Ludwig-Maximilians-Universität München / Institut für Statistik

University of Munich / Department of Statistics

Education in Developing Countries in the Framework of the Millennium Development Goals:

The Issue of Missing Values and Methods of Statistical Imputation

Bachelor Thesis

Supervision: Prof. Dr. Thomas Augustin

Author: Sebastian Steinmüller

28 July 2010

Abstract

Data analysis is often complicated by missing values. Ad-hoc methods for handling missing data like listwise deletion or imputation of the mean frequently lead to efficiency loss or bias in the parameters of the subsequent analysis. To overcome the limitations of these techniques, more elaborate methods like multiple imputation combined with the expectation maximisation algorithm are increasingly applied to data sets containing missing values. Multiple imputation denotes the approach of finding more than one value for each missing data point to express the uncertainty introduced by the imputation. The expectation maximisation algorithm is a method which helps to find the appropriate posterior distribution to draw multiple imputations from. This thesis aims to apply these methods to the country-level data set on the Millennium Development Goals from the UN and to give an overview of missing values in the context of developing aid as well as in general. Various methods for handling missing data will be compared and an outlook on more sophisticated techniques will be given.

Keywords: Missing Values, Imputation Methods, Expectation Maximisation Algorithm, Multiple Imputation, Buck's Method, Millennium Development Goals, Primary Education

Contents

1 Introduction

The Millennium Development Goals (MDGs), a series of obligations accepted by virtually every national government in the world to raise the living standard of their inhabitants, offer abundant opportunities to carry out statistical analysis thanks to the data set created to measure them. Unfortunately, it is subject to missing values which might seriously affect the empirical analysis. Researchers facing missing values often resort to standard methods for handling it, in many cases because statistical software uses them as its default approach. This includes for example the use of only complete cases, which may dramatically reduce the number of available cases and sometimes make the analysis infeasible. However, there is extensive literature on the question how to handle this problem. For instance, King et al. (2001) and Gartner and Scheid (2003) treat the issue of missing values in developing countries in particular and propose likelihood-based approaches to find substitutes for the missing values. Older techniques for handling missing values include imputing unconditional means or finding substitutes by means of a linear regression on available variables. King et al. (2001) present the easy-to-use R-package *Amelia* (Honaker et al., 2007) for the imputation of missing values based on the so-called expectation maximisation (EM) algorithm combined with multiple imputation (MI).

The main goal of this thesis is to apply theoretical background knowledge on missing values and methods of imputation to a selected part of the MDGs - data set and at the same time to provide some insight into the issue of missing values in developing countries. This includes a general overview of the MDGs in the second chapter with a special focus on their statistical indicators. It extends to a section on primary education in developing countries as an own part of the MDGs, followed by a first descriptive summary of the data set. The fourth chapter begins with an introduction of missing values in developing countries and continues with the theory of the descriptive analysis of missingness, which is needed for the subsequent analysis of missingness in the

MDGs-data. The exclusively theoretical chapter 5 is meant to present several methods for dealing with missing values, in particular listwise deletion, omitting variables, imputation by the unconditional mean, Buck's method and, as a more recently developed technique, multiple imputation combined with the expectation maximisation algorithm. The sixth part applies these methods to a modified version of the MDGs - data set and compares their impact on a linear regression analysis. Finally, the thesis provides an outlook on possible improvements for multiple imputation in combination with the EM-algorithm.

2 The Millennium Development Goals

> *We will spare no effort to free our fellow men, women and children from the abject and dehumanizing conditions of extreme poverty, to which more than a billion of them are currently subjected. We are committed to making the right to development a reality for everyone and to freeing the entire human race from want.*
>
> United Nations Millennium Summit (2000, Article 11)

In the “UN Millennium Declaration”, 189 countries promised to engage in the struggle against poverty and its dire consequences at the beginning of the new millennium. To specify these good intentions, the participants of the UN Millennium Summit, hold in September 2000, set eight overarching goals, the MDGs:

1. Eradicate extreme hunger and poverty
2. Achieve universal primary education
3. Promote gender equality and empower women
4. Reduce child mortality
5. Improve maternal health
6. Combat HIV/AIDS, malaria, and other diseases
7. Ensure environmental sustainability
8. Develop a global partnership for development

With the assignment of targets to each of the eight goals and of statistical indicators to every target, the development of the MDGs can be measured in each country. For example, Goal 2, which is to achieve universal primary education, is specified by Target 2A: “Ensure that, by 2015, children everywhere, boys and girls alike, will be able to complete a full course of primary schooling”. This target is measured by the indicators “Enrolment in primary education”, “Completion of primary education” and “Literacy of 15-24 year olds, female and male”[1].

The MDGs are primarily designed to improve the living conditions of people in the so called least developed countries (LDCs) and low income countries (LICs). Most of the poorest people in these countries are caught in a “poverty trap” (Sachs, 2005, p. 56), as well as their countries. That is, they do not possess the means to invest in basic and absolutely necessary goods like water supply, transportation, education and health. The MDGs are also meant to provide the sometimes ineffective and not transparent system of international development aid with a new structure. Special consultant of the UN-Secretary and former director of the Millennium Project Jeffrey Sachs expresses that “the MDGs state real goals that provide not only benchmarks for aid but also milestones for the advice of the international agencies as well.” (Sachs, 2005, p. 82). The MDGs and the subsequent “Declaration of Paris” (OECD, 2008) are based on the idea of development aid as a process of mutual responsibility and transparency to build long-lasting structures in developing countries and of measuring this process constantly. To achieve this goal, both measurable indicators and a statistical infrastructure to collect, assess and statistically evaluate the needed data are necessary. This statistical infrastructure is obviously not always the first concern of countries which even lack means to fund schools and hospitals. Consequently, such countries often fail to obtain and process data needed to assess, for example, the health state of the population, the economic situation or the infrastructure in the country. Countries that are especially in need of external aid at the same time often lack the means to identify these needs. A basic level of

[1] For general information about the MDGs, all indicators and further reading:
`http://www.un.org/millenniumgoals/`

information however, is essential to identify necessary investments. National administration bodies, as well as international agencies like the UN, often therefore have to resort to estimating the missing data based on information they already have about the quantity of interest. This includes data for the same quantity from previous years and neighbouring and comparable countries as well as merely statistical methods taking into consideration the correlation with other variables. The methods for filling in ("imputing") substitutes for missing values are the subject of intensive debate, since the results of any subsequent statistical analysis can be biased by it.

3 Primary education in the framework of the MDGs

> *Everyone has the right to education. Education shall be free, at least in the elementary and fundamental stages. Elementary education shall be compulsory.*
>
> United Nations (1948, Article 26)

Universal primary education is considered to be a vital factor in improving the living conditions of people in developing countries. The UN speak of education as "the vehicle through which societies reproduce themselves" (UN Millennium Project, 2005, p. 23). The call for universal primary education is consequently included in the MDGs through Goal 2, Target 2A: "Ensure that, by the date [2015], children everywhere, boys and girls alike, will be able to complete a full course of primary schooling" (United Nations Millennium Summit, 2000, Article 3). Universal primary education is essential for a society to stand on their own feet and asserts an enormous array of positive social and economic influences. The following chapter gives an overview of the importance of primary education in the framework of the MDGs as well as an introduction to the factors that influence it. Subsequently, this theoretical background will be used to narrow down the data set on the MDGs to variables relating to primary education and to give a short descriptive summary of the data.

3.1 Impacts on and by primary education in developing countries

The children of girls and women with good education are less likely to die until the age of five (UN Millennium Project, 2005, chap. 2), and education of girls and women leads to decreasing birth rates, which again are associated with a higher household income. Quality education can create awareness for health topics like HIV/AIDS and thereby contribute to a better health state of the population. Education, as long as it is equally accessible for girls and boys, can improve the status of girls and women in society in two ways: by helping them to gain access to better paid jobs, thereby making them more independent, and by directly teaching them about their own rights and chances. Education about agriculture and food can help to contribute to a better nutritional situation. And last but not least, education provides people with better chances to access the labour market and ultimately secure a higher income. For the effects of education on these and other social and economic variables, see for example UN Millennium Project (2005), Sachs et al. (2004).

On the other hand, primary education itself is influenced by a broad field of socio-economic and medical variables which have to be taken into consideration to improve primary education. In Sachs et al. (2004), the authors list several variables influencing primary education: Interventions for more efficient agricultural methods reduce the time children have to work in the fields, thus enabling them to access education on a more regular and sustained basis. Healthy and sufficient nutrition is a basic need for cognitive functioning and improved learning success. Maternal education contributes to higher student enrolment rates. Prevention and treatment of HIV/AIDS improve health and attendance of teachers and students and helps to reduce the number of orphans, who are less likely to complete school because of their expected commitment to care for their remaining family and to make their own living. In general, a sound health system will increase attendance and abilities of teachers and students, as do improved sanitary facilities. In many developing countries, women and children are responsible for the transportation of water from remote sources. Thus, the provision of physically accessible water supplies is essential to increase children's attendance rates at school. Enhanced transport infrastructure and services also assist in reducing the time teachers and children have to walk to school. Furthermore the provision of secondary and tertiary education is important for the education of teachers preparing to enter the workforce. Access to electricity enables children to study for longer periods during the night, and modern fuels improve respiratory indoor air. This ultimately results in an improvement of children's health and frees them from time needed to collect traditional fuel.

3.2 Descriptive analysis of the MDGs - data set

The MDGs data set is a time series cross-section for 234 countries and administrative regions over the time period from 1990 to 2009[2]. It contains data for 152 variables, including all the indicators of the MDGs. Focus will be made on the so called developing countries according to the latest list of recipients of official development assistance (ODA), which is in effect for ODA-flows in 2009 and 2010 (see OECD, 2009). The list is issued in intervals of 3 years by the development assistance committee (DAC), a sub-organisation of the Organisation for Economic Co-operation and Development (OECD), which is responsible for co-ordinating international development aid. The countries on the list are divided into four groups according to their gross national income (GNI) per capita. The groups are labeled "Least Developed Countries" (LDCs), "Other Low Income Countries" (LICs), "Lower Middle Income Countries" and "Upper Middle Income Countries", where the upper threshold of the last group and thus for the definition of a developing country are 11455 US-Dollar GNI/capita per year. To be classified as a LDC, a country has to be under the threshold for LICs of 935 US-Dollar GNI/capita and at the

[2] available under `http://mdgs.un.org/unsd/mdg/Handlers/ExportHandler.ashx?Type=Csv`

same time meet certain criteria like particular economic vulnerability and a lack of social resources, e.g. education and health services. We will analyse the assumed influence of several factors on the primary completion rate as described in the previous section by means of a linear regression model, trying to improve the results using various methods of dealing with missing values in the data. Small island states and low-lying coastal areas with a population of less than 5 000 000 will be excluded from the data set, as their social and economic conditions tend to vary vastly from those of other states and this could possibly have biasing effects (Sachs et al., 2004, United Nations Department of Economic and Social Affairs, 2010). Our reduced data set consists of a time series over the years from 1990 to 2008 for 103 countries and 152 variables, resulting in 1957 cases. As already mentioned, many of these values are missing and there is only data for specific years for most of the variables.

There are three indicators and one more variable in the MDGs data set to measure Goal 2, universal primary education for all. Each of them is available for women, men and for both sexes. For all definitions of variables and indicators, see *Millennium Development Goals: Metadata* (2010). There is also an overview of all the MDGs-indicators and variables in printable format issued by the United Nations Development Group (2003). The total net enrolment ratio in primary education is defined as the number of children of official primary school age who are enroled in primary education as a percentage of the total children of the official school age population. It might obscure high dropout rates in primary education in some countries, since it does not take into account if a course of primary education has been completed successfully. The second official indicator is the literacy rate of the 15-24 year olds. The quantity of interest could be denoted by “quality of primary education in a country” and is best measured by the primary completion rate. This is an additional variable of the MDGs-data set and is defined as “the total number of new entrants in the last grade of primary education (according to the International Standard Classification of Education or ISCED97), regardless of age, expressed as percentage of the total population of the theoretical entrance age to the last grade of primary” (*Millennium Development Goals: Meta-*

data, 2010). ISCED97 defines primary education as "programmes normally designed on a unit or project basis to give pupils a sound basic education in reading, writing and mathematics along with an elementary understanding of other subjects such as history, geography, natural science, social science, art and music". The indicator "Percentage of pupils starting grade 1 who reach last grade of primary" would be suitable for measuring the quality of primary education as well, but it has a much higher proportion of missing values than the primary completion rate. There is a whole array of variables in the data set to measure the factors influencing primary education. The variable "Percentage of children under 5 severely underweight" can be used to express the quality and sufficiency of nutrition, while the children under five mortality rate per 1,000 live births and the maternal mortality ratio per 100,000 live births stand for the quality of health care in a country. The gender parity index in tertiary level enrolment (GPI) is the ratio of the number of female students enroled at tertiary level of education to the number of male students and represents the state of gender equality in education. The proportion of the population using improved drinking water sources, the proportion of the population using improved sanitation facilities and the percentage of 15-49 year old people living with HIV will be included as variables as well as the tuberculosis prevalence rate per 100,000 population. The population using solid fuels is the percentage of the population that relies on solid fuels as the primary source of domestic energy for cooking and heating. Those fuels such as wood and dung that are time consuming to collect have negative consequences for indoor air and may contribute to respiratory infections (Sachs et al., 2004, p. 207). The GDP per capita at current prices in US-Dollars is not part of the data set on the MDGs, but will be included using data from the UN as well [3]. It is one of the crucial indicators for the economic performance of a country and therefore possibly influential upon the public funds available for education. The percentage of the population earning below 1 US-Dollar (PPP) per day is a variable used to measure the extent of extreme poverty within a country. The quantity "Internet users per 100

[3] http://data.un.org/Data.aspx?q=gdp+capita&d=SNAAMA&f=grID%3a101%3bcurrID%3aUSD%3bpcFlag%3a1

population" will be included in the analysis to measure the effects of technological progress on primary education (Sachs et al., 2004, p. 203). Table 1 summarises those variables and shows the proportion of missing values in the data set for each of them as well as their correlation with the primary completion rate.
Pearson's correlation coefficients of the variables with the primary completion rate in the last column of table 1 all have the expected signs. The values for "Percentage of the population below $1 (PPP) per day" (-0.767), "Children under five mortality rate per 1000 live births" (-0.837), "Maternal mortality ratio per 100000 live births" (-0.807) and "Proportion of the population using improved sanitation facilities" (0.752) are particularly high. The only really small correlation can be found for the variable "Percentage of people living with HIV, 15-49 years old" (-0.173), this variable should perhaps be excluded from the analysis. The abbreviations in the second column will be used instead of the full variables' names.

Compared to most of the other MDGs, the situation in the field of primary education has improved in the last couple of years. For instance, the net enrolment ratio increased worldwide between the years 2000 and 2007, especially in the key regions of Southern Asia by 15 % and in Sub-Saharan Africa by 11 % (United Nations, 2009, p. 14). Despite the positive trend, 72 million children in primary school age were out of school as of 2007, with almost half of them living in Sub-Saharan Africa. Approximately half of them have never visited a school and are unlikely to ever do so, particularly in Western Asia and Sub-Saharan Africa (United Nations, 2009, p. 15). According to the UN, the number of 72 million children in 2007 who were then out of school can only be reduced to 26 millions until 2015, thus failing to achieve the goal of universal primary education (United Nations, 2009, p. 15). The UN cite economic and gender differences as another obstacle to the aspiration of primary education for all, with children from disadvantaged groups of society being held back from school attendance, for example by school fees.

	Abbr.	N	Mean	Sd	% missing values	Correlation with Prim. Compl
Population below 1$ PPP per day percentage	Dollar.Pov.	329	25.13	24.18	0.83	−0.77
Children under 5 severely underweight percentage	Underweight	313	5.42	5.05	0.84	−0.71
Primary completion rate both sexes	Prim.Compl	713	73.69	25.75	0.64	1.00
Gender Parity Index in tertiary level enrolment, percentage	GPI3	593	86.21	41.53	0.70	0.69
Children under five mortality rate per 1000 live births	CM	514	91.09	66.22	0.74	−0.84
Maternal mortality ratio per 100000 live births	MM	102	475.80	455.36	0.95	−0.81
People living with HIV, 15-49 year olds, percentage	HIV	182	3.02	5.68	0.91	−0.17
Tuberculosis prevalence rate per 100000 population	TBC	1839	287.21	224.82	0.06	−0.57
Population using solid fuels, percentage	Fuels	172	54.59	35.25	0.91	−0.73
Proportion of the population using improved drinking water sources, total	Water	378	73.25	20.27	0.81	0.72
Proportion of the population using improved sanitation facilities, total	Sanitation	370	52.88	29.05	0.81	0.75
Internet users per 100 population	Internet	1280	3.40	6.49	0.35	0.46
Per capita GDP at current prices - US dollars / 100	GDP	1919	16.23	21.07	0.02	0.47

Table 1: Overview of the relevant variables

4 Missing values

4.1 The issue of missing values in developing countries

The problem of missingness in data sets on country-level data can be illustrated by Figure 1 that plots the number of all observed values over all the variables of the data frame on the MDGs in 2006 against the logarithm of the Gross Domestic Product in purchasing power parities (as one possible measure for the wealth of a country) in each country in 2006.

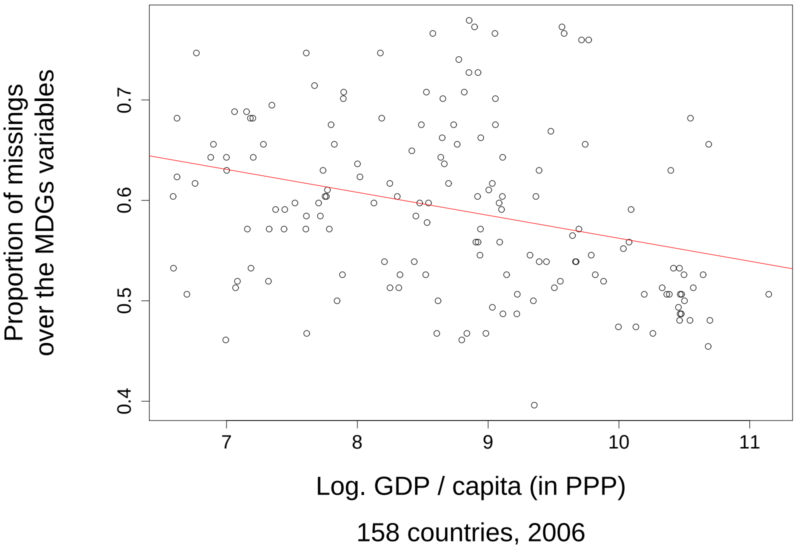

Figure 1: Correlation of wealth and the proportion of missingness

Apparently, data collection tends to be better the wealthier a country is. At first glance, the influence is rather small and non-linear. This could be partly due to the fact that monetary wealth does not account for all of the variation in data availability. Other factors could be influential as well, for example the general secondary and tertiary education level in a country, since the process of data collection and evaluation requires a basic level of education. However, the biggest uncertainty in this correlation is created by the effect of the missing values itself: A country with a high proportion of missing values is also more likely to lack the entry for the GDP. Consequently, countries for which only a small quantity of the data exist, tend to be underrepresented in this graph although they would possibly have a major influence on the correlation. Other factors possibly influencing data availability include diseases like HIV / AIDS and Malaria, rampant especially in parts of sub-Saharan Africa. They not only have devastating effects on health, life expectation and family situation of the people, but also account for high levels of invalidity and early deaths among the working population. This occurs in business as well as in public services, thereby often impeding any self-contained economic and social progress as described in Sachs (2005, p. 193). This probably also holds true to a similar extent for statistical institutes and administration in charge of data collection and evaluation, since they mostly employ highly skilled academics who are difficult to substitute. Figure 2 underlines the correlation of HIV-prevalence rate among the 15-49 year olds and the proportion of missing data in the data frame on the MDGs in 2007.

Just like Figure 1 for the GDP, this graph does not necessarily stand for a causal influence of the HIV-prevalence on the proportion of missing values, since there are possible confounding variables such as the economic situation. Furthermore, the analysis could again be affected by missingness itself. There seems to be nevertheless an indication for a correlation between the two variables in the sense that countries with a higher HIV-prevalence rate tend to perform worse in data collection. This correlation would be even stronger if extremely high values for the HIV-prevalence were excluded from the analysis. The great remoteness of many dwellings in rural areas of developing countries is just another challenge in data collection. The rapidly growing Internet and mobile communication markets could provide a solution for this problem, since they begin to reach even the poorest in the least developed countries (see Figure 3). These technologies could help to transmit information in real time from even remote areas. For example, only one computer

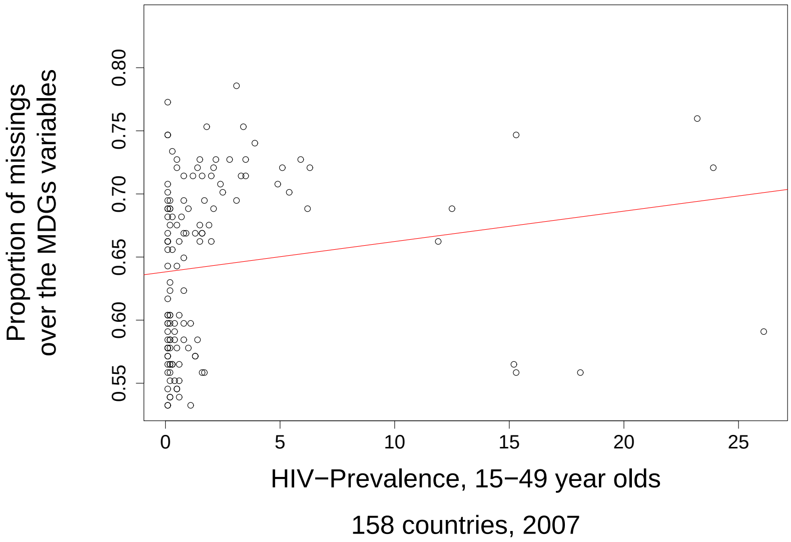

Figure 2: Correlation of HIV-Prevalence and the proportion of missingness

with Internet access or even one mobile phone could be sufficient to inform authorities or agencies about a shortage of mosquito nets in a village, as a first necessary step to solve the problem.

A linear regression model of the proportion of missing values in 158 countries in 2006 on three explanatory variables can be used to quantify the possible effects on the extent of missingness in the data set (see Table 2). 62 of the originally 158 cases have been deleted because of missingness in one of the explanatory variables. As it will be apparent later, the parameters of the performed regression analysis can be assumed to be biased, as the proportion of missing values itself will be influential on the observation status (missing or observed) of the covariates. As it is obviously more likely for countries with a higher proportion of missings to have one or more of the covariates missing, the more extreme values of the dependent variable "proportion of missing values" will probably be underrepresented in the model. Thus, the obtained estimates for regression parameters and the p-values should rather

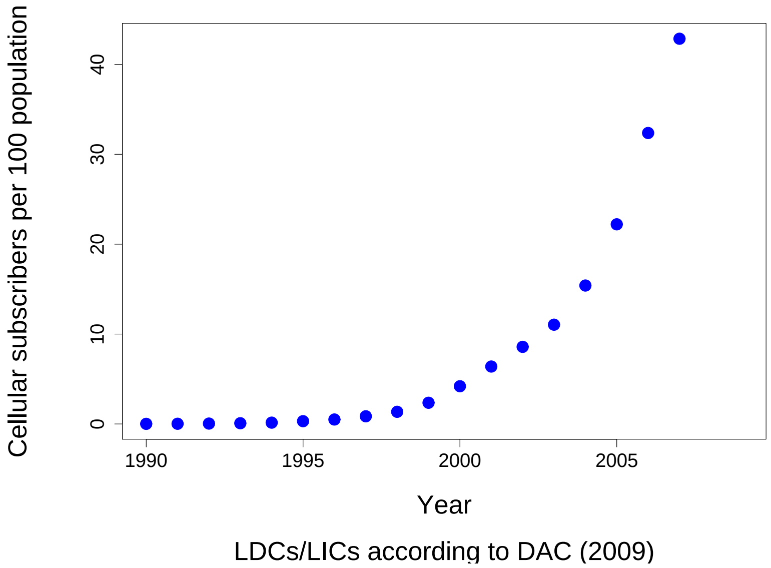

Figure 3: Cellular subscribers in developing countries

Dependent variable: Proportion of missing values in % in the entire MDGs-variables in 2006

	β	p
Intercept	61.055	2.57e-26
BIP/capita (PPP) / 100	-0.019	0.000318
HIV-Prevalence	0.296	0.0237
Enrolment ratio in primary education	-0.04	0.403
N=96 R^2 =0.27		

Table 2: Possible explanatory variables for missingness in developing countries

be taken as a possible indication for factors influencing the proportion of missings than as actual results of a linear model. Nevertheless, the estimates do have the expected signs.
Problems arise especially when variables of interest used in analysis are at the same time presumably influential upon the pattern of missingness in the data, if one exists. Imagine for example an analysis of the effects of the GDP per capita of a country on its tuberculosis-prevalence rate. As we suppose countries with low GDPs to have less scientific and statistical resources and consequently higher levels of missingness in their national data, countries with low GDPs will possibly be underrepresented in the analysis.
Missingness is certainly not the only mechanism compromising data quality. Errors of measurement and especially systematic manipulation of particular quantities towards desirable results are equally serious issues for aggregated data on country level. The latter is in particular due to the fact that the main source for data on country level are national agencies and ministries, which will in many cases be tempted to manipulate measured values in their own interest. This does not only hold for isolated nations like North Korea or Myanmar, but also for democratic governments. Greece's manipulation of data relating to the economic performance of the country, which was uncovered in 2010, is only one example for the unreliability of official sources. However, systematic manipulation and errors of measurement will not be taken into further consideration in this thesis. The data is assumed to be measured without errors or manipulation, although we should keep in mind that this is probably not always the case in reality.

4.2 Theoretical overview of mechanisms leading to missing data

There is a formal description of different mechanisms leading to missing data in Spieß (2008) and Toutenburg et al. (2004) which was originally introduced by Rubin (1976). The most favourable meachanism for any subsequent analysis are values which are missing completely at random (MCAR). That means that the distribution of the missing values in the data set is not influenced by any factor, whether it be a variable of the data of interest itself or any other quantity. MCAR-values can be ignored in analysis, since they do not contain information which could bias results and the observed values are a random sample of the complete, but partly unobserved data. The missingness of a value which is missing at random (MAR) depends on a variable which is not the variable of the missing value itself. Thus, the term "missing at random" can be misleading, since it suggests that the missingness is not influenced by any factors. These values are ignorable if the variable of interest for the analysis is not correlated with the variable influencing the distribution of missingness. The expression "missing not at random" (MNAR) denotes a mechanism leading to missingness which is dependent upon the variable in which missingness is observed itself. These values are non-ignorable in the sense that they will bias the results of an analysis conducted with only the observed values and that the observed values cannot be considered a random sample of the complete, but partly unobserved data. Let us imagine a simple descriptive analysis of the GDP/capita across various country with unobserved values for some of the countries in question. Since there are some indications that the GDP/capita could be influential upon the state of statistical expertise in a country in the sense that poorer countries are more likely to have missing values in aggregated data, the arithmetic mean of the variable using the observed values is possibly biased upwards. The following notation simplifies the one from Spieß (2008, pp. 5 ff.) and might help to understand the described mechanisms leading to missingness. Be Z_{obs} the observed and Z_{mis} the unobserved part of the complete data Z and let R be an additional variable:

$$R = \begin{cases} 1 & \text{, value is observed} \\ 0 & \text{, value is missing} \end{cases}$$

The different mechanisms leading to missingness can then be expressed as follows:

1. MNAR:$P(R|Z_{obs}, Z_{mis}) = P(R)$

2. MAR: $P(R|Z_{obs}, Z_{mis}) = P(R|Z_{obs})$

3. MNAR: $P(R|Z_{obs}, Z_{mis}) = P(R|Z_{obs}, Z_{mis})$

The distinction between MAR- and MCAR-data can be made by looking at the distribution of the variable R, indicating missingness in a particular variable Y, conditioned on the values of additional variables X. If the distribution of R changes significantly over the values of X, this is an indicator for MAR- instead of MCAR-data. However, it is not possible to distinguish MAR or MCAR-data from MNAR-values (see Spieß, 2008, p. 13). This can be illustrated by simple deliberation: Since MNAR is defined as a missingness mechanism R which is conditional on the values of the variable Y containing missing values, the only way of detecting it would be to look at the distribution of R conditioned on Y. This distribution however is observed only incompletely, so that it is not possible to formally find a MNAR-mechanism using only information from the sample. There are various methods to deal with missing values, including the use of external data, for example from NGOs, or the improvement of statistical infrastructure and knowledge in a country on the long run. However, imputation of substitutes for the missing values as a last resort is often necessary for data analysis. We will now run an analysis of the distribution of missing values in the MDGs-data.

4.3 The distribution of missing values in the MDGs - data set

The plot *aggr()* from the R-package *VIM* (Templ and Alfons, 2009) allows to display the percentage of missing values for a variable analogue to Table 1 as in Figure 4. Furthermore, the right part of the graph shows the frequencies of the different combinations of missing values across the variables, where a red cell stands for a missing value and a blue one for an observation. Each row represents a certain combination of missing values across the variables of the reduced data set. The additional graph on the far right displays the frequencies of each combination, with the most frequent combinations at the bottom.

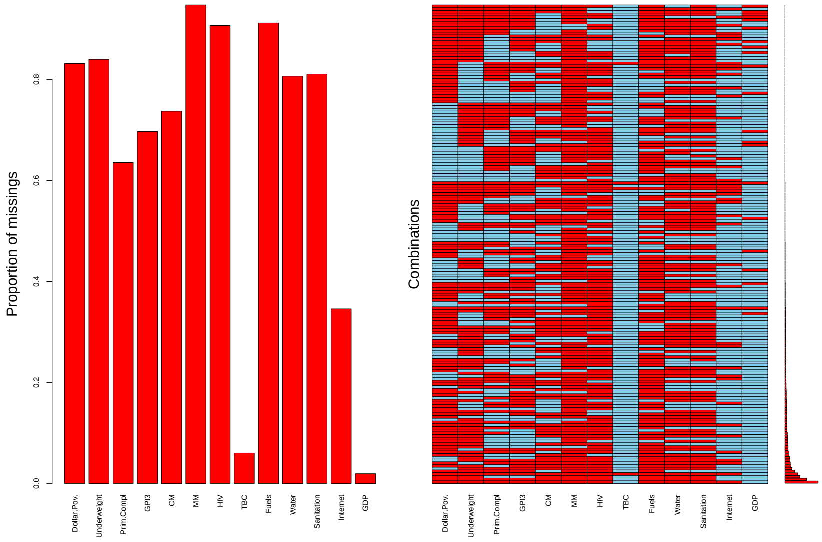

Figure 4: Missingness across the MDGs-variables

Only TBC, the tuberculosis prevalence rate, and GDP have a very small proportion of missing values as displayed in the left section of the graph, whereas all the other variables have an exceedingly high percentage of missingness, mostly above 50 %. This becomes an issue in the right section, where not a single row is entirely blue. Thus, there is no case for which all the variables are observed simultaneously, which makes a complete-case analysis (also named listwise deletion analysis) as the default approach of statistical software like *R-Project* impossible. The next two plots display the proportion of missingness in each variable over time, where every panel represents the fluctuation of missingness over the years from 1990 to 2009.

There are three main patterns of missingness over time in the variables. Measurements for the variables GDP/capita and the tuberculosis prevalence rate are available for almost every country regardless of the year, and their proportion of missingness does not vary substantially over time. The largest part of the variables shows a fluctuating proportion of missing values over time at a high level of 40 % to 90 % missingness. The third group consists in the variables child mortality, HIV-prevalence and fuels, which are observed

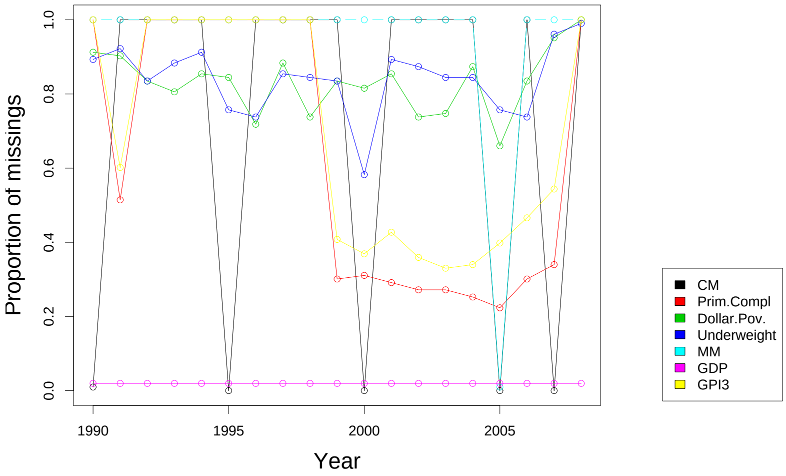

Figure 5: Proportion of missings per year for each variable

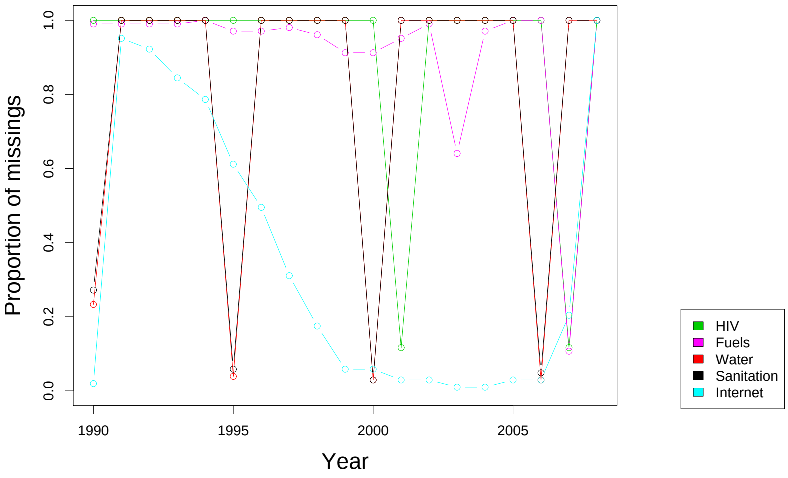

Figure 6: Proportion of missings per year for each variable

completely for some years (i.e. the proportion of missing values is 0) and entirely missing for others. This is not due to political, social or economic factors, but because the UN did not deem necessary an annual measurement of these variables. They are thus not missing in the original sense as they were never intended to be observed. Consequently, it does not seem to make much sense to try to impute substitutional values for them, which in turn causes serious problems for our analysis: Since the resulting time series for those variables will remain entirely incomplete for some years, it will be much more difficult to carry out the imputation and analysis steps taking into consideration the time series character of the entire data set. Analysing the data only for particular years would be an option to overcome this problem. However, the basic idea of the MDGs is to bring about progress over the years, or in a more neutral and statistical sense, to be a time series measuring change. Thus, we will use a modified data set using only the differences between the arithmetic mean of the years from 2000 to 2007 as the first subtrahend and the mean of the years from 1990 to 1999 as the second subtrahend of a variable in a country. The data set thus declines to 103 cases, each consisting of the described difference of a country listed in the ODA-list for 2009 and 2010 (OECD, 2009), not including small island developing states with less than 5000000 inhabitants. This serves as an attempt to maintain at least a part of the time-series character of the data without losing an excessive proportion of the values due to missingness. Table 3 summarises this data set in the same way as table 1 for the original data.

Interpreting the means of the variables as the respective changes between the decade of the 1990s and the years from 2000 to 2007, we can see a positive development in all of them. For instance, the mortality of children under 5 years old decreased by -22.56 deaths per 1000 children averaged over all the 103 countries, and the primary completion rate rose by almost 7 %. The variables HIV and MM (maternal mortality) did not have enough values for computing both the arithmetic mean for the years before 2000 and from this year onwards in any country. Their proportion of missingness is 1.00 and consequently these will be excluded from further analysis. This can be assumed not to bias the analysis in the case of maternal mortality, as it will probably bear collinearity with the remaining variable child mortality. The other variables now have reasonable missingness rates ranging from 0 % to 32 %, with the exception of “Fuels” for which over 77 % of the values are missing. The correlation coefficients of “Internet” and GDP (in 100 US-Dollars) with the primary completion rate do not have the expected signs, as

	Abbr.	N	Mean	Sd	% missing values	Correlation with Prim. Compl
Population below 1$ PPP per day percentage	Dollar.Pov.	72	−5.23	9.29	0.30	−0.05
Children under 5 severely underweight percentage	Underweight	71	−1.20	2.57	0.31	−0.23
Primary completion rate both sexes	Prim.Compl	82	6.89	8.32	0.20	1.00
Gender Parity Index in tertiary level enrolment, percentage	GPI3	70	7.69	14.04	0.32	0.04
Children under five mortality rate per 1000 live births	CM	103	−22.56	19.85	0.00	−0.24
Maternal mortality ratio per 100000 live births	MM	0	NA	NA	1.00	NA
People living with HIV, 15-49 year olds, percentage	HIV	0	NA	NA	1.00	NA
Tuberculosis prevalence rate per 100000 population	TBC	102	1.22	117.74	0.01	−0.20
Population using solid fuels, percentage	Fuels	23	−9.12	18.72	0.78	0.13
Proportion of the population using improved drinking water sources, total	Water	98	6.43	7.29	0.05	0.33
Proportion of the population using improved sanitation facilities, total	Sanitation	96	5.34	6.78	0.07	0.18
Internet users per 100 population	Internet	102	4.89	5.99	0.01	−0.23
Per capita GDP at current prices - US dollars / 100	GDP	101	5.48	10.50	0.02	−0.24

Table 3: Overview of the modified data set

we would expect positive changes in both variables to be linked with positive development in primary education. The variable Fuels does not have the expected sign either. As mentioned, it measures the proportion of households in a country which still use traditional and often health-affecting fuels such as wood in their dwellings. Theoretically, a decrease in this variable should be linked with an improvement of children's and teacher's health and thus in an improved primary education. The positive correlation coefficient in our case could possibly be explained by the high proportion of missingness and by the variable having rather indirect effects on primary education.
We use the plot *aggr()* from the VIM-package again to display graphically the proportion of missing values and the combinations of missings in the data set of differences (Figure 7), this time excluding the HIV-prevalence rate and the maternal mortality ratio.

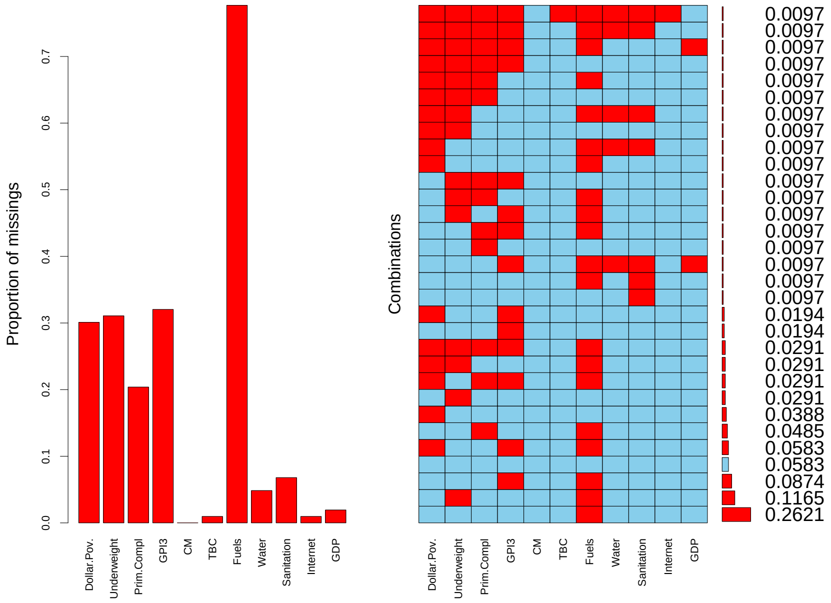

Figure 7: Missingness across the MDGs-variables, differences

The completely blue fourth row counted from the bottom in the right section of the graph shows that approximately 5.8 % of all cases are complete in the modified data set. Nevertheless, the other rows and thus the main part of all cases contain one or more missing values. This enables us to perform

a complete case analysis and compare it to the results obtained by some imputation methods.
As already mentioned, it is impossible do determine whether the mechanism leading to missingness in a variable is MNAR or MAR (Spieß, 2008), and it is still difficult to say whether it is MAR or MCAR. The following graph displays the distribution of GDP/capita conditional on the missingness of the variable "Primary completion rate" by means of two boxplots.

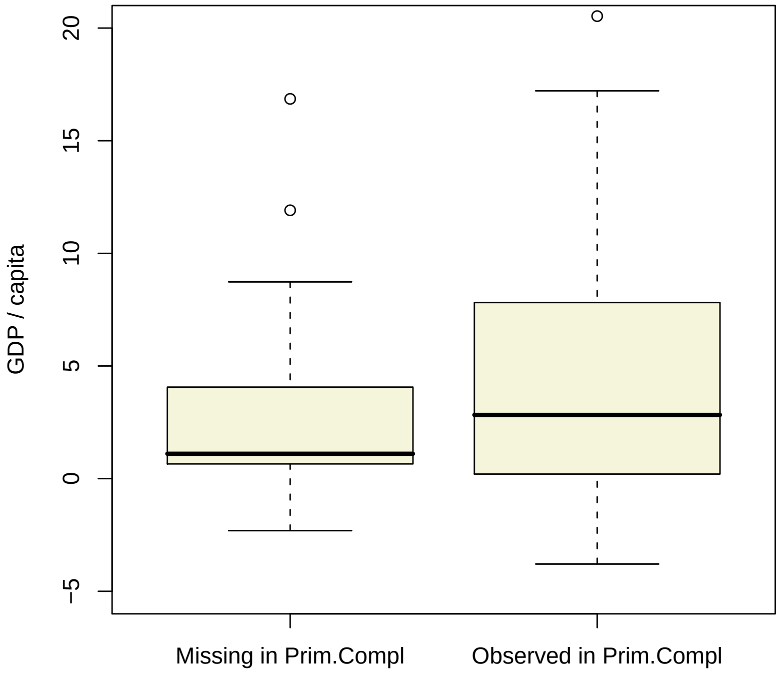

Figure 8: Distribution of GDP according to observation status in Prim.Compl

The median for the left boxplot, representing the cases for which the value of the primary completion rate was unobserved, is smaller than the one for the observed cases. The GDP/capita of the observed values is more scattered, indicating a greater variance.
The underlying idea is to look for differences in a covariate between the group of the cases observed in the dependent variable and the group of the cases

missing in that variable (see Little and Rubin, 2002, p. 41). If there are any differences, this could in turn be an indication of an influence of the covariate upon the distribution of missingness in the dependent variable, i.e. a MAR-mechanism. In the boxplot, the smaller median for the group of the unobserved values could be suspected of contributing to the missingness in the variable “Primary completion rate”. This would support the hypothesis that economically less developed countries tend to have more difficulties in data collection and processing. A formal t-test can be used to test whether the means in the two groups are different. However, the two-sided t-test on equality of the means of GDP/capita in the two groups (primary completion rate observed vs missing) yields a p-value of 0.09, thus not rejecting the null hypothesis (sign. 0.05) of the means being the same. The same holds for the t-test for the variable Internet with a p-value of 0.054. Considering the conservative nature of the t-test and the fact that a wrongly made MCAR-assumption can do far more harm in a regression analysis than a cautious decision for the MAR-assumption, we choose the latter to hold for the data, in particular as the two p-values mentioned are not exceedingly high above 0.05.

5 Methods for data with missing values

Our goal is to compare various methods of imputation for the missing values, in order to be able to conduct a regression of the primary completion rate on the other variables in our reduced data frame. Several methods to handle missing data will be presented in the next sections, beginning with the standard approaches listwise deletion and omitting variables. Alternative ways to deal with missing values include imputation of the mean, imputation by linear regression and more sophisticated methods like the EM-algorithm and multiple imputations, all of which will be considered in the next chapters.

5.1 Listwise deletion and omitting variables

Listwise deletion is the standard approach used by statistical software like *R*, but it bears some serious disadvantages compared to other methods. Its widespread use is primarily due to the fact that it is the easiest way to handle missing data. It consists of checking each case for completeness and using only the cases that do not lack any entry in one of the variables. This is especially a problem if the number of explaining variables is high. In this case, even small proportions of missingness will lead to a drastically reduced number of cases, given that the missingness is scattered throughout the data frame as in Table 4.

Despite the fact that there are four cases with observed values for both "Education" and "Water", the relation between these two variables in a multivariate analysis using listwise deletion would be computed by taking into account only one case, Zambia, which does not have any missing values. This loss of efficiency is accompanied by bias of the estimators for a regression when the structure of missingness is MAR or MNAR, see for example King et al. (2001, p. 52) and Little and Rubin (2002, p. 41). King et al. (2001) discusse the loss of efficiency resulting from listwise deletion even under the MCAR-assumption at length.

	Education	Water	HIV	Fuels	Internet
...	...	...	...	...	...
Nigeria	0.66	NA	0.05	0.82	0.12
Sierra Leone	0.48	0.60	NA	0.95	NA
Somalia	0.73	0.27	0.03	NA	0.16
Viet Nam	0.90	0.72	0.06	NA	0.27
Zambia	0.87	0.54	0.02	0.92	0.22
...	...	...	...	...	...

Table 4: Fictional data

Another approach is to omit one or more explaining variables from the data set and then conduct the regression using only complete cases, in order to obtain a higher proportion of complete cases. However, this method is highly criticised for potentially introducing bias of the estimators by ignoring explaining variables. Furthermore, it appears unsatisfactory, because researchers will usually have a certain idea of a model by which to explain their dependent variable. Simply neglecting some of the explaining variables should only be considered as a method of last resort as long as there are other methods to extract more information from the data. King et al. (2001) provide a concise formal description of the biasing effects of omitting variables on the parameters of a regression for a special case: Let $E(Y) = X_1\beta_1 + X_2\beta_2$ be a linear regression model, where β_1 is the main interesting effect and X_2 are one or more covariates additionally included to control for confounding variables. Furthermore, let missingness be confined to X_2. The so called infeasible estimator $b^I = (b^I_1, b^I_2)^T$ denotes the estimator for β from the regression of Y on X_1 and an entirely observed X_2. Let $b^O_1 = A_1 Y$ be the omitted variable estimator, excluding X_2, with $A_1 = (X_1^T X_1)^{-1} X_1^T$. Then the expectation value of b^O_1 is

$$E(b^O_1) = E(b^I_1 + F b^I_2) = \beta_1 + F\beta_2 \tag{1}$$

where F are regression coefficients of X_2 on X_1. Consequently, $bias(b^O_1) = F\beta_2$ (King et al., 2001, p. 66). King et al. (2001) however state that the loss of efficiency resulting from listwise deletion is in many instances high enough to turn the exclusion of a variable from the analysis into a more appropriate option. The tradeoff between loss of efficiency (introducing variance of the estimators) from listwise deletion and bias resulting from omitting variables can be expressed through the MSE (King et al., 2001, p. 52). Furthermore, listwise deletion may also lead to bias of the estimators if the MCAR-assumption does not hold.

5.2 Imputation by the unconditional mean

Imputation by the unconditional mean denotes the approach of imputing each missing value with the arithmetic mean of the respective variable. Though it seems obvious and easy to handle, it is far from being a perfect method to deal with missing data. It leads to underestimated variances in the completed data set, which in turn biases the estimates for the coefficients as well as the significance levels. Since the empirical variance of a quantity X is defined as $s_x = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$ and imputation by mean affects it by simply increasing n but adding nothing to $(x_i - \overline{x})^2$, every missing value imputed by the mean will decrease the variance. Similar bias holds for the covariance of two variables $s_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})\ (y_i - \overline{y})$. Since the estimates of the linear regression coefficients in the univariate case are defined as $\beta_j = \frac{S_{x,y}}{S_x}$, imputation by unconditional means will obviously yield inconsistent estimates even in the MCAR case. Theoretically, this problem can be overcome by applying the adjustment factor $\frac{n-1}{m-1}$ for the variance and in an analogue way for the covariances, where m is the number of observed values. However, this works only for MCAR-data and yields unsatisfactory estimates for the resulting variance and covariance (Little and Rubin, 2002, p. 44). Furthermore, imputation by the unconditional mean violates some of the assumptions for linear models such as homoscedasticity, the normal distribution of the error terms ϵ_i and σ^2 being the variance of the error terms. Spieß (2008) provides further proofs and detailed explanations.

5.3 Buck's Method

Considering the shortcomings of the method of imputing unconditional means for missing values, imputing means conditioned on the information still available seems to be the obvious next step. Buck (1960) proposes the imputation of conditional means obtained by linear regressions. This includes using the available variables in a particular case to calculate the conditional means for the variables missing in that case. This method yields slightly improved results compared to the imputation of unconditional means, especially regarding estimators for the overall mean (Little and Rubin, 2002, p. 45). In the case of a linear regression as the subsequent analysis however, estimators for the variance are needed again as the basis of estimators for regression coefficients and their variances. For instance, let Y_2 be a variable containing missing values which we try to impute by means of a linear regression of Y_2 on Y_1 in a bivariate data set using the observed values. By partitioning the total variance of Y_2, we obtain

$$\sigma_{22}^2 = \beta_{21}^2 \, \sigma_{11}^2 + \sigma_{22 \cdot 1}^2 \tag{2}$$

where $\beta_{21} = \frac{\sigma_{12}^2}{\sigma_{11}^2}$ denotes the coefficient of the univariate linear regression of Y_2 on Y_1, σ_{11}^2 is the variance of Y_1 and $\sigma_{22 \cdot 1}^2$ the expected variance of Y_2 given Y_1. In terms of the well-known partitioning of the variance in the case of a linear regression model, the first term of the sum is the variability explained by the model and the second term contains the variability not explained by the model, i.e. the residual variance. An imputed value for a missing value of Y_2 in a case with a given value of Y_1 consists in the value for the regression model at y_{1i}, the respective realisation of Y_1 in the i-th case. Hence, the imputed value of Y_2, $\tilde{y_{2i}}$ lies exactly on the regression line, not adding any residual variance to the total sample variance of Y_2. Each imputed value for Y_2 will thus contribute to biasing the sample variance of Y_2 and underestimating σ_{22}^2 by the quantity $\frac{n_{mis,2}}{n-1} \, \sigma_{22 \cdot 1}^2$ in the bivariate case (Little and Rubin, 2002, p. 46), $n_{mis,2}$ being the number of missing values in Y_2. In general, the sample variance from data containing values imputed by

Buck's method underestimates σ_{jj}^2 by the quantity

$$\frac{1}{n-1}\sum_{i=1}^{n}\sigma_{jj\cdot\, obs,i} \tag{3}$$

$\sigma_{jj\cdot\, obs,i}$ being zero in an observed case and the residual variance of a regression of Y_j on the observed variables in the i-th case if Y_j is missing in that case (Little and Rubin, 2002, p. 46). In the case of MCAR-data, consistent estimates of σ_{jj}^2 can be obtained by using the sample variance of the complete cases to correct the bias from (3). However, this requires extensive modification of standard software, which is usually not worth the trouble since there are more appropriate, mostly likelihood-based approaches to impute missing data. Since our goal is to compare several standard methods for our real-world data set, we will nevertheless run a linear model using imputation by Buck's method in chapter 6, but without correction for the variance matrix.

5.4 The technique of multiple imputation

Multiple imputation (MI) denotes an approach in which m values are imputed for each missing, thus creating m completed data sets with the same observed data, but varying imputed values. This technique helps to express the uncertainty associated with any sort of imputation.

Mathematically, MI consists of taking draws from a specified distribution and thus taking into consideration the uncertainty of the process. The approach we will use (King et al., 2001) assumes that the data are MAR and that the joint distribution of the variables is a multivariate normal $N(\mu, \Sigma)$. If D_i is the vector of the p variables in the ith observation with i=1,...,n, the likelihood function for the complete, not entirely observed multivariate normal data is

$$L\left(\mu, \Sigma\left|D\right.\right) \tilde{} \prod_{i=1}^{n} N\left(D_i|\mu, \Sigma\right) \tag{4}$$

Since we are not able to fully observe D, the entire data set without missings, we have to calculate μ and Σ from D_{obs}, the observed data, assuming normal marginal densities:

$$L\left(\mu,\Sigma\left|D_{obs}\right.\right)\,\tilde{}\,\prod_{i=1}^{n}N\left(D_{i,obs}\left|\mu_{i,obs}\right.,\Sigma_{i,obs}\right)\tag{5}$$

where $D_{i,obs}$ is the observed part of the i-th row of D. $\mu_{i,obs}$ and $\Sigma_{i,obs}$ are the corresponding subvector and submatrix for μ and Σ, containing only elements for observed values in D_i. Thus, $\mu_{i,obs}$ and $\Sigma_{i,obs}$ do not change in values over i, but they do change regarding their length and composition. This makes (5) difficult, if not impossible to compute. The actual imputation of missing values in the j-th variable is done by means of a regression of D_j on D_{-j}, where the latter is D without the j-th variable. The parameter estimates $\widehat{\beta}$ of this regression can be calculated directly using μ and Σ. The imputed value for case i in variable j then has the form

$$\tilde{D}_{i,j}=D_{i,-i}\tilde{\beta}+\tilde{\epsilon}_i\tag{6}$$

$\tilde{}$ stands for a random draw from the posterior μ and Σ, where $\tilde{\beta}$ expresses the uncertainty of not knowing exactly μ and Σ and $\tilde{\epsilon}_i$ the uncertainty generated by the world (notation and content from King et al. (2001)).

As mentioned beforehand, the point about multiple imputation consists in generating several values for each missing, thereby introducing uncertainty. As for the number of completed data sets needed to obtain efficient estimators, King et al. (2001) suggest as little as 5 to 10 imputations per missing value, unless the proportion of missingness is exceedingly high. For an example of applied multiple imputation in the context of developing countries, see Gartner and Scheid (2003). In the end, one will usually be interested in some quantity of interest Q like the mean or a regression coefficient, which is m-fold after the process of multiple imputation, where m is the number of imputations. To combine these m data sets regarding Q, it is sufficient to simply take the mean $\overline{q}=\frac{1}{m}\sum_{j=1}^{m}q_j$ of the m slightly different versions of q. Multiple imputation allows us to specify a variance of the multiple imputation point estimate q. Let $SD(q_j)$ be the estimated standard error of q_j from the j-th data set. Then the variance of the multiple imputation point estimate is the average of the m variances from within each data set $SD(q_j)^2$, j=1,...,m, plus the sample variance across the m point estimates $S_q^2 = \sum_{j=1}^{m} \frac{(q_j - \overline{q})^2}{(m-1)}$:

$$SD(q)^2 = \frac{1}{m} \sum_{j=1}^{m} SE(q_j)^2 + S_q^2 \left(1 + \frac{1}{m}\right) \tag{7}$$

The last factor serves as a correction for $m < \infty$ (King et al., 2001, p. 53).

5.5 The expectation maximisation algorithm

Since the computation of (5) is difficult or impossible (King et al., 2001, p. 54), new approaches have been developed to calculate the posterior or at least its parameters to draw samples for multiple imputations from. This includes especially the Imputation-Posterior algorithm (IP) and the Expectation-Maximization algorithm (EM). IP, although being considered a standard for multiple imputation, is said to bear some practical disadvantages which include particularly slow convergence and difficult application due to the use of Markov Chain Monte Carlo methods (King et al., 2001). Thus, we will resort to the EM-algorithm and its implementation in the R-package *Amelia* (Honaker et al., 2007) to find imputations for the MDGs data set. For a detailed explanation and application of the IP-algorithm, we refer the reader to Gartner and Scheid (2003).

The basic idea of EM is to iteratively find the parameters of the distribution of the complete data by maximising the likelihood-function of the complete data given the observed data and starting values for the parameters. This involves calculating the expected log-likelihood of the complete data given the observed data in an E(xpecation)-step and maximising the obtained expectation under the parameters in a M(aximisation)-step. Ideally, this algorithm will converge running it iteratively. Let Z_{obs} be the observed and Z_{mis} the missing part of the complete data Z. Θ denotes the parameters that describe the distribution of the data, for example $\Theta = (\mu, \Sigma)$ for a multivariate normal. The EM-algorithm can now be defined as follows (Dempster et al., 1977):

1. E-step: Compute the conditional expectation

   $$Q\left(\Theta\right) = Q\left(\Theta \left|\Theta^{(i)}\right.\right) = E\left[l\left(Z, \Theta\right)\left|Z_{obs}, \Theta^{(i)}\right.\right] \qquad (8)$$

   where $l\left(Z, \Theta\right)$ denotes the log-likelihood of the complete data. $\Theta^{(i)}$ indicates the i-th iteration of the algorithm.

2. M-step: Find $\Theta^{(i+1)}$ by maximising $Q\left(\Theta\right)$ under Θ and use the new parameter $\Theta^{(i+1)}$ for the next iteration.

For the first iteration step, we have to set a guess for $\Theta^{(0)}$. The EM-approach can be compared to imputing missing values by a linear regression of the particular variable in which a value is missing on the other variables, then rerunning this regression including the newly imputed values and imputing again until convergence. EM has the advantages of converging relatively quickly, deterministically and that the objective function increases with every iteration (King et al., 2001). A major disadvantage of EM is that it yields only the parameters of the underlying posterior, not the distribution itself, thus ignoring the estimation uncertainty. It is possible to get multiple imputations from EM-values by using the posterior variance, but this only takes into consideration fundamental variance, not estimation uncertainty (King et al., 2001, p. 54). Therefore, modified versions of EM are implemented in *Amelia*. EMs (EM with sampling) uses the variance matrix $V\left(\widehat{\Theta}\right)$ (not to be confused with $\widehat{\Sigma}$, the actual variance matrix of the posterior) of the parameter estimates $\widehat{\Theta}$ obtained after running generic EM to express estimation uncertainty. It draws m simulated Θ from a normal with mean $\widehat{\Theta}$ and variance $V\left(\widehat{\Theta}\right)$, uses them to compute the values of $\tilde{\beta}$ in (6) and thereby creates m imputations for every missing. EMs works well in large samples, but the approximation by a normal can cause bias in the standard errors of the multiple imputations in the case of small samples, highly skewed distributions or a high number of variables (King et al., 2001). EMis (EM with importance resampling) tries to overcome these drawbacks by treating draws of Θ from its asymptotic distribution obtained with EMs only as first approximations to the final posterior. It keeps only those draws of Θ with probability proportional to the importance ratio (IR), which is defined as the proportion of the actual posterior to the asymptotic normal distribution at $\tilde{\Theta}$, formally $IR = \frac{L\left(\tilde{\Theta}|Z_{obs}\right)}{N\left(\tilde{\Theta}\middle|\tilde{\Theta},V\left(\tilde{\Theta}\right)\right)}$. EMis is implemented as the default algorithm in *Amelia*, and any reference to EM in the next chapter means EMis.

6 Application of various methods for missing values

It is difficult to determine a best method for treating missing values in the case of the MDGs-data set, since the models use real-world data instead of simulated values. However, assuming the data to be MAR, references such as King et al. (2001) and Little and Rubin (2002) indicate that multiple imputations combined with the EM-algorithm could be the best choice to impute values for the missing entries in the MDGs-data. EM combined with MI has been shown to yield better results than imputation by the unconditioned mean and Buck's method in numerous simulated and real-world examples, for instance in King et al. (2001). Regarding listwise deletion, King et al. (2001) state that there are four conditions which have to hold for it to yield better results than EM combined with MI: The analysis model has to be conditional on X, such as a regression model, which is the case. There should be MNAR-missingness in X, which would lead to wrong results from EMis, a precondition which cannot be tested as described in section 4.2. Furthermore, missingness in X must not be a function of Y (i.e. the primary completion rate) and unobserved variables affecting Y should not exist. At least the latter is likely to be wrong for our data set. Finally, the proportion of missing cases alone, roughly 68 %, introduces loss of efficiency which would equalise any advantages gained by avoiding possible bias of the estimators from MNAR-mechanism in EMis. Taking also into consideration the drawbacks of imputation by the unconditional mean and Buck's method and the advantages of the EM-algorithm combined with multiple imputation, the latter should theoretically be the best choice for our data set.

Table 5 is the summary of a linear regression of the primary completion rate on the other variables in the modified data set according to table 3, using complete cases and excluding maternal mortality, HIV-prevalence rate and tuberculosis-prevalence rate. The latter can be assumed not to contribute a lot of additional information to the analysis, since the general health status in a country is already measured by child mortality. To avoid collinearity, we will omit the variable from the analysis model, but keep it for following imputation techniques, since it can make sense to add variables not included in the analysis to the imputation model (King et al., 2001, p. 57).

Dependent Variable: Primary Completion Rate

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.288	NA	NA	NA
Underweight	-5.559	NA	NA	NA
Dollar.Pov.	0.948	NA	NA	NA
GPI3	0.388	NA	NA	NA
CM	-0.221	NA	NA	NA
Fuels	0.314	NA	NA	NA
Sanitation	NA	NA	NA	NA
Water	NA	NA	NA	NA
Internet	NA	NA	NA	NA
GDP	NA	NA	NA	NA
$R^2 = NaN$				
N=6				

Table 5: Linear model of the primary completion rate , listwise deletion

The results are not very satisfying, as 97 of the 103 cases have been deleted due to missingness in one or more of the variables. *R* does not manage to calculate some of the regression coefficients and fails to compute any of the standard errors and significance levels due to the low number of cases and resulting singularities. The signs for the coefficients of the proportion of people living on less than 1 US-Dollar per day and for the proportion of households using traditional fuels contradict the theoretical causalities explained in Sachs et al. (2004) and the coefficient of the proportion of people living on less than 1 US-Dollar per day is positive instead of negative like the correlation coefficient in table 3. Despite the issues arising with the approach of omitting explaining variables from the regression, it could be useful to consider the results of the regression without including fuels, the variable with the highest proportion of missingness, in order to obtain more complete cases and to be able to compare these results to further analysis.

Table 6 displays the parameters of a linear model excluding the variable "Fuels" in the cases of listwise deletion, imputation by the unconditioned mean, Buck's method and the EM-algorithm combined with MI. The italic values below the actual estimates of the coefficients are the corresponding standard errors and the values below are the p-values. For a comparison of the estimates of the four models, Table 7 shows the arithmetic means of the regression output over all variables excluding the intercept for each of the four methods. The intercept terms of the models in Table 6 can be interpreted as the change in the dependent variable "Primary completion rate" which would occur for a country where none of the predictor variables changes at all between the 1990s and the following decade. The parameter estimates for the variables are the additional changes on the intercept term if the respective covariate changes by 1 unit.
70 observations have been deleted due to missingness in the complete-cases analysis, which means that we can still use 33 countries. The adjusted R^2 is 0.085 and all of the regression parameters and significance levels can be computed now. The estimates for Underweight, the gender parity index in tertiary education (GPI3), Child Mortality, Sanitation, Water and GDP have the expected signs, whereas we would expect a negative coefficient for the proportion of people who live on less than 1 US-Dollars per day (Dollar.Pov.) and a positive one for the Internet users per 100 inhabitants. On the other side, none of the coefficients is anywhere near the common significance threshold of 0.05 and the adjusted R^2 is comparably small. Apart from the estimates for the gross domestic product per capita and the estimate of the proportion of households using improved sanitation facilities, there are no changes in the signs of the coefficents. We will proceed analysing the differences between the regression models using different methods of imputation.

The parameter estimates of the data set using Buck's method of imputing values found by a non-iterative linear regression model all have remarkably small p-values. Without going into detail for each of the predictor variables,

Dependent Variable: Primary Completion Rate

	Listwise Deletion	Imputation by uncond. mean	Buck's method	EM + MI
Intercept	5.359	5.068	4.703	4.709
sd	*3.424*	*1.641*	*1.516*	*2.251*
(p)	*(0.131)*	*(0.003)*	*(0.003)*	*(0.039)*
Dollar.Pov.	0.116	0.124	0.268	0.257
sd	*0.151*	*0.101*	*0.095*	*0.134*
(p)	*(0.449)*	*(0.223)*	*(0.006)*	*(0.058)*
Underweight	-0.268	-0.351	-0.581	-0.464
sd	*0.704*	*0.35*	*0.338*	*0.426*
(p)	*(0.707)*	*(0.318)*	*(0.089)*	*(0.279)*
GPI3	0.186	0.059	0.191	0.145
sd	*0.107*	*0.065*	*0.059*	*0.084*
(p)	*(0.094)*	*(0.364)*	*(0.002)*	*(0.089)*
CM	-0.098	-0.045	-0.069	-0.055
sd	*0.074*	*0.037*	*0.036*	*0.046*
(p)	*(0.197)*	*(0.228)*	*(0.06)*	*(0.235)*
Water	0.146	0.274	0.248	0.288
sd	*0.203*	*0.11*	*0.105*	*0.121*
(p)	*(0.479)*	*(0.015)*	*(0.02)*	*(0.02)*
Sanitation	0.1	-0.001	0.091	0.114
sd	*0.246*	*0.115*	*0.111*	*0.153*
(p)	*(0.687)*	*(0.992)*	*(0.414)*	*(0.456)*
Internet	-0.361	-0.112	-0.172	-0.157
sd	*0.418*	*0.131*	*0.125*	*0.173*
(p)	*(0.397)*	*(0.397)*	*(0.173)*	*(0.366)*
GDP	0.111	-0.116	-0.14	-0.135
sd	*0.377*	*0.074*	*0.071*	*0.089*
(p)	*(0.77)*	*(0.121)*	*(0.052)*	*(0.132)*
N	33	103	103	103
adj. R^2	0.085	0.107	0.261	0.213

Table 6: Estimates of a linear regression with various approaches for missing values

	Coef.	sd	t	p
Listwise deletion	0.173	0.285	0.813	0.472
Imp. by mean	0.135	0.123	1.159	0.332
Buck's method	0.220	0.118	2.028	0.102
EM + MI	0.202	0.153	1.433	0.204

Table 7: Arithmetic means of the parameters of the linear models over all variables

all of the p-values of this model are smaller than their counterparts of the other regression models. They are even significant, i.e. below the 5 % significance threshold, for the predictor variables "Population living on less than 1 US-Dollar per day" (0.006), "Gender parity index in tertiary education" (0.002) and "Proportion of the population using improved drinking water sources" (0.02). The adjusted R^2 is higher than the one of the other models at 0.261. These results provide a good example for the dangers associated with inconsiderately imputing values by means of a method which at a first glance even seems to be an improvement of imputation by the unconditional mean. Since the proportion of missing values in the dependent variable "Primary completion rate" is at 20 %, Buck's method included imputing values for those missings by regressing the variable on the other variables, which are assumed to be the predictor variables in our subsequent regression analysis. Basically, one fifth of the values for the primary completion rate in the new data set are the predictions of a regression from primary completion rate on the rest of the variables. Thus, it should not surprise to find strong indications for a linear relation with primary completion rate as the dependent variable in the new data set, however those findings are obviously mere artefacts resulting only from the method of imputation and not from actual structures in the real world. This example demonstrates the need for careful consideration of the method of imputation, taking into account the assumed missingness mechanism and the goals of the subsequent analysis.

The model using data with values imputed by the unconditional mean of the respective variable in turn has an adjusted R^2 of 0.107, which is not much higher than the one of the listwise deletion model (0.085) despite the fact that the latter uses only 33 cases instead of all 103 cases as does the imputation by mean model. The average over the parameter estimates (Table 7) is by far the lowest of the four methods. As mentioned, imputation by the unconditioned mean biases those estimators and should not be an option to deal with missing values. The p-values are all far above 5 %, at least not suggesting wrong conclusions like Buck's method.

The analysis based on EMis and MI consists of a linear regression model for each of the newly created data sets. The values for the parameter estimates are the arithmetic means of the estimates over all the data sets, whereas the standard errors of the estimates are calculated according to equation (7). The p-values are computed from a t-test using the values of the estimates and standard errors in Table 6. The arithmetic mean over all the absolute estimates for the coefficients is higher than the one for the models using imputation by the unconditional mean and especially listwise deletion, and the p-values are smaller. The most notable change occurs for the estimator of the coefficient for the proportion of the population using improved drinking water sources. It becomes significant in the EMis + MI model and indicates a positive correlation between this variable and the primary completion rate. The change in the corresponding p-value from 0.479 for listwise deletion to 0.02 for EM+MI originates from the parameter estimate doubling for the EM+MI-model compared to listwise deletion and the standard error decreasing drastically from 0.203 to 0.121. However, p-values have to be considered very carefully because of issues arising with multiple testing. Looking at 36 p-values at the same time (9 in every model) drastically increases the chance for at least one of them to fall randomly below the 5 % - level. The estimate for the percentage of the population below $1 (PPP) per day with the already mentioned, unexpected value of 0.257 is close to significance at a p-value of 0.058. Compared to the values of the analysis using listwise deletion, this is mainly due to the much higher value of the coefficient estimate in the MI-model (0.257, as opposed to 0.116 for listwise deletion), whereas the standard errors are approximately of the same magnitude. The p-value for the estimate of the gender parity index in tertiary education is also comparably small at 0.089, but not improving much upon the one of the listwise deletion model and still staying insignificant.

We will now come back to the actual interpretation of the regression output for the effects on primary education. The estimate for the variable "Percentage of people below 1 US-Dollar per day" still has an unexpected sign in all of the models, since it seems unlogical for a higher increase or a less rapid decrease in the proportion of people living in extreme poverty from the 1990s to the years from 2000 onwards to be linked with worsened results or less improvement in the educational performance of a country. However, it is important to take into consideration the fact that the data set consists of differences between the two decades. A country with a comparably high percentage of people living in extreme poverty in the 1990s could make huge improvements in this variable (improvement as in the percentage of people in extreme poverty decreasing between the two decades, resulting in a negative sign for the data set of differences). However, due to the country having little ressources in the 1990s, the improvement of the primary completion rate could be delayed, resulting in less increase in this variable for countries advancing quickly in the decrease of poverty. Figure 9 seems to indicate that there are indeed three groups of countries regarding the effect of extreme poverty on the primary completion rate over time.
Countries which were able to decrease the proportion of people living on less than 1 US-Dollar per day by about 10 or more % made less improvements in primary education the higher the decrease of extreme poverty was, probably due to the fact that they did not have enough ressources by the beginning of the period to invest a lot in education. Countries which reduced the proportion of people living in extreme poverty by 10 to 0 % made the most advancement in primary educational performance. These could be countries with a small percentage of people living in extreme poverty at the beginning of the period compared to other developing countries that managed to further reduce poverty and at the same time had the ressources to drastically improve educational outcome. Countries which deteriorated

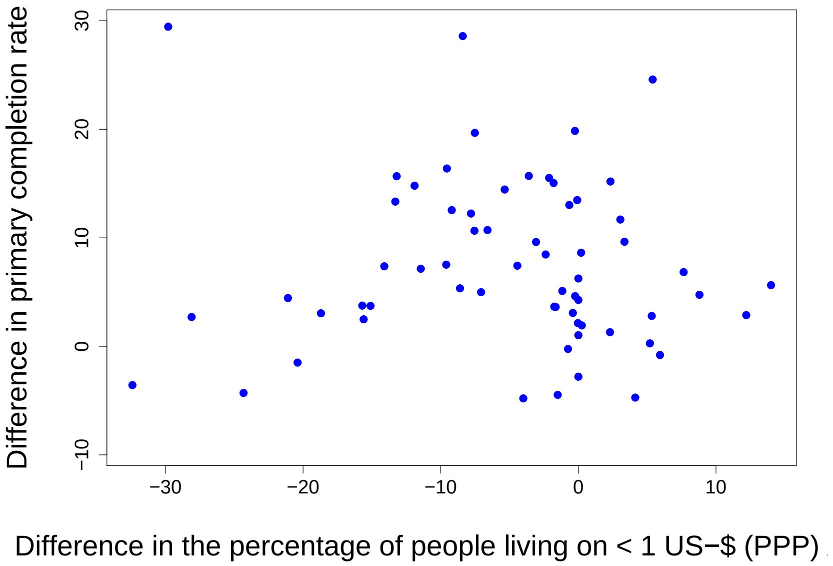

Figure 9: Correlation between Dollar.Pov. and Prim.Compl for differences 2000s - 1990s

regarding extreme poverty, i.e. which had a difference above 0 for the respective variable, mostly did not make good progress in primary education either. The relation between the differences of the primary completion rate and the proportion of people living in extreme poverty seems to be non-linear and reversely u-shaped. This may also explain the rather small p-values of most of the predictor variables. The influences on the dependent variable primary completion rate, if there are any, are possibly not linear for the data set of differences between the two decades.

The arithmetic mean of the primary completion rate over all the available cases for this variable increased by 6.89 between the two decades (see Table 3). Interpreting the significant intercept from the EM+MI-model, 4.709 of this total change would even have occurred if all the other variables included in the model had not changed at all. Note that these two values are not fully comparable, because they stem from two slightly different data sets (original data set of differences for the mean, data completed with EM+MI for the Intercept). The only significant parameter estimate for the explaining variables is the one for the proportion of the population using improved drinking water sources at 0.288. An improvement of 1 % in the proportion of people using enhanced water ressources compared to the intercept brings about a 0.288 % positive change in the primary completion rate using data filled in with EMis and MI. The other parameter estimates are not significant at 0.05-level, although the already discussed estimate for the variable "Percentage of people below 1 US-Dollar per day" is close to significance in the EM+MI-data with a p-value of 0.058.

7 Conclusion

The newly created data set of differences between the two decades largely fails to detect possible relations between the different variables of the MDGs-data set over time, even when missing values are imputed by reasonable methods like multiple imputations. This could be for various reasons: First of all, interpretation of the effect of change over time in one variable on change over time in another variable is more complicated than looking at a time series directly. There might be various layers of values as described in Figure 9 which behave differently, for example according to whether the change in the explaining variable is negative or positive. Furthermore, consideration of the starting values (i.e. the mean of the years from 1990 to 1999) could help to improve the outcome of the analysis. However, the MDGs-data is possibly just too scattered by missingness to maintain the time series character of the data and simultaneously analyse correlation between the variables. Even the UN tend to use it only for univariate time-series analysis. For example, researchers from the African Development Bank define an indicator for a country missing “where the data available are such that two data points with at least 3 years apart cannot be found” (Mubila and Pegoue, 2008, p. 62). This is obviously a very low standard for measuring a time series and underlines the difficulties researchers have to face in the case of data from developing countries.

The high proportion of missingness in turn demonstrates the need for considerate handling of missing values. The drawbacks and dangers of Buck’s method have been demonstrated by the apparently wrong p-values for the linear model applied to the data. Imputation by the unconditional mean has similar disadvantages and in the first place biases estimators. Consequently, it should not be used for creating imputations. Most authors point out that multiple imputation combined with the EM- or IP-algorithm is currently the first choice of general purpose imputation techniques, i.e. methods that are applicable to any data set under certain conditions; see for instance King et al. (2001) and Spieß (2008). It yields unbiased estimators under the

MAR-assumption and helps to reduce the loss of efficiency associated with listwise deletion. One of the main advantages of MI is the correction for underestimation of the variance of the data, which is one of the more serious problems of imputing single values. Another advantage is the comparably weak assumptions which have to be made beforehand for the use of MI - the missingness mechanism being MAR and the distribution of the variables to be jointly multivariate normal. MI as used in *Amelia* even yields satisfactory results when the second assumption is violated, for example in the case of categorical variables (King et al., 2001, p. 53). In the case of the MDGs-data set of differences, it helps to find at least some indications for dependencies between the variables. The findings are significant in only one case (parameter estimate for "Water") and thus have to be interpreted with caution because of the problem of multiple testing.

Despite its advantages compared to other methods, the EM-algorithm becomes instable and slow for data with high proportions of missingness (Little and Rubin, 2002, p. 130). For instance, running *Amelia* on the original data set as in Table 1 resulted in convergence only after a very high number of iterations (up to approximately 1000), great differences in the number of iterations needed for each of the new data sets and sometimes failure to converge at all. It would certainly be difficult to overcome the problem for missingness being as high as in this data set, but there are methods to improve the results of EM + MI taking into consideration prior knowledge of the real distribution of the data. Researchers will often have some idea about the approximate value of a variable for a certain case, and it seems obvious to include this knowledge in the imputation stage. Apart from taking advantage of time-series, which can already be used in the default options of *Amelia* to find imputations, there is also the possibility to include Bayesian priors for missing entries in the data matrix (Honaker and King, 2010). This can either be done by specifying a point prior with a standard deviation or a confidence range for the missing value. The final imputation for this data point is a weighted mean of the prior value set beforehand and the model-based imputation. The priors are included in the E-step of the EM-algorithm and are the more influential upon the final value for the imputation the smaller their variance set by the researcher. The model-based imputation in turn will downweight the prior when the predictive strength of the model is high. In the context of development aid, this approach is of special interest in the case of values which are missing for a country for known reasons, this is values which are not missing at random. Let us imagine a country which entered into civil war for a certain year. Data collection will obviously be poor for this year, as there will certainly be other priorities than a high-level statistical infrastructure. For example, let the variable "percentage of the population with access to improved sanitation" be missing in that particular year. Since it can at least be assumed not to have improved compared to previous years, it would make sense to include a prior according to such knowledge with an appropriate standard deviation reflecting the uncertainty of the guess (or alternatively a confidence interval instead of a point guess). This approach could be particularly useful for the process of data collection in organisations like the UN. Missing values in UN-data sets are often the results of national statistical agencies failing to deliver the data for certain years and variables. The UN however are able to resort to their own expertise or to second-hand data from NGOs which can be included as priors, constituting guesses instead of fixed values in the final data set. The best approach to a specific missing-values problem is certainly not to follow a particular method straight forward, but to take into careful consideration circumstances such as the most likely missing mechanism, the proportion of missingness, the assumed distribution of the data and possible prior knowledge of the missing values. At the end of the day, the point about imputation techniques is not to indiscriminately "invent" new data or even manipulate the existing data, but to carefully gain access to a much bigger part of the data set than with a complete cases analysis. An image taken from Honaker and King (2010) summarises this idea:

> *If archaeologists threw away every piece of evidence, every tablet, every piece of pottery that was incomplete, we would have entire cultures that disappeared from the historical record. We would no longer have the Epic of Gilgamesh, or any of the writings of*

Sappho. It is a ridiculous proposition because we can take all the partial sources, all the information in each fragment, and build them together to reconstruct much of the complete picture without any invention. Careful models for missingness allow us to do the same with our own fragmentary sources of data.

Honaker and King (2010, p. 563)

References

Buck, S. F. (1960). A Method of Estimation of Missing Values in Multivariate Data, Suitable for Use with an Electronic Computer, *Journal of the Royal Statistical Society Ser. B, 22, 302-306* .

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. (1977), pp. 1-38* . `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.7580&rep=rep1&type=pdf`, Last Accessed: 26.07.2010.

Gartner, H. and Scheid, S. (2003). Multiple Imputation von fehlenden Werten mit Daten über Unterernährung und Kindersterblichkeit, *Sonderforschungsbereich 386, Paper 322* . `http://epub.ub.uni-muenchen.de/1703/1/paper_322.pdf`, Last Accessed: 26.07.2010.

Honaker, J. and King, G. (2010). What to do about Missing Values in Time Series Cross-Section Data, *American Journal of Political Science, Vol. 54, No. 2, April 2010* pp. 561–581. `http://gking.harvard.edu/files/pr.pdf`, Last Accessed: 25.07.2010.

Honaker, J., King, G. and Blackwell, M. (2007). *Amelia: Amelia II: A Program for Missing Data.* R package version 1.1-23, `http://gking.harvard.edu/amelia`, Last Accessed: 18.07.2010.

King, G., Honaker, J., Joseph, A. and Scheve, K. (2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation, *American Political Science Review* **95**(1): 49–69. `http://gking.harvard.edu/files/evil.pdf`, Last Accessed: 18.07.2010.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*, Probability and Statistics, second edn, Wiley, New Jersey.

Millennium Development Goals: Metadata (2010). `http://mdgs.un.org/unsd/mdg/Metadata.aspx`. Last Accessed: 26.07.2010.

Mubila, M. and Pegoue, A. (2008). Toward a Methodology for Computing a Progress Composite MDG Index, *Le Journal statistique africain, numéro 7, novembre 2008* pp. 60–72. http://www.afdb.org/fileadmin/uploads/afdb/Documents/Publications/4-towards.pdf, Last Accessed: 21.07.2010.

OECD (2008). The Paris Declaration on Aid Effectiveness and the Accra Agenda for Action, http://www.oecd.org/dataoecd/11/41/34428351.pdf. Last Accessed: 31.05.2010.

OECD (2009). Dac list of oda recipients, http://www.oecd.org/dataoecd/32/40/43540882.pdf. Last Accessed: 18.07.2010.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org, Last Accessed: 25.07.2010.

Rubin, D. B. (1976). Inference and Missing Data (with Discussion), *Biometrika, 63(3)* pp. 581–592.

Sachs, J. (2005). *The End of Poverty. How We Can Make It Happen in Our Lifetime*, Penguin Books, London.

Sachs, J. D., McArthur, J. W., Schmidt-Traub, G., Kruk, M., Bahadur, C., Faye, M. and McCord, G. (2004). Ending Africa's Poverty Trap, *Brookings Papers on Economic Activity* **35**(2004-1): 117–240. http://www.unmillenniumproject.org/documents/BPEAEndingAfricasPovertyTrapFINAL.pdf, Last Accessed: 18.07.2010.

Spieß, M. (2008). *Missing-Data Techniken: Analyse von Daten mit fehlenden Werten*, Lit Verlag, Hamburg .

Templ, M. and Alfons, A. (2009). *VIM: Visualization and Imputation of Missing Values*. R package version 1.3.2, http://cran.r-project.org/package=VIM, Last Accessed: 25.07.2010.

Toutenburg, H., Heumann, C. and Nittner, T. (2004). *Statistische Methoden bei unvollständigen Daten*, Department für Statistik, Ludwig-Maximilians-Universität München. http://www.stat.uni-muenchen.de/sfb386/papers/dsp/paper380.ps, Last Accessed: 18.07.2010.

UN Millennium Project (2005). *Toward Universal Primary Education : Investments, Incentives and Institutions*, Task Force on Education and Gender Equality. http://www.unmillenniumproject.org/documents/Education-complete.pdf, Last Accessed: 18.07.2010.

United Nations (1948). Universal Declaration of Human Rights, http://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf. Last Accessed: 18.07.2010.

United Nations (2009). The Millenium Development Goals Report 2009, *UN Millennium Project* . http://www.un.org/millenniumgoals/pdf/MDGReport2009ENG.pdf, Last Accessed: 18.07.2010.

United Nations Department of Economic and Social Affairs (2010). Website of the Small Island Developing States Network, http://www.sidsnet.org/2.html. Last Accessed: 18.07.2010.

United Nations Development Group (2003). Indicators for Monitoring the Millennium Development : Definitions, Rationale, Concepts and Sources. http://mdgs.un.org/unsd/mdg/Resources/Attach/Indicators/HandbookEnglish.pdf, Last Accessed: 26.07.2010.

United Nations Millennium Summit (2000). *United Nations Millennium Declaration*, United Nations Department of Public Information, New York. http://www.un.org/millennium/declaration/ares552e.htm, Last Accessed: 18.07.2010.

A Data Sets

The following data sets were used for this thesis:

Data set on the Millennium Development Goals, `http://mdgs.un.org/unsd/mdg/Handlers/ExportHandler.ashx?Type=Csv`,
last accessed: 26.07.2010

UN-Data: Per capita GDP at current prices - US dollars, `http://data.un.org/Data.aspx?q=gdp+capita&d=SNAAMA&f=grID%3a1013bcurrID%3aUSD%3bpcFlag%3a1`, last accessed: 26.07.2010

UN-Data: GDP per capita in 2006, current international dollars (PPPs) (WB estimates), `http://data.un.org/Data.aspx?q=GDP+per+capita+2006&d=CDB&f=srID%3A29922%3Byr%3A2006`, last accessed: 26.07.2010

B Contents CD

1. Data

    (a) Modified data set on the Millennium Development Goals: mdgs_komplett.csv

    (b) Per capita GDP at current prices - US dollars: GDPPC_USD_countries90_08.csv

    (c) GDP per capita, current international dollars (PPPs): GDP_Capita_2006.csv

    (d) 10 data sets with imputed values created by *Amelia*: amelia_ldclics5_meandif21.csv', ... , 'amelia_ldclics5_meandif210.csv

2. R-Code: SebastianSteinmueller_BT.r

3. PDF-file of the bachelor thesis: SebastianSteinmueller_BT.pdf

C R-Code

```
##### R-Version 2.5.1 (2007-06-27):
#  R Development Core Team (2007). R: A language and environment for
#  statistical computing. R Foundation for Statistical Computing,
#  Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

# setwd("YourWD")

######  mdgs_komplett.csv  is the original MDGs-data set, downloaded from
######   http://mdgs.un.org/unsd/mdg/Handlers/ExportHandler.ashx?Type=Csv .
####### Name it  mdgs_komplett.csv  and modify it as follows:
######  delete all columns apart from
######  Series ,  Country ,  MDG ,  1990 ,  1991 , ... ,  2009
######  and arrange them in exactly this order.
##### the following code transposes the file to a generic cross section
##### time-series with variables as columns and cases
#####  (year+country) as rows. This file is written to "MDGS_T_02_01.csv"
####  and can be used for the code of the thesis.


mdgs1<-read.csv("mdgs_komplett.csv", na.strings=c("NA",""))


mdgs2<-subset(mdgs1, select=c(-3))

mdgs_split<-split(mdgs2, mdgs2$Country)
 #length(levels(mdgs1$COUNTRY))
countrylist<-list()
for (i in 1:length(levels(mdgs2$Country))) {
#transpose i-th country
  country_t<-data.frame(t(mdgs_split[[i]][,-c(2)]))

  #variable names
  names(country_t)<-mdgs_split[[i]][,1]
#values for  country  and  year
  Country<-rep(mdgs_split[[i]][1,2], length(country_t[,1]))

  country_t2<-cbind(Country, Year=c(0, 1990,   1991,   1992,
   1993   , 1994 ,
  1995  , 1996 ,
  1997 ,  1998,   1999  , 2000 ,  2001   ,2002  , 2003 ,  2004
  ,2005 ,  2006 ,
  2007  , 2008 ,  2009) , country_t)


#cdelete colnames and write split in countrylist s:
  countrylist[[i]]<-country_t2[-1,]



}

mdgs_t<-countrylist[[1]]
for (i in 1:(length(levels(mdgs2$Country))-1)){

  mdgs_t<-rbind(mdgs_t, countrylist[[i+1]])
  #mdgs_t<-cbind(Year=rownames(countrylist[[i+1]]), mdgs_t)

}
write.csv2(mdgs_t, file="MDGS_T_02_01.csv", row.names=TRUE)


###################################################
### chunk number 1:
###################################################
options(width=60)


###################################################
### chunk number 2:
###################################################
#setwd("YourWD")
mdgs_ges1<-read.csv2("MDGS_T_02_01.csv")
```

```
###ldcs and lics according to the DAC 2009


ldcnames<-c("Afghanistan", "Angola", "Bangladesh", "Benin", "Bhutan",
"Burkina Faso", "Burundi", "Cambodia", "Central African Republic",
"Chad", "Comoros", "Democratic Republic of the Congo", "Djibouti",
"Equatorial Guinea", "Eritrea", "Ethiopia", "Gambia", "Guinea",
"Guinea-Bissau",
"Haiti", "Kiribati", "Laos", "Lesotho", "Liberia", "Madagascar",
"Malawi", "Maldives", "Mali", "Mauritania", "Mozambique", "Myanmar",
"Nepal", "Niger", "Rwanda", "Samoa", "Sao Tome and Principe", "Senegal",
 "Sierra Leone", "Solomon Islands", "Somalia", "Sudan", "Tanzania",
"Togo", "Tuvalu", "Uganda", "Vanuatu", "Yemen", "Zambia")

licnames<-c("Cote d Ivoire", "Ghana", "Kenya", "Korea, Democratic
People s Republic of", "Kyrgyzstan", "Nigeria", "Pakistan", "Papua New Guinea",
"Tajikistan", "Uzbekistan", "Viet Nam", "Zimbabwe")

ODA_list_names<-c("Afghanistan", "Cote d Ivoire", "Albania",
"Anguilla","Angola", "Ghana", "Algeria", "Antigua and Barbuda",
"Bangladesh", "Kenya", "Armenia", "Argentina", "Benin", "Korea,
 Democratic People s Republic of", "Azerbaijan", "Barbados",
 "Bhutan" ,
"Kyrgyzstan", "Bolivia", "Belarus", "Burkina Faso", "Nigeria",
 "Bosnia and Herzegovina", "Belize", "Burundi", "Pakistan", "Cameroon",
"Botswana", "Cambodia", "Papua New Guinea", "Cape Verde", "Brazil",
"Central African Republic", "Tajikistan", "China", "Chile", "Chad",
"Uzbekistan", "Colombia", "Cook Islands", "Comoros", "Viet Nam", "Congo",
"Costa Rica", "Democratic Republic of the Congo", "Zimbabwe",
"Dominican Republic", "Croatia", "Djibouti", "Ecuador", "Cuba",
"Equatorial Guinea", "Egypt", "Dominica", "Eritrea", "El Salvador",
"Fiji", "Ethiopia", "Gabon", "Gambia", "Georgia", "Grenada", "Guinea",
 "Guatemala", "Jamaica", "Guinea-Bissau", "Guyana", "Kazakhstan",
"Haiti", "Honduras", "Lebanon", "Kiribati", "India", "Libya", "Laos",
"Indonesia", "Malaysia", "Lesotho", "Iran", "Mauritius", "Liberia",
"Iraq", "Madagascar", "Jordan", "Mexico", "Malawi", "Maldives",
"Marshall Islands", "Mali", "Micronesia, Federated States of", "Nauru",
"Mauritania", "Moldova", "Oman", "Mozambique", "Mongolia", "Palau",
"Myanmar", "Morocco", "Panama", "Nepal", "Namibia", "Serbia", "Niger",
"Nicaragua", "Seychelles", "Rwanda", "Niue", "South Africa", "Samoa",
 "Sao Tome and Principe", "Paraguay", "Saint Kitts and Nevis",
"Senegal", "Peru", "Saint Lucia", "Sierra Leone", "Philippines",
"Saint Vincent and the Grenadines", "Solomon Islands", "Sri Lanka",
"Suriname", "Somalia", "Swaziland", "Trinidad and Tobago" ,"Sudan",
"Syria", "Turkey", "Tanzania", "Thailand", "Uruguay", "Venezuela",
"Togo", "Tonga", "Tuvalu", "Tunisia", "Uganda", "Turkmenistan",
"Vanuatu", "Ukraine", "Yemen", "Zambia")


### without small island development states population < 5 000 000:
ldcnames2<-c("Afghanistan", "Angola", "Bangladesh", "Benin", "Bhutan",
"Burkina Faso", "Burundi", "Cambodia",  "Central African Republic",
"Chad",  "Democratic Republic of the Congo",   "Eritrea", "Ethiopia",
"Guinea",
"Haiti",  "Laos", "Lesotho", "Liberia", "Madagascar", "Malawi",  "Mali",
"Mauritania", "Mozambique", "Myanmar",
"Nepal", "Niger", "Rwanda",  "Sierra Leone",  "Senegal", "Somalia",
"Sudan", "Tanzania",
"Togo",  "Uganda",  "Yemen", "Zambia")

licnames2<-c("Ghana", "Kenya", "Korea, Democratic People s Republic of",
 "Kyrgyzstan", "Nigeria", "Pakistan", "Papua New Guinea",
"Tajikistan", "Uzbekistan", "Viet Nam", "Zimbabwe")


ODA_list_names2<-c("Afghanistan", "Cote d Ivoire", "Albania", "Angola",
"Ghana", "Algeria", "Bangladesh", "Kenya", "Armenia",
"Argentina", "Benin", "Korea, Democratic People s Republic of", "Azerbaijan",
 "Bhutan" ,
"Kyrgyzstan", "Bolivia", "Belarus", "Burkina Faso", "Nigeria",
"Bosnia and Herzegovina", "Burundi", "Pakistan", "Cameroon",
"Botswana", "Cambodia", "Papua New Guinea",  "Brazil",
"Central African Republic", "Tajikistan", "China", "Chile", "Chad",
"Uzbekistan", "Colombia", "Viet Nam", "Congo", "Costa Rica",
"Democratic Republic of the Congo", "Zimbabwe",
"Dominican Republic", "Croatia", "Djibouti", "Ecuador", "Cuba",
"Equatorial Guinea", "Egypt",  "Eritrea", "El Salvador",
 "Ethiopia", "Gabon", "Gambia", "Georgia",  "Guinea", "Guatemala",
 "Kazakhstan", "Haiti", "Honduras", "Lebanon",
 "India", "Libya", "Laos", "Indonesia", "Malaysia", "Lesotho",
 "Iran", "Liberia",
"Iraq", "Madagascar", "Jordan", "Mexico", "Malawi", "Mali",
"Mauritania", "Moldova", "Oman", "Mozambique", "Mongolia",
"Myanmar", "Morocco",
"Panama", "Nepal", "Namibia", "Serbia", "Niger",
"Nicaragua", "Rwanda",  "South Africa",  "Paraguay", "Senegal",
 "Peru",  "Sierra Leone", "Philippines", "Sri Lanka",
 "Somalia", "Swaziland", "Sudan", "Syria", "Turkey", "Tanzania",
 "Thailand", "Uruguay", "Venezuela",
"Togo",   "Tunisia", "Uganda", "Turkmenistan",  "Ukraine", "Yemen", "Zambia")

### sub-saharan Africa:
ssanames<-c("Benin", "Burkina Faso", "Ghana", "Madagascar",
"Malawi",  "Mali", "Mauritania", "Senegal", "Cameroon",
"Central African Republic", "Chad", "Congo", "Cote d Ivoire",
 "Eritrea", "Ethiopia", "Guinea", "Kenya", "Mozambique",
 "Niger", "Nigeria",  "Rwanda",  "Sierra Leone", "Tanzania",
  "Togo", "Uganda", "Zambia", "Angola", "Burundi",
   "Democratic Republic of the Congo",
 "Sudan", "Zimbabwe", "Liberia", "Somalia")


ldclics<-subset(mdgs_ges1, Country %in% ODA_list_names2)
attach(ldclics)




######################### functions for the proportion of
###                       missings by variables:

library(Amelia)
library(VIM)

propmis<-function(x){
     propm<-matrix(data=0, nrow=length(colnames(x)), ncol=1)
     rownames(propm)=colnames(x)
     r<-length(rownames(x))
      for(i in 1:length(colnames(x))){
          ifelse(names(table(is.na(x[,i])))=="TRUE",
          propm[i,1]<-(as.numeric(table(is.na(x[,i]))[1]))/r,
          propm[i,1]<-(r-as.numeric(table(is.na(x[,i]))[1]))/r)
          }
          help1<-data.frame(colnames(x), propm)
          sorted1<-help1[order(propm),]
          #sorted1<-as.data.frame(propm)[order(propm),]
          return(list(unsorted=propm, sorted=subset(sorted1, select=2)))
          }



###########  function for cross- section time series, missings by Country
######## and Year in all variables, y is Year:
########## propmis_split$Year_Country: matrix with the proportion of missings
##### in a specific year for a country
######### across all the variables in the data set,
######### propmis_split$Country: missingness per country over all the
###### years and variables,
######## propmis_split$Year: missingness per Year over all the
######## countries and variables

propmis_split<-function(x, y){
      xsplit<-split(x, y)
      p<-length(colnames(xsplit[[1]]))
      .countryvec<-vector(length=length(rownames(xsplit[[1]])))
      yearmatrix<-matrix(data=0, nrow=length(rownames(xsplit[[1]])),
      ncol=length(xsplit))
       colnames(yearmatrix)=unique(x[,which(colnames(mdgs_ges1)=="Year")])
       rownames(yearmatrix)=xsplit[[1]][,which(colnames(mdgs_ges1)=="Country")]
      for(i in 1:length(xsplit)){
          for(k in 1:length(rownames(xsplit[[i]]))){
              .countryvec[k]<-as.numeric(table(is.na(xsplit[[i]][k,]))[2])/p
              }
          yearmatrix[,i]<-.countryvec
          }
          return(list(Year_Country=yearmatrix, Country=rowMeans(yearmatrix),
          Year=colMeans(yearmatrix)))
}



###function which displays the missings in all the variable for each year
#### (Year and Country have to be included in the dataframe
###### "dat" as variables) :

propmis_varyear<-function(dat){
    propm_varyear<-data.frame(rep(0,length(unique(dat$Year))))
    for(f in 1:ncol(dat)){
        vardf<-data.frame()
        for (i in 1:length(unique(dat$Country))){
              vardf<-rbind(vardf, subset(dat[,f],
              dat$Country==unique(dat$Country)[i]))}
        rownames(vardf)<-unique(dat$Country)
        colnames(vardf)<-unique(dat$Year)

        propm<-matrix(data=0, nrow=length(colnames(vardf)), ncol=1)
           rownames(propm)=colnames(vardf)
           r<-length(rownames(vardf))
            for(i in 1:length(colnames(vardf))){
                ifelse(names(table(is.na(vardf[,i])))=="TRUE",
                propm[i,1]<-(as.numeric(table(is.na(vardf[,i]))[1]))/r,
                propm[i,1]<-(r-as.numeric(table(is.na(vardf[,i]))[1]))/r)
                }

        propm_varyear[,f]<-propm

    }
    colnames(propm_varyear)<-colnames(dat)
    rownames(propm_varyear)<-rownames(propm)
  return(propm_varyear)
}


### creating data set to compare proportion of missings per
##  Country to various variables:
```

```
propmis_split_ges<-propmis_split(mdgs_ges1, Year)

missings_Year_Country<-propmis_split_ges$Year_Country

missings2006<-missings_Year_Country
[,which(colnames(missings_Year_Country)==  2006 )]



##### download the .csv (semicolon) from
#####  http://data.un.org/Data.aspx?q=GDP+per+capita+2006&d=CDB&f=srID%3A29922%3Byr%3A2006
##### and name it  GDP_Capita_2006.csv
gdp_capita_2006<-read.table("GDP_Capita_2006.csv",
 sep=";", dec=",", header=T, row.names=1)
colnames(gdp_capita_2006)<-c( Year  ,
 GDP.per.Capita.in.PPP ,  Value.Footnotes )

missings_gdp_2006<-merge(gdp_capita_2006, missings2006, by="row.names")
row.names(missings_gdp_2006)<-missings_gdp_2006$Row.names
colnames(missings_gdp_2006)<-c(colnames(missings_gdp_2006)[1:4],
  missings2006 )


enrolment_intUsers_2006<-subset(mdgs_ges1, Year==2006,
 select=c(2,3,35,90,135))
row.names(enrolment_intUsers_2006)<-enrolment_intUsers_2006[,1]

hiv_prev_2007<-subset(mdgs_ges1, Year==2007, select=c(2,70))
row.names(hiv_prev_2007)<-hiv_prev_2007[,1]

explain_missings1<-merge(enrolment_intUsers_2006,
missings_gdp_2006, by="row.names")
row.names(explain_missings1)<-explain_missings1[,2]

explain_missings<-merge(explain_missings1,
 hiv_prev_2007,  by="row.names")
row.names(explain_missings)<-explain_missings[,1]
```

```
###################################################
### chunk number 3:
###################################################

ldclics4<-subset(ldclics, select=c(1,2,3,4,5,8,10,11,15,19,20,
21,33,c(35:64),70,73,74,75,76,79,80,81,82,
c(83:90),106,114,117,130,131,133,135), Year %in% c(1990:2008))

ldclics5<-ldclics4[,c(2,3,4,11,20,29,35,38,44,60,61,62,63,67)]
colnames_ldclics5_1<-c("Country", "Year", "Population below 1$
PPP per day percentage",
"Children under 5 severely underweight percentage",
"Primary completion rate both sexes",
"Gender Parity Index in tertiary level enrolment, percentage",
"Children under five mortality rate per 1000 live births",
"Maternal mortality ratio per 100000 live births",
"People living with HIV, 15-49 year olds, percentage",
"Tuberculosis prevalence rate per 100000 population",
"Population using solid fuels, percentage",
"Proportion of the population using improved drinking water sources, total",
"Proportion of the population using improved sanitation facilities, total",
"Internet users per 100 population"
)
colnames(ldclics5)<-colnames_ldclics5_1


### add GDP / Capita

##### download the semicolon-separated file from
##### http://data.un.org/Data.aspx?q=gdp+capita&d=SNAAMA&f=grI
##### D%3a101%3bcurrID%3aUSD%3bpcFlag%3a1 and name it
##### "GDPPC_USD_countries90_08.csv"
##### For it to work here, you have to modify it:
##### delete all comments which are not part of the actual data and
###  keep only the years from 1990 to 2008.
##### You will also have to change some of the country names in the file
###  to match the names from the DAC-list (for example China)
```

```
gdp90_08<-read.csv2("GDPPC_USD_countries90_08.csv",row.names=1)
colnames(gdp90_08)<-1990:2008

ldclics5<-data.frame(ldclics5, GDP=numeric(length=length(rownames(ldclics5))))

for (i in 1:length(ldclics5$GDP)){
                  if(ldclics5$Country[i]%in%rownames(gdp90_08)){
                  l<-which(rownames(gdp90_08)==ldclics5$Country[i])
                  y<-which(colnames(gdp90_08)==ldclics5$Year[i])
                  ldclics5$GDP[i]<-as.numeric(gdp90_08[l,y])}
                  else{ldclics5$GDP[i]<-NA}
                  }

ldclics5<-data.frame(ldclics5[,c(1:5)], GPI3=ldclics5[,6]*100,
 ldclics5[,7:14], GDP=ldclics5[,15]/100)



########
colnames_ldclics5_2<-c("Country", "Year", "Dollar.Pov.",
"Underweight",
"Prim.Compl",
"GPI3",
"CM",
"MM",
"HIV",
"TBC",
"Fuels",
"Water",
"Sanitation",
"Internet",
"GDP"
)

N<-vector("numeric")
N<-for(i in 3:length(colnames(ldclics5))){N<-c(N, as.numeric
(table(is.na(ldclics5[,i]))[1]))}
```

```
corvec<-vector("numeric")
corvec<-for(i in 3:length(colnames(ldclics5))){corvec<-c(corvec,
 cor(ldclics5[,i], ldclics5$Primary.completion.rate.both.sexes,
use="complete.obs", method="pearson"))}
table_ldclics<-data.frame(    = c(colnames_ldclics5_1[c(-1,-2)],
 Per capita GDP at current prices - US dollars / 100 ), "Abbr."=
colnames_ldclics5_2[c(-1,-2)],N, "Mean"=mean(ldclics5[,c(3:15)],
 na.rm=T), "SD"=sd(ldclics5[,c(3:15)],
 na.rm=T), "Prop. missings"=propmis(ldclics5[,c(3:15)])$unsorted,
 "Correlation with Prim Compl"=corvec )
colnames(table_ldclics)<-c(    , "Abbr.", "N", "Mean",   "Sd",
 "% missing values" ,"Correlation with Prim. Compl")
library(xtable)


###################################################
### chunk number 4:
###################################################
xtable_ldclics<-xtable(table_ldclics, digits=c(0,0,0,0,2,2,2,2),
caption="Overview of the relevant variables", label="tab1")
align(xtable_ldclics)<-"|m{5cm}|m{3cm}m{2cm}m{2cm}m{2cm}
m{2.5cm}m{2.5cm}|r|"
print(xtable_ldclics, hline.after=c(-1:nrow(table_ldclics)),
floating.environment = "sidewaystable", floating.placement="h!",
 include.rownames=FALSE)
colnames(ldclics5)<-colnames_ldclics5_2



###################################################
### chunk number 1:
###################################################

par(mar=c(8,10,8,2), mgp=c(6, 1, 0), cex=1.4)
plot(explain_missings$missings2006˜
log(explain_missings$GDP.per.Capita.in.PPP),
 cex.axis=1.5, cex.main=2.2, cex.sub=2, cex.lab=2,
xlab="Log. GDP / capita (in PPP)\n", ylab="Proportion of
missings\nover the MDGs variables",  sub="158 countries, 2006")
```

```
lm_missings_gdp<-lm(explain_missings$missings2006~
log(explain_missings$GDP.per.Capita.in.PPP))

abline(coef=lm_missings_gdp$coef, col=  red )


###################################################
### chunk number 2:
###################################################

#### % missings on HIV prevalence rate 2007

missings2007<-missings_Year_Country[,18]
missings2007<-as.data.frame(missings2007)
row.names(missings2007)<-rownames(missings_Year_Country)

hiv_prev2007<-explain_missings$People.living.with.
HIV..15.49.years.old..percentage
hiv_prev2007<-as.data.frame(hiv_prev2007)
row.names(hiv_prev2007)<-rownames(explain_missings)

hiv_missings2007<-merge(missings2007, hiv_prev2007, by="row.names")

par(mar=c(8,10,8,2), mgp=c(5.5, 1, 0), cex=1.4)
plot(hiv_missings2007$missings2007~hiv_missings2007$hiv_prev2007,
 cex.axis=1.5, cex.main=2.2, cex.sub=2, cex.lab=2,
xlab="HIV-Prevalence, 15-49 year olds\n", ylab="Proportion of
missings\nover the MDGs variables", sub="158 countries, 2007")

lm_missings_hiv<-lm(hiv_missings2007$missings2007~
hiv_missings2007$hiv_prev2007)

abline(coef=lm_missings_hiv$coef, col=  red )


###################################################
### chunk number 3:
###################################################
```

```
### cellular subscribers in developing countries:
split_year<-split(ldclics, Year)

phones_mean<-vector(mode=  numeric , length= length(split_year))
for(i in 1:length(split_year)){phones_mean[i]<-
mean(split_year[[i]]$
Mobile.cellular.telephone.subscriptions.per.100.population,
 na.rm=T)}
par(mar=c(8,10,8,2), mgp=c(6, 1, 0), cex=1.4, pch=19)
plot(phones_mean~c(1990:2009), type=  p , col= blue , cex=2, cex.axis=1.5,
cex.main=2.2, cex.sub=2, cex.lab=2, xlab="Year\n",
ylab="Cellular subscribers per 100 population",
sub="LDCs/LICs according to DAC (2009)")


###################################################
### chunk number 4:
###################################################

##### possible influences on data availability:
lm_missings2<-
lm(explain_missings$missings2006~explain_missings$GDP.per.Capita.in.PPP+
explain_missings$People.living.with.HIV..15.49.years.old..percentage+
explain_missings$Total.net.enrolment.ratio.in.primary.education..both.sexes)


###################################################
### chunk number 5:
###################################################
par(cex=1.2, cex.main=1.7, mar=c(5,4,2,2))
aggr(ldclics5[,c(-1,-2)], numbers=T, cex.axis=0.9, cex.lab=1.4)


###################################################
### chunk number 6:
###################################################
```

```
propmis_varyear5<-propmis_varyear(ldclics5)


########## next two figures: time series of the proportion
########## of missingness in specific variables:

par(mar=c(5, 4, 4, 14)+0.1, xpd=T, cex=1.8, cex.lab=1.4)
plot(propmis_varyear5$"CM"~rownames(propmis_varyear5), xlab="Year",
 ylab="Proportion of missings", type="o", col=1)
points(y=propmis_varyear5$"Prim.Compl",x=rownames(propmis_varyear5),
type="o", col=2)
points(y=propmis_varyear5$"Dollar.Pov.",
x=rownames(propmis_varyear5), type="o", col=3)
points(y=propmis_varyear5$"Underweight",
x=rownames(propmis_varyear5), type="o", col=4)
points(y=propmis_varyear5$"MM",
x=rownames(propmis_varyear5), type="b", col=5)
points(y=propmis_varyear5$"GDP",
x=rownames(propmis_varyear5), type="o", col=6)
points(y=propmis_varyear5$"GPI3",
x=rownames(propmis_varyear5), type="o", col=7)

legend(x=2011,y=0.33,legend=c("CM", "Prim.Compl",
"Dollar.Pov.", "Underweight", "MM", "GDP", "GPI3"), fill=c(1:7))


###################################################
### chunk number 7:
###################################################
par(mar=c(5, 4, 4, 13)+0.1, xpd=T, cex=1.8, cex.lab=1.4)

plot(propmis_varyear5$"HIV"~rownames(propmis_varyear5),
xlab="Year", ylab="Proportion of missings", type="o", col=3, ylim=c(0,1))
points(y=propmis_varyear5$"Fuels",x=rownames(propmis_varyear5), type="b", col=6)
points(y=propmis_varyear5$"Water",x=rownames(propmis_varyear5), type="o", col=2)
points(y=propmis_varyear5$"Sanitation",x=rownames(propmis_varyear5),
type="b", col=1)
```

```
points(y=propmis_varyear5$"Internet",x=rownames(propmis_varyear5),
type="b", col=5)

legend(x=2011,y=0.22,legend=c("HIV", "Fuels", "Water", "Sanitation",
 "Internet"), fill=c(3,6,2,1,5))


###################################################
### chunk number 8:
###################################################

ldclics5_meandif<-data.frame()


for(i in 1:length(unique(ldclics5$Country))){
    for(k in 3:length(colnames(ldclics5))){
      l<-k-2
        co<-unique(ldclics5$Country)[i]
        ldclics5_meandif[i,l]<-mean(subset(subset(ldclics5,
         Year %in% 2000:2008),
Country==co, select=k), na.rm=T)-mean(subset(subset(ldclics5,
 Year %in% 1990:1999),
Country==co, select=k), na.rm=T)
        }
    }


colnames(ldclics5_meandif)<-colnames(ldclics5[c(-1,-2)])
rownames(ldclics5_meandif)<-unique(ldclics5$Country)

write.table(ldclics5_meandif, file="ldclics5_meandif.csv",
sep=";", dec=",", col.names=NA, row.names=TRUE)

ldclics5_meandif<-read.csv2("ldclics5_meandif.csv", row.names=1)

N<-vector("numeric")
N<-for(i in 1:length(colnames(ldclics5_meandif)))
{if(names(table(is.na(ldclics5_meandif[,i])))=="TRUE")
{N<-c(N,0)}
else { N<-c(N,as.numeric(table(is.na(ldclics5_meandif[,i]))[1]))}
}

corvec<-vector("numeric")
corvec<-for(i in 1:length(colnames(ldclics5_meandif))){corvec<-c(corvec,
cor(ldclics5_meandif[,i], ldclics5_meandif$Prim.Compl,
use="complete.obs", method="pearson"))}
table_ldclics2<-data.frame(    = c(colnames_ldclics5_1[c(-1,-2)],
  Per capita GDP at current prices - US dollars / 100 ),
"Abbr."=colnames_ldclics5_2[c(-1,-2)],N,
"Mean"=mean(ldclics5_meandif, na.rm=T), "SD"=sd(ldclics5_meandif,
 na.rm=T), "Prop. missings"=propmis(ldclics5_meandif)$unsorted,
  "Correlation with Prim Compl"=corvec )
colnames(table_ldclics2)<-c(    , "Abbr.", "N", "Mean",   "Sd",
"% missing values" ,"Correlation with Prim. Compl")



###################################################
### chunk number 9:
###################################################
xtable_ldclics2<-xtable(table_ldclics2, digits=c(0,0,0,0,2,2,2,2),
caption="Overview of the modified data set", label="tab3")
align(xtable_ldclics2)<-"|m{5cm}|m{3cm}m{2cm}m{2cm}m{2cm}m{2.5cm}m{2.5cm}|r|"
print(xtable_ldclics2, hline.after=c(-1:nrow(table_ldclics)),
floating.environment = "sidewaystable", floating.placement="h!",
include.rownames=FALSE, NA.string=  NA )


###################################################
### chunk number 10:
###################################################
par(cex=1.5, cex.main=1.7, mar=c(5,4,4,2))
aggr(ldclics5_meandif[c(-6,-7)], numbers=T, cex.lab=1.25, cex.axis=0.87)


###################################################
### chunk number 11:
###################################################
```

```
#### comparison of the distribution of GDP/capita conditional on the
#####  observation status in Prim.Compl

mis_primcompl<-vector("integer")
for(i in 1:nrow(ldclics5_meandif)){
ifelse(is.na(ldclics5_meandif$Prim.Compl[i]), mis_primcompl[i]<-0,
 mis_primcompl[i]<-1) }

ldclics5_mis<-data.frame(ldclics5_meandif, Missings.Prim.Compl=mis_primcompl)

boxplot(ldclics5_mis$GDP~ldclics5_mis$Missings.Prim.Compl, ylim=c(-5, 20),
names=c(  Missing in Prim.Compl  ,  Observed in Prim.Compl ),
ylab=  GDP / capita , col="beige")


###################################################
### chunk number 12:
###################################################

##### t- tests for MAR of the variable Prim.Compl with various variables
######  (H0: there are no differences between the mean of the cases for
which Prim.Compl is missing
#######   and the mean of the cases where Prim.Compl is observed)


ttest_gdp<-t.test(x=split(ldclics5_mis$GDP, ldclics5_mis$Missings.Prim.Compl)
[[1]], y=split(ldclics5_mis$GDP, ldclics5_mis$Missings.Prim.Compl)[[2]],
 alternative="two.sided", var.equal=FALSE)

ttest_internet<-t.test(x=split(ldclics5_mis$Internet,
ldclics5_mis$Missings.Prim.Compl)[[1]], y=split(ldclics5_mis$Internet,
ldclics5_mis$Missings.Prim.Compl)[[2]], alternative="two.sided",
var.equal=FALSE)

ttest_cm<-t.test(x=split(ldclics5_mis$CM, ldclics5_mis$Missings.Prim.Compl)
[[1]], y=split(ldclics5_mis$CM, ldclics5_mis$Missings.Prim.Compl)[[2]],
 alternative="two.sided", var.equal=FALSE)
```

```
ttest_gpi3<-t.test(x=split(ldclics5_mis$GPI3,
ldclics5_mis$Missings.Prim.Compl)[[1]],
 y=split(ldclics5_mis$GPI3, ldclics5_mis$Missings.Prim.Compl)[[2]],
 alternative="two.sided", var.equal=FALSE)

ttest_dollar<-t.test(x=split(ldclics5_mis$Dollar.Pov.,
ldclics5_mis$Missings.Prim.Compl)[[1]], y=split(ldclics5_mis$Dollar.Pov.,
ldclics5_mis$Missings.Prim.Compl)[[2]], alternative="two.sided",
var.equal=FALSE)

ttest_underweight<-t.test(x=split(ldclics5_mis$Underweight,
ldclics5_mis$Missings.Prim.Compl)[[1]], y=split(ldclics5_mis$Underweight,
ldclics5_mis$Missings.Prim.Compl)[[2]], alternative="two.sided",
var.equal=FALSE)



###################################################
### chunk number 1:
###################################################

#### imputation by the unconditional mean:

impmean<-ldclics5_meandif
for (i in 1:length(colnames(impmean))){
    for(k in 1:length(rownames(impmean))){
        if(is.na(impmean[k,i])) {impmean[k,i]<-
        mean(ldclics5_meandif[,i], na.rm=T)}
        }
    }
impmean_lm1<-lm(impmean$Prim.Compl~impmean$Underweight+
impmean$"Dollar.Pov."+impmean$GPI3+impmean$CM+impmean$Fuels+
impmean$Sanitation+impmean$Water+impmean$Internet+impmean$GDP,
 data=impmean)
impmean_lm2<-lm(Prim.Compl~Dollar.Pov.+Underweight+GPI3+CM+Water+
Sanitation+Internet+GDP, data=impmean)


###################################################
```

```
### chunk number 2:
###################################################

##### imputation by Buck s method:
ldclics5_meandif2<-ldclics5_meandif[,c(-6,-7, -9)]


bucks<-function(x){
  x2<-x
  reglist<-list()
  coveclist<-list()
  m<-1
  for(i in 1:nrow(x)){

    for(k in 1:ncol(x)){
        if(is.na(x[i,k])){
            covec.i<-vector("character")
            for(g in 1:ncol(x)){
                if(!is.na(x[i,g])){if(g!=k){covec.i<-
                c(covec.i, colnames(x)[g])}}}
            covec2.i<-paste(covec.i, collapse="+")
            covec3.i<-paste(colnames(x)[k], covec2.i, sep=" ˜ ")

            if(covec3.i %in% coveclist){
            reg.k<-reglist[[which(coveclist==covec3.i)]]
            }
            else{
            reg.k<-lm(formula=covec3.i, data=x)
            reglist[[m]]<-reg.k
            coveclist[[m]]<-covec3.i
            m<-m+1
            }

            x2[i,k]<-predict(reg.k, x)[i]

            }
            }
    } #i
    return(x2)
```

```
   } #end function


ldclics5_reg<-bucks(ldclics5_meandif2)


### linear model using data set imputed by buck s method:

buck_lm<-lm(Prim.Compl~., data=ldclics5_reg[,c(-6)])


###################################################
### chunk number 3:
###################################################


amelia_ldclics5_meandif2<-amelia(ldclics5_meandif2, m=10,
outname="amelia_ldclics5_meandif2", write.out=T)


areglist<-list(0)
for(i in 1:10){
reg.i<-lm(Prim.Compl~., data=amelia_ldclics5_meandif2[[i]][,c(-6)])
areglist[[i]]<-reg.i
}


reglist_mean<-matrix(nrow=9, ncol=4, data=0)
for(i in 1:10){
  reglist_mean<-(reglist_mean+summary(areglist[[i]])$coef)}

reglist_mean<-reglist_mean/10



betamat<-matrix(nrow=9, ncol=10, data=0)
for(i in 1:10){
betamat[,i]<-summary(areglist[[i]])$coef[,1]}
```

```
crossvar<-apply(betamat, 1, var)

complvar<-(reglist_mean[,2])^2+crossvar*(1+1/10)

amelia_reg<-matrix(nrow=9, ncol=4, data=0)
amelia_reg[,1]<-reglist_mean[,1]
amelia_reg[,2]<-sqrt(complvar)
amelia_reg[,3]<-amelia_reg[,1]/amelia_reg[,2]
amelia_reg[,4]<-(1-pt(abs(amelia_reg[,3]), df=94))*2
colnames(amelia_reg)<-c("Estimate", "Std. Error", "t value", "Pr(>|t|)")
rownames(amelia_reg)<-c("(Intercept)",
 colnames(amelia_ldclics5_meandif2[[i]][,c(-3,-6)]))
radj_ame<-0
for(i in 1:10){radj_ame<-radj_ame+mean(summary(areglist[[i]])$adj.r)}
radj_ame<-radj_ame/10


###################################################
### chunk number 4:
###################################################
#detach(ldclics)
#attach(ldclics5)
meandif_lm1<-lm(ldclics5_meandif$"Prim.Compl"~
Underweight+Dollar.Pov.+ldclics5_meandif$"GPI3"+ldclics5_meandif$CM
+ldclics5_meandif$Fuels+ldclics5_meandif$Sanitation+ldclics5_meandif$Water+
ldclics5_meandif$Internet+ldclics5_meandif$GDP, data=ldclics5_meandif)


###################################################
### chunk number 5:
###################################################
meandif_lm2<-lm(Prim.Compl~Dollar.Pov.+Underweight+GPI3+CM+
Water+Sanitation+Internet+GDP, data=ldclics5_meandif)
 ra<-summary(meandif_lm2)$adj.r
xa<-c(ra, NA,NA,NA)
names(xa)<-c(  Estimate ,   Std. Error ,   t value ,   PR(>|t|) )
lm2_table<-rbind(summary(meandif_lm2)$coefficients, xa)
lm2_table<-rbind(lm2_table, c(NA,NA,NA,NA))
```

```
rownames(lm2_table)[11]=  N = 33
rownames(lm2_table)[10]=  adj. R-Square =


###################################################
### chunk number 6:
###################################################
comptable<-matrix(nrow=4, ncol=4, data=0)

comptable[1,1]<-mean(abs(summary(meandif_lm2)$coef[c(2:9),1]))
comptable[1,2]<-mean(abs(summary(meandif_lm2)$coef[c(2:9),2]))
comptable[1,3]<-mean(abs(summary(meandif_lm2)$coef[c(2:9),3]))
comptable[1,4]<-mean(abs(summary(meandif_lm2)$coef[c(2:9),4]))

comptable[2,1]<-mean(abs(summary(impmean_lm2)$coef[c(2:9),1]))
comptable[2,2]<-mean(abs(summary(impmean_lm2)$coef[c(2:9),2]))
comptable[2,3]<-mean(abs(summary(impmean_lm2)$coef[c(2:9),3]))
comptable[2,4]<-mean(abs(summary(impmean_lm2)$coef[c(2:9),4]))

comptable[3,1]<-mean(abs(summary(buck_lm)$coef[c(2:9),1]))
comptable[3,2]<-mean(abs(summary(buck_lm)$coef[c(2:9),2]))
comptable[3,3]<-mean(abs(summary(buck_lm)$coef[c(2:9),3]))
comptable[3,4]<-mean(abs(summary(buck_lm)$coef[c(2:9),4]))

comptable[4,1]<-mean(abs(amelia_reg)[c(2:9),1])
comptable[4,2]<-mean(abs(amelia_reg)[c(2:9),2])
comptable[4,3]<-mean(abs(amelia_reg)[c(2:9),3])
comptable[4,4]<-mean(abs(amelia_reg)[c(2:9),4])

row.names(comptable)<-c("Listwise deletion","Imp. by mean",
"Buck s method","EM + MI")
colnames(comptable)<-c("Coef.","sd","t","p")


###################################################
### chunk number 7:
###################################################
xcomptable<-xtable(comptable, digits=c(0,3,3,3,3),
caption="Arithmetic means of the linear models over all
variables", label="tab6")
align(xcomptable)<-"|l|rrrr|"
print(xcomptable, floating.placement="h")


###################################################
### chunk number 8:
###################################################
par(mar=c(10,7,7,2), mgp=c(5, 1, 0), cex=1.4)
plot(ldclics5_meandif$Dollar.Pov., ldclics5_meandif$Prim.Compl,
cex.axis=1.5, cex.main=2.5, cex.sub=2.2, cex.lab=2.2,
 pch=19, col=4, cex=1.3,
xlab="Difference in the percentage of people living on < 1 US-\$
 (PPP) / day", ylab=  Difference in primary completion rate )
```

Affidavit

I hereby confirm that I prepared this bachelor thesis independently, by exclusive reliance on the literature and tools indicated therein.

Munich, 28 July 2010

Sebastian Steinmüller

Eidesstattliche Erklärung

Hiermit erkläre ich, dass es sich bei der vorliegenden Bachelorarbeit um eine selbständig verfasste Arbeit handelt und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt wurden.

München, den 28. Juli 2010

Sebastian Steinmüller