics* **70**(1) 65–94.

SRIDHARAN, E. (2004). The Growth and Sectoral Composition of India's Middle Class: Its Impact on the Politics of Economic Liberalization. *India Review* **3**(4) 405–428.

SWAN, T. W. (1956). Economic Growth and Capital Accumulation. *Economic Record* **32**(1) 334–361.

TAUCHEN, G. (1986). Statistical Properties of Generalized Method-of-Moments Estimators of Structural Parameters Obtained from Financial Market Data. *Journal of Business and Economic Statistics* **4**(4) 397–416.

WACHTER, M. L. (1970). Relative Wage Equations for U.S. Manufacturing Industries 1947-1967. *The Review of Economics and Statistics* **52**(4) 405–410.

WATSON, G.S. (1964). Smooth Regression Analysis. *Sankhyā, Series A* **26** 359–372.

WERE, M. AND NAFULA, N. N. (2003). An Assessment of the Impact of HIV/AIDS on Economic Growth: The Case of Kenya. CESifo Working Paper Series 1034.

WINDMEIJER, F. (2005). A Finite Sample Correction for the Variance of Linear Efficient Two-Step GMM Estimators. *Journal of Econometrics* **126**(1) 25–51.

XIA, B. (2010). Status, Inequality and Intertemporal Choice. *The B.E. Journal of Theoretical Economics* **10**(1).

YANG, L. AND TSCHERNIG, R. (1999). Multivariate bandwidth selection for local linear regression. *Journal of the Royal Statistical Society, Series B* **61**(4) 793–815.

YI, S. (2001). Asymptotic Stability of the OSCV Smoothing Parameter Selection. *Communications in Statistics - Theory and Methods* **30**(10) 2033–2044.

ZILIAK, J. P. (1997). Efficient Estimation with Panel Data when Instruments Are Predetermined: An Empirical Comparison of Moment-Condition Estimators. *Journal of Business and Economic Statistics* **15**(4) 419–431.

Curriculum Vitae

Contact Information

Name:	Max Köhler
Adress:	Georg-August Universität Göttingen, CRC Poverty, Equity and Growth, Wilhelm-Weber-Str. 2, 37073 Göttingen
Phone:	+49 (0551) 39 4794
E-Mail:	Max.Koehler@wiwi.uni-goettingen.de

Personal Information

Date of Birth:	11th Febuary 1983
Place of Birth:	Hamburg
Citizenship:	German

Employment History

since 10/2008	CRC Poverty, Equity and Growth

Education

10/2003 – 09/2005	Universität Dortmund, Vordiplom in Wirtschaftsmathematik
10/2005 – 09/2008	Julius-Maximilians-Universität Würzburg, Diplom in Wirtschaftsmathematik
since 10/2008	Georg–August–Universität Göttingen: Doctoral studies

Publications and Working Paper

- *The Africa-Dummy in Growth Regressions* (2011) – Courant Research Centre: Poverty, Equity and Growth - Discussion Papers (with S. Sperlich and J. Vortmeyer)
- *A Review and Comparison of Bandwidth Selection Methods for Kernel Regression* (2011) – Courant Research Centre: Poverty, Equity and Growth - Discussion Papers (with A. Schindler and S. Sperlich)
- *A Review and Comparison of Bandwidth Selection Methods for Kernel Regression* (2011) – submitted in International Statistical Review (with A. Schindler and S. Sperlich)

Göttingen, 6.2.2012

Eidesstattliche Erklärung

Ich versichere an Eides statt, dass ich die eingereichte Dissertation

ECONOMETRIC STUDIES ON FLEXIBLE MODELING OF DEVELOPING COUNTRIES IN GROWTH ANALYSIS

selbstständig verfasst habe. Anderer als der von mir angegebenen Hilfsmittel und Schriften habe ich mich nicht bedient. Alle wörtlich oder sinngemäß den Schriften anderer Autorinnen und/oder Autoren entnommenen Stellen habe ich kenntlich gemacht.

Göttingen, 6.2.2012