Charles University in Prague

Faculty of Social Sciences
Institute of Economic Studies

MASTER'S THESIS

Education and HIV: Evidence from Sub-Saharan Africa

Author: **Bc. Tadeáš Kopecký**

Supervisor: **doc. PhDr. Julie Chytilová, Ph.D.**

Academic Year: **2016/2017**

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain a different or the same degree.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, January 5, 2017

Signature

Acknowledgments

I would like to express my deepest gratitude to my supervisor, doc. PhDr. Julie Chytilová, Ph.D., for providing me with valuable and useful advise during the whole process of writing the thesis. Also I would like to thank to the Demographic and Health Survey for providing me with the data. Last but not least I deeply thank to my family and friends. Without their support this thesis could have never been written.

Abstract

The HIV/AIDS epidemic remains a large threat for developing countries, especially for Sub-Saharan Africa. To be able to fight the epidemic, we need to understand the socio-economic drivers of it to distinguish the groups of people at the highest risk of the HIV. We performed an econometric analysis using logistic regression dealing with the relationship between education and several HIV connected factors - HIV status, HIV knowledge and sexual behavior - based on a large sample from 21 Sub-Saharan African countries from Demographic and Health Survey data collection from years 2008-2014. The education appears to be non-lineary correlated with the HIV status as people with primary and secondary education are at the highest risk of being HIV positive. These results can be nevertheless influenced by e.g. survivorship bias as the education appears to have a positive effect on both HIV knowledge and protective sexual behavior. It is thus advised to promote education in the Sub-Saharan Africa. At the same time, it is needed to target the help primary to the groups at the highest risk of being HIV positive to prevent further spread of HIV and to help families of the HIV positive individuals. Moreover, we found that there is no significant difference in the correlation between education and HIV status between rich and poor and high and low HIV prevalence countries.

JEL Classification	C21, I12, I15, I21,
Keywords	HIV, AIDS, education, sexual behavior, Africa

Author's e-mail	`tadeas.kopecky@gmail.com`
Supervisor's e-mail	`julie.chytilova@fsv.cuni.cz`

Abstrakt

Epidemie HIV/AIDS představuje velkou hrozbu pro rozvojové země, zejména pro subsaharskou Afriku. Aby bylo možné proti epidemii efektivně bojovat, je nutné pochopit sociálně-ekonomické faktory epidemie k určení nejrizikovějších skupin lidí. V práci jsme provedli ekonometrickou analýzu (logistická regrese) zabývající se vztahem mezi vzděláním a několika faktory spojených s HIV - HIV statusem, znalostí HIV a sexuálním chováním - založenou na vzorku 21 zemí ze subsaharské Afriky z sběru dat z let 2008-2014 organizací Demographic and Health Survey. Objevili jsme nelineání korelaci mezi vzděláním a HIV statusem, kdy lidé se základním a středním vzděláním mají nejvyšší riziko být HIV pozitivní. Výsledky mohou být nicméně ovlivněny několika faktory, zejména tzv. survivorship bias. Dále bylo zjištěno, že vzdělání má pozitivní vliv na znalost HIV i na sexuální chování jednotlivců. Je tedy potřebné dále cíleně podporovat vzdělávání v subsaharské Africe. Současně je třeba cílit pomoc na skupiny lidí, které mají nejvyšší riziko být HIV pozitivní s cílem zamezit dalšímu šíření HIV a poskytnout pomoc rodinám s HIV pozitivními jedinci. Dále jsme zjistili, že není žádný signifikantní rozdíl v korelaci mezi vzděláním a HIV statusem mezi státy s vysokou a nízkou prevalencí HIV a mezi bohatými a chudými státy.

Klasifikace JEL	C21, I12, I15, I21,
Klíčová slova	HIV, AIDS, vzdělání, sexuální chování, Afrika

E-mail autora	`tadeas.kopecky@gmail.com`
E-mail vedoucího práce	`julie.chytilova@fsv.cuni.cz`

Contents

List of Tables

List of Figures

Acronyms

AIDS Acquired immune deficiency syndrome

AME Average marginal effect

ART Antiretroviral treatment

DHS Demographic and Health Survey

GDP Gross domestic product

GNI Gross natinal income

HIV Human immunodeficiency virus

MAE Marginal effect on average

MLE Maximul likelihood estimation

OLS Ordinary least squares

UNAIDS Joint United Nations Programme on HIV/AIDS

Master's Thesis Proposal

Author	Bc. Tadeáš Kopecký
Supervisor	doc. PhDr. Julie Chytilová, Ph.D.
Proposed topic	Education and HIV: Evidence from Sub-Saharan Africa

Motivation The World Health Organization reports 35 million people infected by the HIV worldwide in 2013. From all infected people, 71% live in Sub-Saharan Africa. Only in 2013, 1.5 million deaths were connected to the HIV in that area. Although the antiretroviral therapy can help people to have a decent life even if they are infected by the virus, it does not completely prevent the virus to spread. Moreover, only 11.7% of people HIV positive people in low and middle income countries have access to the treatment. Therefore, there is a need to fight the virus also in different ways than medically, e.g. through prevention programmes or by improving factors that encourages the virus to spread.

It was shown by David and Li (2008) that HIV infection has a negative impact on social capital of the country. Mwakolabo (2007) shows that being HIV positive increases a chance of an individual to fall into the poverty. Walker (2002) estimates a decline in income of between 48 and 78 percent when one member of the household dies from HIV/AIDS. Ka?ová (2005) shows on the example from Namibia that AIDS/HIV infection has devastating effect on the development of the country's economy. Also, Greener (2002) points out that the high mortality of young HIV infected people reduces the taxable population and, as a consequence slow down overall economy of the country. Therefore, lowering the HIV infection could have a positive effect on poor Sub-Saharan countries' overall development. To be able to fight the infection, we need to understand the correlation between the HIV status and socioeconomic status of the individuals. Then, we can target the help and prevention programmes more efficiently to specific groups of people. The analyses concerning this topic were limited in the past due to the data availability. With current possibilities of cheap and quick HIV testing, we can find large datasets from many developing countries.

In the thesis, we want to focus mainly on the correlation between the HIV status and the level of education. The findings about this correlation are mixed. There

is an evidence of a positive correlation, e.g. Gregson et al. (2001). On the other hand, we can find many papers suggesting negative correlation, e.g. Glynn et al. (2004). Moreover, we can find papers that suggest no significant correlation, e.g. de Walque (2008). Nevertheless, many of these analyses were done with limited datasets. We want to further analyze the correlation by using larger dataset from more Sub-Saharan African countries compared to the previous work. Mainly, we want to analyze how the strength of the correlation varies across the countries with different overall HIV prevalence or with the different level of GDP (e.g. in the analysis of five African countries by Forston (2009), we can find positive correlation for three countries with overall HIV prevalence between 4.6%-6.2% whereas for two countries with the overall HIV prevalence below 2%, there is no significance relationship found).

Besides studying the correlation between the HIV status and the education level, we also want to analyse the impact of the education on the knowledge of the prevention and on the actual preventive behaviour. These variables tend to have positive relationship to the level of education as shown by de Walque (2008) on the data from five Sub-Saharan African countries. In the thesis, we want to see whether these results are consistent among the countries with different overall HIV prevalence or whether e.g. even less educated people in countries with high prevalence of HIV tend to have more knowledge about the prevention as the HIV can have higher priority in education in such countries. Also, we want to see whether the knowledge about the prevention truly implies to at least some extent, the preventive behavior itself and how is the preventive behavior influenced by the development level of the countries (e.g. in poor countries, people can have knowledge about prevention, but do not necessary have access to it).

Hypotheses

Hypothesis #1: Consistency of the strength of the HIV status and the education correlation among similar countries, based on the overall HIV prevalence and/or development level.

Hypothesis #2: Knowledge of the HIV prevention is more common even for less educated people in the countries with larger overall HIV prevalence.

Hypothesis #3: Knowledge of the HIV prevention does not imply preventive behavior itself in poor countries.

Methodology We plan to use the data from the Demographic and Health Survey (DHS) from years 2010-2013. The DHS collects individal-level data from many developing countries. It provides the data about socioeconomic factors of the respondents such as marital status, ethnicity, level of education, level of poverty, sexual

behaviour, etc. Also, for some countries, it provides the data about HIV status of the respondents.

Our goal is to estimate the correlations with newer and larger datasets than in the previous works that were limited by the availability of the data. The new datasets are available for more countries. We can find countries that differ significantly in the GDP (USD300 - USD14500 per capita) or in the overall HIV prevalence (3% - 23.3% of total population). Based on these differences, we want to investigate whether the strenght of the correlation varies among the states with different overall HIV prevalence or with different level of GDP.

We plan to choose 15 Sub-Saharan African countries that have overall HIV prevalence above 3%. To conduct the analysis, we will use the bivariate and multivariate logistic regressions both separately on the countries and on the pooled data controlling for age, residence, marital status, etc. To distinguish the differences among the different countries and regions, we will use multilevel model. Also, we will run various regressions for variables concerning the attitude to and knowledge of the HIV with the level of education as the dependent variable. We will also mention possible problems of estimating the HIV and education correlations based on Beegle and de Walque (2009).

Expected Contribution Using the newer and larger datasets, we can extend the current literature on the HIV and education correlates as past works were mainly done with limited number of the data. With larger amount of data available, we can explore whether the results are consistent among the countries or whether the factors influencing the HIV status are individual for different countries. Based on the results, we believe it will be easier to allocate groups of people that are in greater risk of HIV infection and certain policies (financial aid, prevention programmes) can be addressed directly to these groups of people.

Outline

1. Literature Review: We will briefly summarize the current literature on the topic.
2. Data Describtion: We will describe the datasets used in our analysis.
3. Methodology: We will explain econometric models that will be used for the analysis. Also, we will comment on possible problems of the estimation.
4. Results: We will discusse the results.
5. Conclusion: We will summarize the results and their implications.

Core bibliography

BEEGLE, K. & D. DE WALQUE (2009): "Demographic and Socioeconomic Patterns of HIV/AIDS Prevalence in Africa." World Bank Policy Research Working Paper No. 5076

DE WALQUE, D. (2009): "Does Education Affect HIV Status? Evidence from five African Countries." The World Bank Economic Review, 23(2): pp. 209-233

FORSTON, J. G. (2008): "The Gradient in Sub-Saharan Africa: Socioeconomic Status and HIV." Demography, 45(2): pp. 303-322

GLYNN, J.R., M. CARAËL, A. BUVÉ, S. ANAGONOU, L. ZEKENG, M. KAHINDO & R. MUSONDA (2004): "Does Increased General Schooling Protect Against HIV Infection? A Study in Four African Cities." Tropical Medicine and International Health 9: pp. 4-14

GREGSON, S., H. WADDELL & S. CANDIWANA (2001): "School Education and HIV Control in Sub-Saharan Africa: From Discord to Harmony?" Journal of International Development 13: pp. 467-485

MWAKOLABO, A. (2007): "Implications of HIV/AIDS for Rural Livelihoods in Tanzania: The Example of Rungwe District." African Studies Review 50(3): pp. 51-73

Author

Supervisor

Chapter 1

Introduction

The HIV/AIDS (human immunodeficiency virus/acquired immune deficiency syndrome) infection is a large worldwide threat for populations of many countries. The most severely hit region is with no doubt Sub-Saharan Africa as many countries from that region are amongst the countries with the highest HIV prevalence in the world. Moreover, the mortality connected to the HIV is very high in Sub-Saharan Africa due to the lack of the treatment possibilities that in turn has negative impacts on the whole economies of the countries and on the families of the HIV positive individuals.

Even though the medical progress in the HIV treatment has experienced a significant improvements over the past years, the treatment is not available for everyone in need (especially in poor regions such as Sub-Saharan Africa). Thus, it is important to fight the HIV infection also in different ways than medically. To do so, we need to understand the correlation between the HIV status and socio-economic factors of the individuals. Then we can target the help and prevention programmes more efficiently to specific groups of people that are at the highest risk of the HIV. This can help to prevent the further spread of the HIV or help the families with HIV positive members to keep their living standards. The analyses concerning this topic were limited in the past due to the data availability. With current possibilities of cheap and quick HIV testing we can find large datasets from many developing countries.

In Sub-Saharan Africa, the education is still unreachable for large share of population. Looking at most of the health threats, the education is often regarded as a kind of a social vaccine when people with more education are more likely to avoid the threat of the infections or diseases. Nevertheless, the research dealing with the HIV status and education delivers rather mixed results. Thus,

in the thesis we take an advantage of a new data collection about the HIV status and other socio-economic indicators from 21 Sub-Saharan African countries and perform an econometric analysis dealing with the relationship between education and factors that are connected to the HIV infection - HIV status, knowledge about the HIV and sexual behaviour. Moreover, we will analyse whether the relationships are the same for the countries with different HIV prevalence and with different level of wealth.

The thesis is organized followingly: Chapter 2 presents the general information about the HIV epidemic and its impact on the economics of both countries and individuals. Chapter 3 contains literature review. The hypotheses are described in Chapter 4. Data description with an overview of the variables used in the analysis is presented in Chapter 5. Chapter 6 contains a brief methodology overview together with the discussion about the limitations of the analysis. Chapter 7 then provides the empirical results. Chapter 8 presents several robustness checks. The results are discussed in Chapter 9. The Conclusion then summarizes the results.

Chapter 2

HIV/AIDS epidemics

Even though the HIV/AIDS epidemic is a well-known worldwide problem, we believe that to better understand the purpose of the analysis presented in the thesis, it is important to introduce the disease from the broader perspective. Thus, in this chapter, we summarize the current state of the epidemic, briefly address the medical aspects of the disease and finally we present various economic impacts that the HIV/AIDS epidemic has on both nations and individuals.

2.1 Current state of HIV/AIDS epidemics

The HIV/AIDS disease is worldwide spread. Nevertheless, some parts of the world such as middle and low income countries, are hit more severely than others. The region with the highest HIV/AIDS prevalence is, as clearly depicted in the Figure 2.1, Sub-Saharan Africa which is also the subject of the thesis.

The HIV was firstly discovered in the USA in 1981. It is estimated that the HIV/AIDS has caused death to more than 34 million people so far. According to UNAIDS (2016a) [1], there were approximately 36.7 million people worldwide infected by HIV in 2015. Even though the progress of the disease has shown decreased momentum in the recent years, UNAIDS estimates that the number of newly infected people in 2015 accounted for 2.1 million people worldwide. Despite the strong global campaign targeted against the HIV infection, there

[1] UNAIDS (Joint United Nations Programme on HIV/AIDS) is a leading organization in the fight against the HIV/AIDS. The UNAIDS work is focused on preventing the spread of HIV, providing treatment for HIV positive individuals or providing information about the HIV/AIDS with an ultimate goal to end the world HIV epidemic.

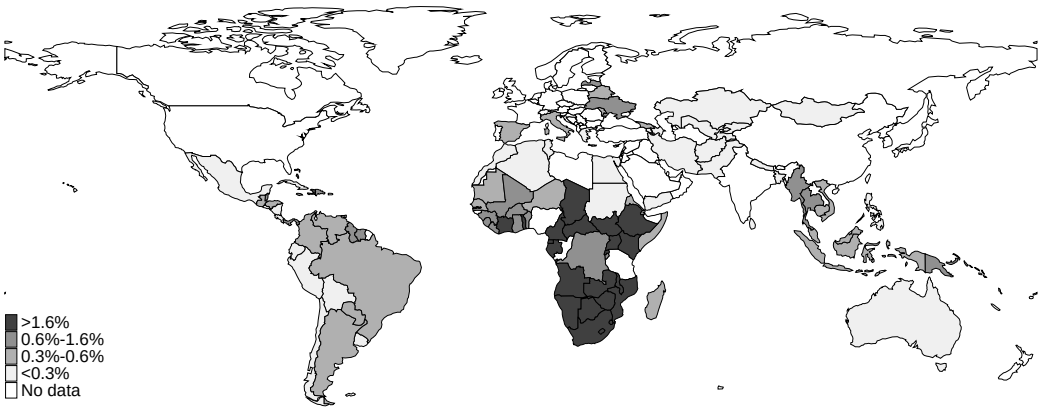

Figure 2.1: World HIV prevalence

is still around 50% of HIV positive individuals who do not know their status and nearly 60% do not have access to the HIV treatment.

The extremely unfavourable situation of the epidemic in the Sub-Saharan Africa can be illustrated by the following numbers. The region accounted for more than 25 million HIV positive people in 2015. The region also accounted for more than 65% of newly infected people worldwide in the same year. Moreover, more than 72% (810 thousand people) of deaths worldwide connected with the HIV happened in the Sub-Saharan region.

2.2 Medical view and treatment possibilities

The HIV/AIDS is a summarizing name for the condition that is caused by the human immunodeficiency virus. The virus is transmitted by unprotected sexual activities, blood transfusion, contaminated needles, and from mother to children during pregnancy, birth or breastfeeding.

Newly infected individual may not experience any symptoms and can live for a relatively long period without knowing they are HIV positive. More often, there is an infection connected with the symptoms such as fever, nausea, rash or headache. These symptoms however may be misdiagnosed and the HIV may not be detected. After a certain time period that varies among individuals (usually 3 to 20 years) of relative calm, the HIV influence the immune system of the infected individual to larger and larger extent increasing risk of tumours or infections such as tuberculosis. This state is referred to as AIDS. As the body of the infected individual is weaker and more prone to the infections, the AIDS stadium is followed by death in a matter of few years.

As the HIV is a relatively new disease, there is currently no medical treat-

ment that could completely cure HIV or AIDS. Nevertheless, the antiretroviral treatment (ART) is an efficient way to slow the progress of the HIV/AIDS, sometimes even to the extent of achieving normal life length. In the recent years, the cost of ART has dropped significantly. According to UNAIDS (2012), the cost of an HIV treatment for a year for one individual was around USD 10,000 in 2000 and dropped to only USD 100 in 2012.

2.3 Look into future

With rising availability of the ART and rising awareness about the disease, the epidemic is expected to further decline in the future. The UNAIDS target 90-90-90 (UNAIDS 2016b) plans that 90% of HIV positive people will know their status, will have access to the ART and will have suppressed the viral loads (i.e. the amount of virus in their body will drop below life-threatening level) by 2020. Following the plan, it is believed that the world epidemic will end in 2030. The plan is relatively ambitious since it is estimated that nowadays, 50% of HIV positive individuals do not even know their status.

2.4 Social and economical impacts of HIV/AIDS epidemics

The HIV/AIDS infection has a significant negative impact on the economy of countries through decreasing available human capital. Greener (2002) argues that due to the higher mortality connected with the HIV/AIDS infection, the countries with high HIV prevalence suffer from low workforce growth, changing age structure of workforce towards young inexperienced workers, loss of skilled workers and decreasing taxable population. Moreover, as the HIV/AIDS is connected with an overall higher chance of being infected by other infections such as tuberculosis, the productivity is also decreasing due to the often sick leaves. Altogether, the HIV/AIDS infection leads to the increasing expenditures on:

- Health care - costs paid by both governments and individuals for the HIV treatment and other medical costs.
- Training - costs paid by governments and companies to train new workforce to replace sick or dead workers.

- Sick pay - costs paid by governments and companies for temporary sick workers.

The increased expenditures in turn reduce savings of households and reduce both public and private investment. On the government level, the expenditures for HIV treatment put a pressure onto the state budget and contribute to the deficit spending.

Altogether, the above-mentioned consequences of the HIV slow down overall economy performance. Many studies have been developed to assess the impact of the HIV infection on country gross domestic product (GDP). Generally, the studies find negative relationship between the HIV prevalence and the GDP growth, although not always the relationship is significant. For example, Dixon *et al.* (2001) estimate on the sample of 41 African countries for period between 1960 and 1998 that growth rates tend to be 2-4% a year smaller due to the HIV infection than they would be without it. Kirigia *et al.* (2001) studied impact of HIV on GDP of 45 African countries and they found insignificant but negative relationship between GDP and HIV mortality. Igwike & Hussain (2012) studied the relationship between the HIV prevalence rates and various macroeconomic variables such as GDP, savings, consumption, labour supply and investment with the conclusion of negative and significant relation.

Various studies also examine the micro-economic impacts of the HIV infection. As shown by Mwakalobo (2007) on a case study from Tanzania, being HIV positive significantly increases the risk of the household to fall below the poverty line due to the medical expenses, loss of the job, funeral expenses and decreasing household savings. Béchu (1998) shows on the data from Cote d'Ivore that households that include an HIV positive person spend twice more on the medical expenses compared to other households. Moreover, almost 80% of all household medical expenditures were used for the HIV positive person rather than for other ill household members. Bachmann & Booysen (2003) studied 404 households in South Africa and based on their econometric analysis, they found statistically significant differences between households with and without HIV person. Households with the HIV positive member(s) were on average poorer than other households. Furthermore, they found that income of the households fell more dramatically in the households with HIV positive members during the 6 months observation period.

As stated by Mahajan *et al.* (2008), the fear of stigma of being HIV positive is making the fight against the HIV infection even more complicated. As people

often do not have proper information about the HIV/AIDS, they fear the HIV positive individuals. In turn, the HIV positive individuals try to hide their status, do not visit medical treatment or even do not go for testing. These conditions contribute for the further spread of the HIV/AIDS disease.

This chapter briefly illustrated that the HIV/AIDS infection is a factor that prevents many countries from better development and brings whole households into the poverty. Especially in the case of poor Sub-Saharan Africa, fighting the HIV/AIDS can improve the unfavourable macroeconomic situation in the region. As the ART is still not available for all people, it is important to fight the infection also by prevention and better targeted educational programs. For this, it is important to understand which groups of people are at the higher risk to be able to deliver the assistance more efficiently.

Chapter 3

Literature review

In this chapter we briefly present and summarize recent studies connected with the topic of the thesis. Special attention is given to those studies that use the data from the Demographic and Health Survey (DHS) which is used also for the analysis presented further in the thesis. Nevertheless, we also comment on other studies we find relevant for the topic.

3.1 Education and HIV status

As Africa is hit by the HIV/AIDS to large extent, most studies about the HIV/AIDS and socio-economical correlates are done on the datasets from African countries. Nevertheless, most of the studies lack quality data. Usually, the datasets come only from small subsample of population (e.g. data only about pregnant women - Fylkesnes *et al.* (2001) or data about a risky social group such as sexual workers - Nagot *et al.* (2002)) and come from certain region in the country or even from one city. The results are therefore difficult to generalize. However, with the new data about HIV/AIDS status from the DHS, there are new possibilities to further explore the correlations.

The HIV/AIDS status and education relationship is rather unclear based on the current literature. The findings about this correlation are mixed and very sensitive to estimation techniques used. For most health problems the education and wealth behave as a social vaccine, meaning that higher education and wealth relates to better overall health of individuals. Nevertheless, the HIV/AIDS seems to behave differently when many papers found positive and significant relationship. For example, according to Fox (2010) who summarizes findings from various papers that analyse the wealth of both individuals and

countries with respect to HIV/AIDS, the evidence suggest that more wealthy individuals are more likely to be HIV positive.

De Walque (2009) used in his analysis DHS data from 5 countries (Burkina Faso, Cameroon, Ghana, Kenya and Tanzania) from years 2003 and 2004. This is one of the first analysis using large DHS dataset. De Walque pooled the data and estimated the regression with HIV status as a dependent variable. He controlled for various socio-economic factors such as education, wealth, religion, marital status or age. The relationship between education and HIV status appeared insignificant in his pooled regression. Nevertheless, with restricted sample to only individuals from the urban area, he found negative gradient between education and HIV status.

Another study using the same dataset as De Walque was done by Fortson (2008). Surprisingly, the results are not consistent with those suggested by De Walque. Fortson finds positive gradient between education and HIV status meaning that more years of education relate to higher chance of being HIV positive. Fortson uses different specification of the estimation when she expects the non-linear effect of education. Thus, she includes education squared into the regression. Moreover, she studies also relationship between wealth and HIV status with a conclusion that there is no significant relation between those two variables.

Smith *et al.* (2012) suggest that contrary to majority of previous studies that claimed that education increases a chance of being HIV positive, there has been gradual shift towards education as a social vaccine in case of HIV infection.

The results appear to be highly heterogeneous across countries or even cities as for example Corno & De Walque (2007) when analysing DHS dataset from Lesotho from 2004, find that there is an overall negative relationship between level of education and HIV status. Glynn *et al.* (2004) studied data from 4 cities from Kenya, Zambia, Cameroon and Benin. The results were heterogeneous as for cities from Kenya and Zambia, they found no significant relation between HIV status and education. On the other hand, in a city from Cameroon, the more educated women appeared to have significantly lower risk of being HIV positive than those with less education. The same results appeared for men from a city in Benin.

A study conducted by Fylkesnes *et al.* (2001) in Zambia found declining trend in the HIV prevalence connected with the increasing level of education among Zambian antenatal women since 1990s. Also, authors claim that low

education is associated with stable or rising HIV prevalence among the observed group of women (women 15-29 years old). Moreover, the effect of urban setting appears to be a positive factor connected with HIV prevalence when the population living in urban areas tend to use condom more often compared to the rural population.

The effect of education on the HIV status appears to vary also with the type of place individuals live in. Smith *et al.* (1999) studied effect of education on HIV prevalence in Rakai district in Uganda. The conclusion of their study is that the higher education is significantly associated with the higher risk of being HIV positive in the rural areas, whereas in the main road trading centres, the association appeared insignificant.

One possibility why the results are often contradicting each other and inconclusive can be truly large heterogeneity in the socio-economical patterns of the disease. Also the correlation could be changing over time and thus delivering contradictive results. The studies concerning these correlations bear also several limitations which will be discussed later in the thesis.

Nevertheless, due to the small datasets often used for the analyses, we believe that estimating the relationship using larger and newer datasets could bring up new and additional results. These can be then used to better address the problem and target the help more efficiently.

3.2 Education and HIV knowledge

The knowledge about prevention and transmission of the HIV/AIDS is a crucial factor that can influence the momentum of the disease. Nevertheless, it seems that the knowledge about the basic facts regarding the HIV/AIDS is still poor, especially in developing countries.

For illustration, Nasir *et al.* (2008) examined the HIV/AIDS knowledge among students of dentistry in Sudan. As dentistry is closely related field to medicine, we would expect the students to have good knowledge about HIV transmission. Nevertheless, only 49% of students involved in the study had correct knowledge about transmission through contamination.

Another study carried out by Negin *et al.* (2012) focused on the older part of the African population. As this group of people is still sexually active and represents an important part in the structure of the society and families, it is also important that they would possess knowledge how to prevent themselves from the HIV. Based on the survey taken in nine African sites, the authors

found that people older than 50 years have generally lower knowledge about HIV than people aged 24-50 years.

Education is a natural candidate for better knowledge about the HIV/AIDS. For example, Fako *et al.* (2010) conducted research in Botswana regarding the predictors of knowledge about HIV/AIDS among young people. They used the data from 1,294 students with finding that education in terms of quality (students from better schools attained higher knowledge) and level (university students had better knowledge than high school students) is an important predictor of the knowledge about HIV/AIDS.

That education plays an important role for the HIV knowledge is shown also by Mwamwenda & Kariuki (2014) who compares HIV knowledge regarding transmission of the disease, among secondary school and university students in Nairobi, Kenya. Even though both groups attained relatively good knowledge of HIV, the knowledge was significantly higher for university students.

Moreover, De Walque (2009) states that education is "one of the most consistent predictors of knowledge" when he uses a proxy for knowledge about HIV whether healthy looking person can be HIV positive using DHS data. Also Agüero & Bharadwaj (2014) show on the data from Zimbabwe that the education is a strong predictor of the comprehensive HIV knowledge and that education decreases the chances of having a general misconception about the HIV.

From the previous studies it appears that education is a factor that positively influences the knowledge about the HIV. In the thesis we want to further examine the effect using larger country-representative dataset and examine the consistency of the results among different countries.

3.3 Education and sexual behaviour

As the majority of HIV transmission is through sexual intercourse, the sexual behaviour is one of the most important determinants of the future of the epidemic. The risky sexual behaviour such as not using condom or having many sexual partners contributes with no doubt greatly to the further spread of the HIV.

Hargreaves *et al.* (2008) conducted research in South Africa among young people of age from 14 to 25 years. They found that the school attendance has a positive effect on safe sexual behaviour measured by condom use, number of unprotected intercourses in the past or by number of sexual partners.

Baker *et al.* (2011) found robust and positive relationship between education and condom use using a dataset from 9 Sub-Saharan countries. Moreover, the positive effect of education on the use of condom in non-spousal partnerships is presented also in the analysis by Lagarde *et al.* (2001) who used data from four cities from Cameroon, Benin, Zambia and Kenya with the sample size of 4,624 non-spousal partnerships.

On the other hand, De Walque (2009) explores the relationship between sexual behaviour and education in Sub-Saharan Africa on DHS data with rather mixed results. He found a positive relation between education and use of condom. However, at the same time, the number of sexual partners increased with the increasing years of schooling.

Chapter 4

Hypotheses

As the results from previous studies about the education and HIV/AIDS status are mostly mixed and inconclusive, we want to use the larger dataset to explore the topic in more detail. In the thesis, we use data from 21 Sub-Saharan African countries to examine the relationship between education and factors connected with the HIV/AIDS infection, i.e. HIV status, HIV knowledge and sexual behaviour. Firstly, we assume that the education and HIV status would be negatively correlated, i.e. that education would have the effect of a social vaccine. Regarding HIV knowledge, we expect to find the positive effect of education. Lastly, we assume that the effect of education on sexual behaviour would have a preventive effect, i.e. that more educated people would understand the need of protection regarding sex life and would act accordingly.

Further, our hypothesis is that similar countries in the terms of wealth and overall HIV prevalence would achieve similar results in terms of correlation of education and the HIV prevalence, i.e. that the magnitude of the correlation would be consistent across the countries. Therefore, we divide the countries based on their wealth measured by Gross National Income (GNI) per capita and by overall HIV prevalence.

Lastly, we want to test whether the effect of education on the HIV knowledge differs with different overall HIV prevalence. The effect of education could be either stronger or weaker based on the following rationale: either the higher HIV prevalence countries consider the HIV threat as more important and would therefore provide more information about the HIV in schools which would increase the effect of education on the HIV knowledge; or the higher HIV prevalence countries could provide the information generally to public via other channels than school which would in turn diminish the effect of education.

The hypotheses are summarized as following:

Hypothesis #1: The education and HIV status is negatively correlated.

Hypothesis #2: The education has a positive effect on knowledge about the HIV/AIDS.

Hypothesis #3: The education has a preventive effect regarding sexual behaviour.

Hypothesis #4: The correlation between education and HIV status of an individual is consistent regardless the country wealth and/or country overall HIV prevalence.

Hypothesis #5: The effect of education on knowledge about HIV is stronger/weaker for countries with higher HIV prevalence.

Chapter 5

Data description

In the first part of this chapter we present a general overview of the dataset we use later for the analysis. In the second part we provide the description of the dependent and independent variables as well as data summary statistics.

5.1 Demographic and Health Survey

The main goal of the thesis is to estimate the relationship between the level of education and factors connected with the HIV infection such as HIV status, HIV knowledge and sexual behaviour. For this purpose we use a large dataset from 21 countries containing several socio-economic variables as well as indicators about sexual behaviour, HIV knowledge and HIV status of the individuals. We believe that this analysis would further contribute to exploring the topic as to our best knowledge, the earlier work on this topic used smaller datasets (e.g. maximum of 6 countries for exploring the correlations between education and HIV status). Furthermore, the majority of earlier research is based on the datasets from mid-2000 whereas we take advantage of the new DHS data collection from years 2008-2014.

DHS began collecting the data from the developing countries in 1984. In the thesis we work with the fifth and sixth data collection from the Measure DHS Phase III and IV from period 2008-2014. DHS datasets contain many socio-economic factors such as marital status, religion, education, proxy for wealth, ethnicity, place of residence, literacy and several other variables describing the sexual behaviour of the individuals. The information about the HIV/AIDS status began to be collected in 2001 and nowadays, it is available for more than 25 countries from the Sub-Saharan Africa. Usually, the DHS dataset

from one country contains information about more than 10,000 individuals. The HIV/AIDS status is not however tested for each DHS participant. Usually, approximately half of the respondents from DHS are tested for the HIV status.

DHS is organized in such manner that the respondents should be reflecting the country population patterns; i.e. the respondents are not only from one region or city, but the data collection is carried out throughout the whole country. Also the DHS datasets provide weights that should be used for the analysis to guarantee that the sample truly reflects the country population. All surveys are unified, meaning that we can easily pool all the data and perform the analysis for the large part of the Sub-Saharan Africa.

The sample covers men and women of age from 15 to 59 years. The survey for each country is divided to households, men and women datasets. Additional dataset about HIV status contains only one variable stating whether individual is HIV positive or negative. It can be linked to the respective women or men questionnaire using the individual unique identification numbers. The basic socio-economic datasets and datasets containing the HIV status are free of charge for researchers. Nevertheless, the precise purpose and outline of the research has to be specified and consequently approved by the DHS organization. The data are strictly confidential and cannot be distributed to the third parties.

For the analysis, we decided to use the datasets from 21 Sub-Saharan countries. We did not use additional three countries with available datasets (Tanzania, Uganda, Mozambique) as they unfortunately come from a different type of survey and have slightly different structure compared to the most of the DHS datasets. Neither we used the data from Niger as the dataset did not contain all information we needed for the analysis. Nevertheless, we believe that 21 countries can be considered as large enough sample to produce meaningful results.

The Table 5.1 shows countries we use for the analysis with the year in which the data collection took place in the parenthesis. Also the Table 5.1 provides the number of female and male respondents and the number of HIV positive individuals in the datasets. For each dataset, there is a larger number of female respondents. The total number of respondents for each country varies from 6,906 from Kenya to 29,007 from Zambia. From the HIV prevalence in the dataset, we can observe a heterogeneity of HIV prevalence in the Sub-Saharan countries. The lowest number of HIV positive people are in the datasets from Mali, Senegal or Burkina Faso. On the other hand, the highest number of

Table 5.1: Dataset description

Country	Females	HIV positive females	%	Males	HIV positive males	%
Burkina Faso (2010)	8,350	100	1.20%	7,039	60	0.85%
Burundi (2010)	4,510	108	2.39%	4,078	56	1.37%
Cameroon (2011)	7,254	434	5.98%	6,948	215	3.09%
Congo (2013-14)	9,316	133	1.43%	8,322	44	0.53%
Cote d'Ivore (2011-12)	4,656	209	4.49%	4,352	127	2.92%
Ethiopia (2011)	15,517	358	2.31%	13,015	182	1.40%
Gabon (2012)	5,490	315	5.74%	5,502	168	3.05%
Gambia (2013)	4,487	93	2.07%	3,284	43	1.31%
Ghana (2014)	4,687	119	2.54%	4,161	45	1.30%
Guinea (2012)	4,692	108	2.30%	3,688	56	1.52%
Kenya (2008-09)	3,811	318	8.34%	3,095	154	4.98%
Lesotho (2009)	3,894	997	25.60%	3,075	543	17.66%
Liberia (2013)	4,377	74	1.69%	3,805	45	1.18%
Malawi (2010)	7,398	890	12.03%	6,512	530	8.14%
Mali (2012-13)	5,110	66	1.29%	3,751	31	0.83%
Namibia (2013)	4,984	814	16.33%	3,874	419	10.82%
Rwanda (2010)	6,952	266	3.83%	6,296	154	2.45%
Senegal (2010-11)	5,590	44	0.79%	4,327	26	0.60%
Sierra Leone (2013)	7,865	141	1.79%	6,735	81	1.20%
Togo (2013-14)	4,807	127	2.64%	4,365	72	1.65%
Zambia (2013-14)	15,433	2,328	15,08%	13,574	1,573	11.59%
Zimbabwe (2010-11)	7,852	1463	18.63%	6,045	811	13.42%

HIV positive people is from countries such as Namibia, Zambia, Lesotho or Zimbabwe. Also, the HIV prevalence appears to be higher for women for all 21 countries. AVERT (2016) [1] summarizes several reasons why women are generally at higher risk of HIV/AIDS infection in the Sub-Saharan Africa:

- Gender inequality - women often live in the abusive conditions when the protection against HIV/AIDS is limited or not possible.
- Poor access to the healthcare - in some developing countries, women do not have the same access to the healthcare as men.
- Poor access to the education - again, in some developing countries, women do not have the same possibilities for education as men and cannot obtain information about the HIV/AIDS.
- Legal restriction - there are laws in some countries that prevent women from getting a proper access to the prevention or access to the ART.

[1] AVERT is an organization providing information and advice on HIV/AIDS

5.2 Dependent variables

In this section we briefly describe the dependent variables we use in the regressions, i.e. HIV status, HIV knowledge and sexual behaviour variables (number of sexual partners and use of condom during the last sexual intercourse).

5.2.1 HIV status

The HIV status variable is a binary one equal to one if the respondent is HIV positive and zero otherwise.

The DHS survey collects the HIV/AIDS status data based on a blood drop sample from the individuals' fingers. The blood samples are analysed for the HIV virus and if the HIV/AIDS positive status occurs, the sample is tested again together with 5-10% of negative results to decrease the probability of a test failure. The HIV/AIDS testing is simple and can include a large number of participants as opposed to many previous data collections that usually picked only certain group of individuals for testing such as pregnant women or high risk population (e.g. drug addicts). The testing is completely anonymous providing relatively high response rate that can be found for each respective country in the Appendix in the Table A.1. The overall response rate is higher than 80% for both men and women in all countries, although it tends to be higher for less educated people. More about response rates and their possible impacts on the analysis is discussed later in the thesis.

5.2.2 Knowledge about HIV

The DHS datasets offer relatively wide variety of variables describing knowledge about the HIV infection where people are asked yes/no questions concerning HIV transmission and general misconception connected with HIV. For the purpose of the analysis, we construct variable describing comprehensive knowledge about the HIV as suggested by DHS. The variable is constructed from 5 categorical variables:

- Condom use - equals to 1 if a respondent thinks that the use of condom reduces risk of the HIV transmission.
- Stable sexual partner - equals to 1 if a respondent thinks that having one stable sexual partner reduces risk of the HIV transmission.

- Healthy looking person - equals to 1 if a respondent thinks that healthy looking person can be HIV positive.
- Mosquito bite - equals to 1 if a respondent thinks that the HIV infection can be transmitted by a mosquito bite.
- Sharing food - equals to 1 if a respondent thinks that the HIV infection can be transmitted by sharing food with an HIV positive individual.

In some questionnaires, the questions about general misconception (i.e. mosquito bite and sharing food) are replaced with similar question *HIV infection can be transmitted by supernatural events* (i.e. by witchcraft).

Combining all these variables mentioned above, we can obtain variable *comprehensive knowledge about HIV* that is equal to one if respondent answered 'yes' to questions regarding condom use, one sexual partner and healthy looking person and 'no' to questions regarding general misconceptions.

5.2.3 Sexual behaviour

To assess the effect of education on the sexual behaviour, we use two dependent variables describing to some extent the sexual behaviour - number of sexual partners and use of condom during the last sexual intercourse. Even though these two variables do not precisely describe whether the individuals sexual behaviour is dangerous or not (e.g. the use of condom is not necessary in some situation when the partners are trying to conceive a baby), they can at least serve as a proxy for it.

The variable use of condom during the last sexual intercourse is a binary one equal to 1 if respondent used condom during the last sexual intercourse during the past 12 months and 0 otherwise. The variable number of sexual partners can equal from 1 to 99, when value 99 indicates 99 lifetime sexual partners and more.

5.3 Independent variables

To control for as many factors as possible that could influence or be correlated with the above mentioned dependent variables, we take an advantage of the fact that DHS questionnaires provide many socio-economic variables about the respondents. For the analysis we use such variables we believe are relevant

for the problem regarding the HIV/AIDS. The complete description of the variables used follows:

- Years of education - represents a total number of completed years of education.
- Age - represents respondents age. It enters the regression in a form of dummies for 5 age categories (15-25 years, 26-35 years, 36-45 years, 46-55 years and more than 55 years).
- Women - equals to 1 if the respondent is a woman.
- Religion - represents respondents' religion. It enters the regression in the form of four dummy variables for no religion, Islamic religion, Christianity and other religion that accounts for animist/traditional religion.
- Marital status - represents current marital status of the respondent. It enters the regression in the form of dummy variables for never in union/marriage, currently in union/marriage and omitted one, formerly in union/never married.
- Wealth - proxy for wealth consisting of owning certain assets (more information below).
- Urban - equals to 1 if the respondent lives in the urban are (i.e. city).
- Region - represents region the respondent is from. It enters the regression in the form of dummy variables for each region.
- Last sexual partner - indicates the relationship the respondent had with the last sexual partner. It enters the regression in the form of dummy variable for indicating stable partner, random partner, sexual worker and other.

Generally it is problematic to measure wealth of the individuals. DHS datasets provide wealth index that divides the individuals into 5 groups according to their wealth. It is, however, difficult to interpret this index as there is no exact interpretation of e.g. difference between the first and the second wealth group. Therefore, we constructed proxy for wealth in a similar way as Case *et al.* (2004), who also analysed DHS datasets.

The proxy is constructed from the information about the presence of eight assets in the respondents' households, namely: flushing toilet, electricity, television, car, bicycle, motorcycle, refrigerator and radio. The new proxy for wealth then represents a fraction of how many assets respondent owns, i.e. if respondent owns only a car and a bicycle, the proxy for wealth has a value of $\frac{2}{8}$. Even though this measure is far from being objective and accurate wealth measure, it is easily interpretable. The similar approach to the wealth measure was adopted also by Fortson (2008) when using the DHS data.

The dependent variables could be also influenced by the ethnicity when certain group of people could be somehow disadvantaged in one or more areas connected with the HIV infection. Unfortunately, in our analysis, we are unable to control for the ethnicity effect as the DHS dataset provides data about ethnicity only for 13 countries out of 21. Nevertheless, we believe that having the data, controlling for ethnicity could improve the accuracy of the results.

5.4 Subsamples

To test the hypotheses dealing with the different effect of education in rich/poor countries and high/low HIV prevalence countries, we need to divide the sample into two subsamples. This division will be always arbitrary and the threshold we use will certainly to some extent influence the results. To divide rich and poor countries, we use the threshold used by World Bank based on the Gross National Income (GNI) per Capita using Atlas method. The current threshold is USD 1,025 (World Bank 2017). This threshold serves as a division between low and middle income countries and is used by the World Bank e.g. for lending eligibility of countries. The Atlas method uses Atlas conversion factor instead of simple exchange rates to reduce the impact of exchange rate fluctuations. This allows for cross-country comparison. The GNI per capita of each country based on World Bank estimates can be found in the Appendix in the Table A.2.

The division of the sample based on the HIV prevalence would be more problematic as there is no officially used threshold between high and low HIV prevalence. In this case we decided to use the threshold based on the median value, i.e. 1.8% HIV prevalence. The HIV prevalence of each country based on UNAIDS estimates can be found in the Appendix in the Table A.2.

The high and low GNI per capita countries and high and low HIV prevalence countries are thus following:

- High GNI countries - Cameroon, Gabon, Ghana, Kenya, Lesotho, Namibia, Zambia
- Low GNI countries - Burkina Faso, Burundi, Congo, Ethiopia, Gambia, Guinea, Liberia, Malawi, Mali, Rwanda, Senegal, Sierra Leone, Togo, Zimbabwe
- High HIV prevalence countries - Cameroon, Gabon, Kenya, Lesotho, Namibia, Zambia, Malawi, Rwanda, Togo, Zimbabwe
- Low HIV prevalence countries - Burkina Faso, Burundi, Congo, Ethiopia, Gambia, Ghana, Guinea, Liberia, Mali, Senegal, Sierra Leone

Figures 5.1 and 5.2 show a map of Africa with the above mentioned countries. The most countries from the dataset are located in the western and southern part of the African continent. It appears that the southern part of the African continent is hit more severely by the HIV infection compared to the western part. Additionally, the high/low GNI and the high/low HIV prevalence subsamples overlap to some extent - the richer countries appear to have higher HIV prevalence.

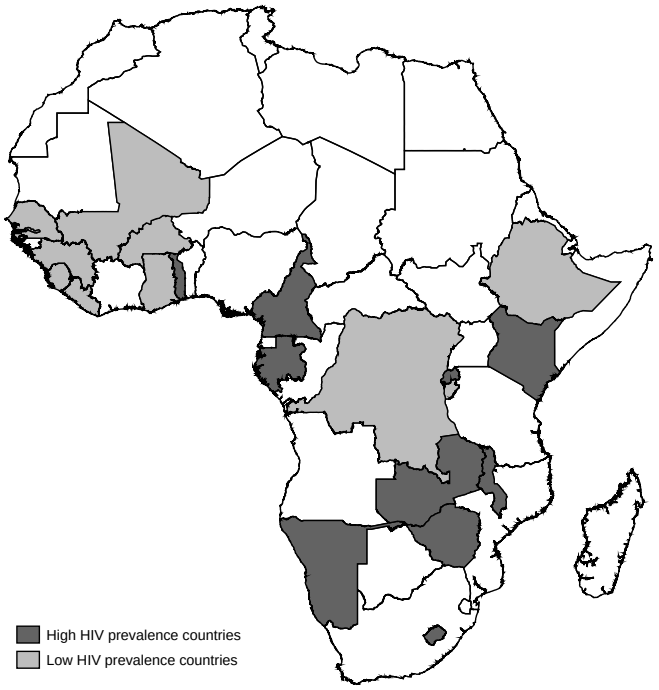

Figure 5.1: Countries by HIV prevalence

Figure 5.2: Countries by GNI per capita

5.5 Data summary statistics

Table 5.2 provides with the data summary statistics such as sample mean, minimum, maximum and total number of observations of the main variables used in the regressions. The summary statistics is divided into two subsamples for women and men to observe the differences between these two groups. In the dataset, there are up to 142,331 observations for women and up to 121,489 observations for men. Only age and urban variables contains all the observations. Other variables have some missing values. DHS explain that there are generally two reasons for missing values in the datasets - either the question was not asked during the interview (i.e. interviewer error) or the respondent refused to answer. Most missing values are (except for variables describing sexual behaviour) for wealth proxy variable when missing values comprise approximately 2.3% for women and 5.7% for men. Nevertheless, we believe that it would not affect the results significantly due to generally large dataset with more than 200,000 observations.

The variables regarding sexual behaviour, i.e. number of sexual partners and use of condom are not available for the whole dataset. The reason is that DHS collects the data about the last sexual intercourse only from those respondents who had sexual intercourse in the past 12 months. Also, the number of

Table 5.2: Data summary statistics

	Number of observations	Sample mean	SD	Minimum	Maximum
1. MEN					
HIV status	121,454	0.044	0.205	0	1
Schooling years	121,419	6.27	4.539	0	24
Age	121,489	30.779	11.929	15	64
Urban	121,489	0.36	0.48	0	1
Wealth	114,233	0.246	0.217	0	1
Number of sexual partners	93,930	6.904	12.765	1	95
Condom use at last intercourse	87,333	0.21	0.407	0	1
Last intercourse with stable partner	87,333	0.936	0.245	0	1
Last intercourse with random partner	87,333	0.032	0.176	0	1
Last intercourse with sex worker	87,333	0.004	0.067	0	1
2. WOMEN					
HIV status	142,307	0.066	0.247	0	1
Schooling years	142,279	4.998	4.408	0	22
Age	142,331	28.492	9.678	15	64
Urban	142,331	0.365	0.481	0	1
Wealth	138,746	0.244	0.22	0	1
Number of sexual partners	118,988	2.27	4.634	1	95
Condom use at last intercourse	101,076	0.111	0.314	0	1
Last intercourse with stable partner	101,076	0.992	0.087	0	1
Last intercourse with random partner	101,076	0.005	0.072	0	1
Last intercourse with sex worker	101,076	0.001	0.024	0	1

sexual partners variable does include only those respondents who had at least one sexual partner.

As stated earlier, there is a relatively large difference in the HIV prevalence between women and men when 6.6% of women and only 4.4% of men are HIV positive in the dataset. The average number of years of schooling is higher for men with 6.3 years on average (compared to women with only 5.0 years). This finding is in line with the overall status of women in Africa when the education is often not available for them for political, social or other reasons. Nevertheless, the standard deviation of schooling years is relatively large. The percentage of people living in the urban area is almost the same (36%) for both genders. The wealth index is also very similar which is something we would expect as we believe that majority of people involved in the survey live in the household containing more members than just one and that the households of respondents would overlap to some extent. The average value of wealth index indicates that on average, respondent owns 2.5 out of 8 assets. The average number of sexual partners is higher for men than for women (almost 7 for men compared to 2.3 for women). This difference could be driven either by higher promiscuity of men or by the fact that men generally tend to overstate their true number of sexual partners. Regarding the last sexual intercourse, only 21% of men and 11% of women used condom. On the other hand, majority of the last sexual intercourse was with a stable partner.

Chapter 6

Methodology

In this chapter we provide a brief overview of the estimation methodology used in the regressions. In the second part of the chapter we comment on possible limitations of the analysis.

6.1 Binary response model

As all dependent variables we use for the analysis, except for number of sexual partners, are binary variables that can attain only values 0 and 1, the binary response model need to be used in the regressions. For better understanding the whole analysis, we briefly address the theory behind the binary response models. The following section is based on Wooldridge (2015).

The general form of the binary response model can be described by the following equation:

$$\text{Prob}(y = 1|\mathbf{x}) = G(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k) = G(\beta_0 + \mathbf{x}\beta), \tag{6.1}$$

where G is a function that can attain values between 0 and 1, i.e. $0 < G(z) < 1$ for all $z \in \mathbb{R}$, $\mathbf{x}$ represents the independent variables, and $\beta_0, ...\beta_k$ are parameters we want to estimate. The assumption about the function G guarantees that the estimated response probabilities would be in the desired interval between 0 and 1.

Several functional forms are proposed for function G by the literature. Nevertheless, the two mostly often used G functions are the logistic function

$$G(z) = \frac{exp(z)}{[exp(z) + 1]} = \Lambda(z), \tag{6.2}$$

and the standard normal cumulative distribution function

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(\nu) d\nu \tag{6.3}$$

where $\phi(z)$ is the standard normal density

$$\phi(z) = \frac{1}{2\pi} exp(-z^2/2) \tag{6.4}$$

The logistic function specification leads to logit model (logistic regression) and the standard normal cumulative distribution function specification leads to probit model (probabilistic regression). Both functions possess similar characteristics:

- Both functions $G(z)$ are increasing.
- The increase is most profound around $z = 0$
- Both functions, as stated earlier, can attain only values between 0 and 1 which can be easily verified by computing limits at $-\infty$ and $+\infty$, i.e. for both G(z), as $z \to -\infty$ then $G(z) \to 0$ and as $z \to +\infty$ then $G(z) \to 1$.

Due to the similarity of the functions, we cannot consider logit model as better compared to probit and vice versa. Usually, the models yield very similar results and the decision about using one or other is often on the researchers experience and intuition.

The logit and probit models are derived from the latent variable model

$$y^* = \beta_0 + \beta \mathbf{x} + e, \tag{6.5}$$

where y^* is a latent variable and the function $1 = [\cdot]$ is an indicator function that attains values 0 or 1. Thus, we can write

$$y = \begin{cases} 1 & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0 \end{cases} \tag{6.6}$$

In the latent variable model, we assume that e and $\mathbf{x}$ are independent and e has a logistic distribution in case of the logit model and a standard normal distribution in case of the probit model and zero mean. We can then derive the response probability of y followingly:

$$\begin{aligned} P(y=1|\mathbf{x}) = P(y^* > 0|\mathbf{x}) =& P[e > -(\beta_0 + \beta\mathbf{x})|\mathbf{x}] \\ &= 1 - G[-(\beta_0 + \beta\mathbf{x})] = G(\beta_0 + \beta\mathbf{x}) \end{aligned} \tag{6.7}$$

As the model is non-linear, the interpretation of the coefficients is not as straightforward as for the ordinary least squares (OLS) estimation. In some areas, logit model is preferred for the analysis as the coefficients can be easily transformed into the odds ratios. The logistic regression (due to the odds ratios) is thus often used e.g. in survey research, in epidemiology, and to express the results of some clinical trials.

Another possibility is to use the marginal effects to interpret logit and probit coefficients. The marginal effect of continuous variables can be expressed as following:

$$\frac{\partial P(y=1|\mathbf{x})}{\partial x_j} = \frac{\partial G}{\partial x_j}(\beta_0 + \beta\mathbf{x})\beta_j = g(\beta_0 + \beta\mathbf{x})\beta_j \tag{6.8}$$

Based on the properties of $G(\cdot)$, the marginal effect would have the same sign as the coefficient from the logit or probit regression. The marginal effect for binary variables can be expressed as following (assuming change of x_1 from 0 to 1):

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + ... + \beta_k x_k) - G(\beta_0 + \beta_2 x_2 + ... + \beta_k x_k) \tag{6.9}$$

Also, this approach can be applied to other discrete variables when we plug in different values into the equations 6.8 and 6.9. As the magnitude of the marginal effects depends on $\mathbf{x}$, we need to plug some values for each x_j to be able to use marginal effects. Usually values such as means or medians are plugged to compute the marginal effects. One of the most common approaches is to plug sample mean to obtain Marginal effect on average (MAE). However, having a lot of binary variables results in somehow strange results as we would plug means of binary variables which does not correspond with any respondent from the sample. Thus, a possible solution is to use Average marginal effect (AME) that averages the individual marginal effects across the sample.

The Maximum Likelihood Estimation (MLE) is used to estimate the logit and probit models. Under the general conditions, the MLE is consistent, asymptotically normal and asymptotically efficient. This allows us to test hypotheses in the same way as for OLS regression.

As the models are estimated by the Maximum Likelihood Estimation, we cannot use regular goodness-of-fit R^2 statistics as for the OLS as the estimates are obtained by the iterative process of maximum likelihood estimation as opposed to minimizing the variance in OLS. Nevertheless, several pseudo R^2 were suggested by the literature that should indicate goodness-of-fit of the model. As Stata (Stata version 13) is used for the analytical part of the thesis, McFadden's R^2 is provided together with the regression outputs. As described by Long & Freese (2006), the McFadden's R^2 is calculated followingly

$$R^2 = 1 - \frac{ln[\hat{L}(M_{full}]}{ln[\hat{L}(M_{intercept}]}, \tag{6.10}$$

where the term in the nominator represents value of the log-likelihood of the full model and the term in the denominator represents value of the log-likelihood of the model with only an intercept. The McFadden's pseudo R^2 can be viewed as a certain indicator of improvement of the full model compared to the model with the intercept only. Similarly to the R^2 at the OLS, the R^2 closer to one indicates better fit of the model.

The DHS datasets contain weights for all observations. The weights should correct for over-sampling of regions with small population and under-sampling of regions with large population. The over-sampling is done to attain representative number of observations even for regions with small population. On the other hand, under-sampling is done in regions with large population to save costs. Also, the weights are designed to correct for different response rates in different regions. Using the weights in the regression analysis should transform the data into the nationally representative sample.

In the regressions, the data are weighted according to the DHS manual (Rutstein & Rojas 2006). In Stata, we use *pweight* command to add weights into the regressions. The *pweight* command treats the weights as probability weights, explained in Stata documentation as "*...weights that denote the inverse of the probability that the observation is included because of the sampling design.*"

Having pooled data from 21 countries means that we are very likely to have the individual observations related within certain groups. The unobservable effects of the respondents could be correlated either at the country or at the regional level, i.e. respondents from the same country (region) would be similar in some unobservable characteristics compared to the respondents from other groups. Thus, we assume a correlation between respondents from certain

cluster while we do not assume any correlation with respondents from other clusters. Not controlling for this effect could have a consequence in underestimated standard errors in the regressions resulting into the incorrect significance of the estimated coefficients. To control for this, we use clustered standard errors in all our regression when we assume clustering at the highest level, i.e. at the country level.

Part of the thesis is an analysis whether the relationship between education and HIV status (HIV knowledge) is consistent (different) based on country wealth and/or different overall HIV prevalence. To test the group differences, we incorporate the interaction term into the regressions. We add a dummy variable indication the group membership and the interaction term(s) of the group membership dummy with the variable which different effect across the subsamples we want to observe, education in our case. The interaction term would thus look followingly:

$$\text{interaction term} = \text{group membership dummy} * \text{education}$$

By adding the group membership dummy into the regression, we allow the intercept to vary across the subsamples. By adding the interaction term, we allow for the different effect of education across the subsamples. The education would then have a different effect across the groups if the interaction term(s) would be statistically significant in the regression.

In the regressions, we use independent variables similar to those used by De Walque (2009) or Fortson (2008) who also used the DHS datasets for their analyses. Before the estimation itself, it is important to note that we do not believe that the independent variables will be sufficient enough to fully explain the dependent variables as we face several constraints in the regressions. As it would seem rational to include sexual behaviour variables, we will not do it in regressions for reasons described later in this chapter. Also, even though the DHS datasets provide exhaustive amount of data, some data that we believe would be helpful to the analysis such as ethnicity are not available.

For the estimation, we use logit specification. Later in the part regarding robustness tests, we also run probit regressions to see whether the different estimation technique would yield different results. As the logit coefficients have not straightforward interpretation except for significance and general direction of the coefficient, all presented results are in a form of Average marginal effects.

As the data comes from different regions, we add dummy variables for each

region in all regressions to control for the fixed effect. We do not report the coefficients at the region dummies as the main focus of this work is somewhere else. We report only the joint significance of the region fixed effect.

6.2 Possible limitations of the analysis

Even with the support of such high quality and large datasets as we have from the DHS survey, the analysis comes with several limitations. We try to address and control for those when possible and at least comment for those that are not possible to control.

Firstly, we are unable to distinguish the causal relationship between the variables of interest, i.e. between HIV status and education level. On one hand, we could claim that the level of education affects the HIV status for example through better knowledge about it or by better overall social status due to the higher education. On the other hand, it is also possible that HIV status would have an impact on one's educational level. It is not uncommon that especially in the Sub-Saharan Africa, being HIV positive comes with a stigma as mentioned earlier in Chapter 2. Therefore, being HIV positive in a young age could have severe impact on one's possibility to achieve higher education. As our data are cross-section, we cannot observe the timing of the variables of interest and thus we cannot establish proper causal relationship. When interpreting the data, we should therefore see the results rather as correlations than as causal relationships.

Secondly, the data may be affected by the survivorship bias. Generally, we can expect the more educated individuals to be wealthier. The wealthier individuals are then more likely to achieve some kind of HIV treatment that can prolong the period between the initial infection and beginning of the AIDS stage. This fact can significantly bias the results as it could strengthen the possible positive relationship between HIV status and education attainment. On the other hand, the HIV status lowers one's wealth, that could then to some extent correct the potential bias (e.g. Shelton *et al.* (2005)).

Based on the two limitations described above, the results from the regression with the HIV status as dependent variable, we would not be able to distinguish the effect the education has on the HIV status but rather we will be able to evaluate which groups of people are now at the highest risk of the HIV. To these groups the aid should be then targeted to avoid further spread of the

infection and to prevent the families with the HIV positive members to e.g. fall into the poverty.

Additionally, there may be a potential bias due to the DHS questionnaires response. People who complete the DHS survey and are then asked to come for the HIV testing can refuse to participate. As long as the non-responsiveness is random and not influenced by the HIV status of the individuals, the analysis would be unaffected. The problem arises when the individuals who know that they are HIV positive refuse to be tested. As shown in the Table A.1 in the Appendix, the HIV testing response rates tend to be lower for the individuals with higher educational attainment in some countries. Nevertheless, the overall response rate is still relatively high (more than 80% for all countries) so that the effect of the potential bias should be minimized by controlling for education and other socio-economic indicators.

A disputable area of the analysis is the decision about inclusion or exclusion the sexual behavior variables into the regression with the HIV status as the dependent variable. The cross-sectional analysis could then suffer from "*reverse causality and endogeneity*" (De Walque 2009). For example, we can expect that individuals who are engaging into more risky sexual behaviours would be more likely to use condoms as they understand the higher risk connected with their behaviour. Nevertheless, majority of previous studies using the DHS data does not control for the sexual variables in the regressions. Thus, in the thesis the sexual behaviour variables are also not included into the regression.

The use of condom during the last sexual intercourse can be an indication whether the individual protects themselves. On the other hand, in marriage, there is sometimes no need to use the condom for many reasons such as trying to conceive a baby or using other means of contraception. To partly control for this, we include in the regression dummies indicating the relationship of the respondent with the last sexual partner.

The number of sexual partners can suffer from the inaccuracy as the reported number can be over- or underestimated. Generally, men are expected to overestimate the number of sexual partners when women tend to underestimate it. Nevertheless, the sexual behaviour of individuals is generally difficult to describe as there will always be a possibility that people did not respond based on the truth and wanted to appear better in front of the person taking the survey.

Chapter 7

Empirical results

In this chapter we present the empirical results of the analysis. The chapter is organized followingly: first section provides results from bivariate and multivariate analysis dealing with the HIV status and education correlation; second section provides results from the analysis dealing with the HIV knowledge and education relationship; third section provides results from the analysis dealing with the sexual behavior and education. For brevity, the results from pooled regressions with the dependent variables HIV status, HIV knowledge and sexual behavior proxies are all presented together in the Table 7.2. Note that all coefficients are in a form of average marginal effects.

7.1 HIV status and education

To evaluate the relation between education and HIV status, we begin with simple two-way frequency statistics. The Table 7.1 presents the number of HIV positive individuals within five groups - no education, 0-5 years of education, 6-9 years of education, 10-14 years of education and more than 15 years - for subsamples of women and men. Note that the division into the presented years groups is arbitrary and does not correspond to the certain level of schooling (e.g. primary, secondary, etc.) as each country has different lengths of the respective school levels.

Large share of the respondents does not have any education - 21% for men and 33% for women. The largest share of the respondents is in the group of 6-9 years of education - education that could to some extent represent primary schooling. Only a very small fraction of respondents has more than 15 years of

Table 7.1: Education and HIV status two-way statistics

	Total number	HIV positive	% HIV positive
1. MEN			
0 years	25,381	573	2.2%
1-5 years	23,059	889	3.7%
6-9 years	38,742	2,189	5.3%
10-14 years	24,447	1,454	5.6%
> 15 years	4,441	209	4.5%
2. WOMEN			
0 years	44,757	1,178	2.6%
1-5 years	27,029	1,729	6.0%
6-9 years	39,247	4,120	9.5%
10-14 years	19,481	2,108	9.8%
> 15 years	2,422	184	7.1%

schooling which would likely represent higher education such as university or college.

The percentage of the HIV positive individuals is larger in all categories for women than for men which is in line with the overall higher HIV prevalence among women. The share of HIV positive individuals is similar for men and women only for people with no education. Then for people with at least some education, the difference between men and women in HIV prevalence is more profound. From simply looking at the two-way analysis, there is an evidence that share of HIV prevalence increases with the increasing education categories up to the group with more than 15 years of schooling. After that level of education (> 15 years), the HIV prevalence is somehow lower. It can indicate some non-liner relationship between the HIV prevalence and education. To allow for this effect in the regression, we add education squared as one of the independent variables.

The Figures 7.1 and 7.2 present the relations of the HIV prevalence and the education level for the four subsamples - rich and poor countries; high and low HIV prevalence countries - we will later use in the regressions. The pattern is somehow similar for the subsamples based on the GNI per capita. On the other hand, looking at the subsamples from high and low HIV prevalence countries, there seems to be differences as for low HIV prevalence countries, there is not much variation in the HIV prevalence with changing years of schooling. Whether these differences are statistically significant will be evaluated further in this chapter.

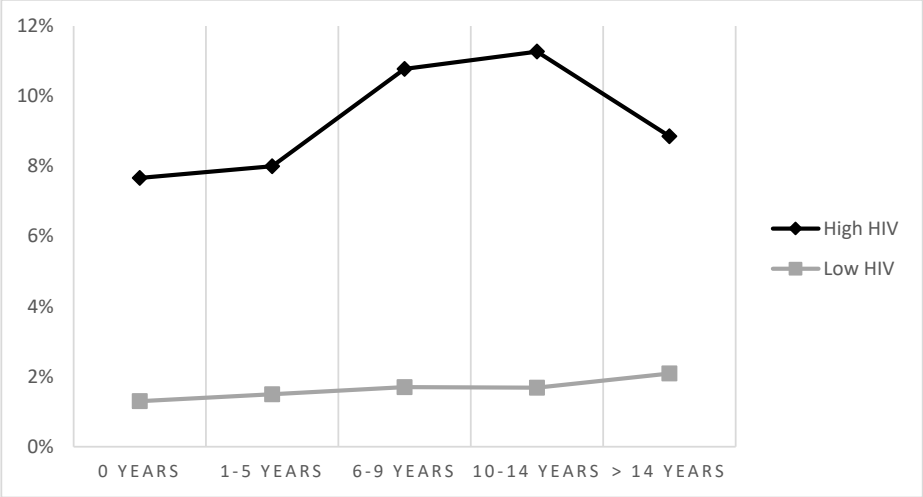

Figure 7.1: HIV prevalence and education - HIV subsamples

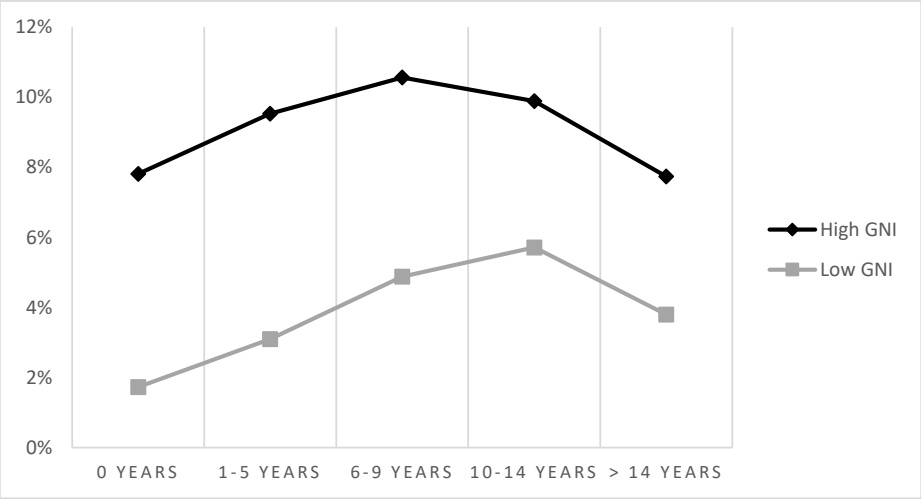

Figure 7.2: HIV prevalence and education - GNI subsamples

The first column of the Table 7.2 presents the results of logit estimation with HIV status as a dependent variable. For the regression we pooled all available data from 21 countries which translates into 250,057 observations. As the independent variables, we use schooling years together with the squared schooling years to allow for non-linear relation, age group dummies, religion dummies, wealth proxy, urban dummy and dummy for women together with the set of dummies for regions. We weight the data by frequency weights provided by the DHS and cluster the standard errors on the country level.

The McFadden's pseudo R^2 is relatively low as it equals 0.21. Nevertheless, we do not draw any conclusions from it due to the relative complexity of our problem and the fact that earlier works on similar topic also attained low levels

Table 7.2: Pooled observations regressions

	Dependent variables			
	HIV status	HIV knowledge	Use of condom	Number of sex partners
Schooling years	0.00468***	0.0261***	0.00924***	0.139**
	(0.000647)	(0.00121)	(0.000815)	(0.0429)
Schooling years squared	-0.000308***			
	(0.0000480)			
Age 15-25	-0.0656***	-0.0360*	0.258***	-4.688***
	(0.00588)	(0.0157)	(0.0116)	(0.981)
Age 26-35	-0.00535	0.0204	0.164***	-3.098***
	(0.00433)	(0.0136)	(0.0113)	(0.708)
Age 36-45	0.0114**	0.0254	0.113***	-2.369***
	(0.00427)	(0.0134)	(0.0118)	(0.605)
Age46-55	0.00440	0.0179	0.0680***	-1.353**
	(0.00452)	(0.0136)	(0.0145)	(0.386)
Women	0.0227***	-0.0233	-0.0947***	-3.963***
	(0.00220)	(0.0160)	(0.00588)	(0.638)
Urban	0.0343***	0.0550***	0.0431***	0.139
	(0.00387)	(0.00716)	(0.00768)	(0.126)
Wealth	-0.0297***	0.102***	0.0273*	-0.360
	(0.00827)	(0.0136)	(0.0110)	(0.349)
No religion	0.0151**	0.00896	0.0334***	0.345
	(0.00554)	(0.0131)	(0.00882)	(0.367)
Christian	0.00713	0.0607***	0.0337**	-0.811*
	(0.00684)	(0.0125)	(0.0108)	(0.299)
Islamic	-0.000857	0.0414**	0.0182	-1.089**
	(0.00729)	(0.0152)	(0.0113)	(0.382)
Stable sexual partner			-0.0103	
			(0.0136)	
Random sexual partner			0.118***	
			(0.0223)	
Sexual worker			0.187***	
			(0.0476)	
HIV status			0.0974***	
			(0.0102)	
Intercept				10.40***
				(1.086)
Regional dummies χ^2-test	1.0e+12	1.6e+10	4.6e+11	2.1e+05
Prob > χ^2	0.0000	0.0000	0.0000	0.0000
N	250,057	252,108	180,261	203,447
R squared	0.2084	0.1074	0.2389	0.1526

All logit regressions (except of regression for number of sexual partners for which OLS is used)
Robust and clustered standard errors in parentheses
** p<0.05, ** p<0.01, *** p<0.001*
Base group - men, more than 55 years old, living in rural area, other religion
Use of condom - base group is HIV negative and other partner (i.e. relative, other)

of R^2.

Age group dummies are significant only for groups 15-25 and 36-45 years. Nevertheless, the dummies are jointly significant with $\chi^2 = 374.23$ and p-value < 0.001. There is a negative correlation between age and HIV status for youngest people when the age group of 15-25 years is 6.6% less likely to be HIV positive compared to the oldest age group (people > 55 years old). The reason behind this effect could be in the fact that it takes some time to become HIV positive and the older the people get, the more chance they have to become infected (due to e.g. bad luck, more sexual partners, etc.). This would be partly offset by the mortality of the HIV infected people, but with modern medicine, people can live relatively long time being HIV positive. Thus, the older age groups could accumulate more HIV positive people (i.e. the survivorship bias).

As expected, we find positive and statistically significant coefficient at women dummy. Based on the AME, being a woman is associated with 2.3% higher probability of being HIV positive compared to men. This is in line with the overall worse situation of women in Sub-Saharan Africa regarding the HIV infection that was discussed in the previous chapters. Also, living in the city appears to be positively and significantly correlated with the HIV status (living in the city is associated with 3.4% higher probability of being HIV positive compared to living in the rural areas) which is in line with the assumption that people in the city live more dangerous lives regarding e.g. drugs or sexual activity. Nevertheless, wealth appears to be negatively correlated with the HIV status. The religion dummies are insignificant with exception of no religion. They are, however, jointly significant when $\chi^2 = 12.52$ and p-value is equal to 0.0058. Also, the regional dummies are jointly significant as can be seen from the table.

Regarding the education, both number of schooling years and number of schooling years squared is statistically significant at p-value 0.001. Also, the two variables are jointly significant at p-value 0.001. Thus, we can observe a non-liner relation between education measured by number of schooling years and the HIV status. The education appears to be overall positively correlated with the HIV status. The strength of the correlation ceteris paribus increases with increasing number of schooling years and it reaches its peak at around 9 years of schooling which corresponds approximately with finished primary school. Then, the correlation begins to decrease, For people with more than 16 years of schooling, the education appears to be negatively correlated with the HIV status (the Figure 7.3 presents the direction of the correlation for each

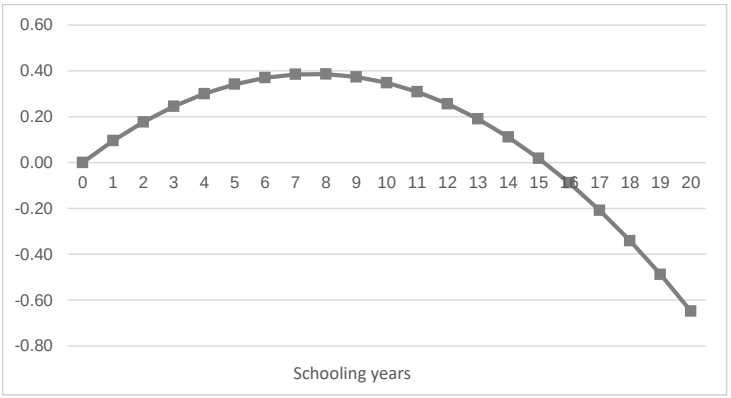

Figure 7.3: Education and HIV status correlation

year of schooling). Looking at the AME of education, the chances of being HIV positive are 1.7% higher for people with 9 years of education compared to people with no education. As for the people with higher education, the 16 years of schooling is associated with 0.4% lower probability of being HIV positive compared to people with no education.

The results show that the correlation between education and the HIV status is still somehow strange and difficult to grasp. From the shape of the correlation, we conclude that people with no education are at lower risk than the people with primary or secondary education. The reason for overall positive correlation between education and the HIV status could be explained by the fact that more educated people possess certain behaviour patterns that increase probability of becoming HIV positive, or that people who get more education, have certain characteristics that increase the probability of infection.

Possible explanation is that education can influence the different modes of transmission. According to Fortson (2008), drug use is very low in Africa, so it is unlikely to drive the results. The blood transfusions are not common enough to drive the results either. Also, the vertical transmission (through pregnancy) would not play important role as the coefficient is robust when excluding people younger than 35 years that could be affected by this as HIV become more spread only after 1980. Therefore, it appears that the effect of education could be somehow connected with the sexual behaviour patterns of the people with different level of education. Alternatively, the overall positive correlation between the HIV status and education could be truly caused by the survivorship bias. People with better education tend to have better jobs and thus more financial resources to pay for their treatment which can significantly prolong their life. Thus, there would be accumulated many HIV positive people

in categories that can afford the treatment. On the other hand, the poor people (i.e. people with not education) would have much larger mortality connected with the HIV infection due to the unavailability of treatment for them. It would then in turn decrease overall number of HIV positive people in the dataset.

Next, we proceed further to verify our hypotheses about the differences between high and low GNI and HIV prevalence countries. The countries are divided into the subsamples as presented in Chapter 5. The tables with results are for brevity presented in the Appendix only (Tables A.3 and A.4) and again contains AMEs rather than plain logit results.

Firstly, we examine whether the correlation between education and HIV status is consistent among the countries with different HIV prevalence. We run two separate regression for both subsamples. The results are provided in the Appendix in the Table A.3. The coefficients at education are significant in both regressions and positive. The coefficient at squared terms are negative and statistically significant also for both regressions, implying consistent direction of the relationship between HIV status and education with the pooled regression. The magnitudes of the coefficients appear to be similar.

To verify whether the effect of education is consistent across the two subsamples, we run additional regression that includes high HIV prevalence country dummy (equals to 1 if country is in the high HIV prevalence subsample) and an interaction term of education with high HIV prevalence country dummy (also for the squared term of education). The results of this regression can be found in the Table A.3 in the third column. Both interaction terms are not statistically significant. Moreover, the two interaction terms are not jointly significant when $\chi^2 = 1.36$ with p-value equal to 0.5077. Thus, there is no difference in the strength of the correlation between education and HIV status for countries with low and high HIV prevalence. This conclusion support our hypotheses that the education is a consistent factor connected with the HIV status among countries even with different HIV prevalence.

To see the effect of the country wealth on the magnitude and direction of the correlation between education and HIV status, we apply similar approach as above. We run two separate regressions for subsamples divided according to the World Bank threshold described in the Chapter 5. The results can be found in the Appendix in the Table A.4. The coefficients at education are significant in both regressions and positive. The coefficient at squared terms are negative and statistically significant also for both regressions, implying consistent direction of the relationship between HIV status and education with

the pooled regression. The magnitudes of the coefficients appear to be similar.

To see whether the effect of education differs across the subsamples, we estimate the regression with pooled data and with the high GNI country dummy and interaction terms of education and education squared with the high GNI country dummy. The results can be found in the Table A.4 in the third column. Both interaction terms are not statistically significant. Moreover, the two interaction terms are not jointly significant when $\chi^2 = 1.19$ with p-value equal to 0.55. Thus, there is no difference in the strength of the correlation between education and HIV status regarding wealth of the country. This conclusion support our hypotheses that the education is a consistent factor connected with the HIV status among countries even with different wealth.

7.2 HIV knowledge and education

As stated earlier, variable comprehensive knowledge about HIV is based on five variables describing knowledge about certain characteristics of the HIV/AIDS. The Figure 7.4 shows the level of knowledge about HIV for men and women in the dataset. The bars represent the share of the people in the dataset that answered 'correctly' to the answer regarding HIV knowledge. It appears that men have generally better knowledge about the HIV as for all questions, the share of correctly answered questions is higher for them than for women. This is in line with the fact that in Africa, men have generally easier access to education and other sources of information than women. The differences are not large though and for most of the variables, they are no larger than 5 percentage points with the exception of the knowledge about the condom as a possibility of prevention against HIV which is almost 10 percentage points larger for men than for women.

The biggest misconception about the HIV appears to be the transmission of the HIV by mosquito bites. Approximately 30% of people in the dataset believe to this statement. On the other hand, only around 15% believe that one can become HIV positive if they would share food with an HIV positive individual.

Although the share of people having answered to the questions correctly is relatively high, the share of population in the dataset that is considered to have a comprehensive knowledge about the HIV is only around 36% for women and 42% for men. Thus, it appears that even though people have some

information about the HIV transmission and infection in general, they lack more comprehensive knowledge about it.

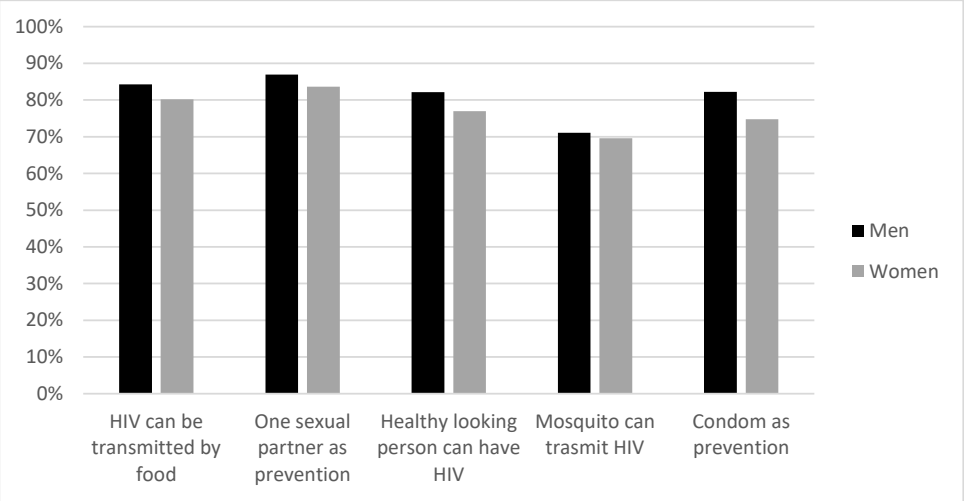

Figure 7.4: Individual HIV knowledge variables

The Table 7.3 presents education and comprehensive HIV knowledge two-way statistics in the similar way as above for education and HIV prevalence. There is a clear pattern that the groups of people with higher education attains higher shares of people that possess the comprehensive HIV knowledge. Thus, we expect to find positive linear relationship between education and HIV knowledge further in the regression. Interestingly, the percentage of people with comprehensive knowledge about the HIV is similar for men and women with at least some schooling. In the group of people with no education, men attain almost 7 percentage points larger share than women. Therefore, despite the fact that for all 5 individual HIV knowledge variables men attained better results, the comprehensive knowledge presents less variation between men and women.

Table 7.3: Education and comprehensive HIV knowledge two-way statistics

	Total number	HIV knowlege	% HIV knowledge
1. MEN			
0 years	18,897	7,073	27.2%
1-5 years	15,776	8,181	34.1%
6-9 years	22,500	18,437	45%
10-14 years	15,698	10,207	60.6%
> 15 years	3,354	1,296	72.1%
2. WOMEN			
0 years	36,399	9,545	20.8%
1-5 years	19,426	9,337	32.5%
6-9 years	23,977	19,399	44.7%
10-14 years	13,278	8,311	61.5%
> 15 years	1,897	710	72.8%

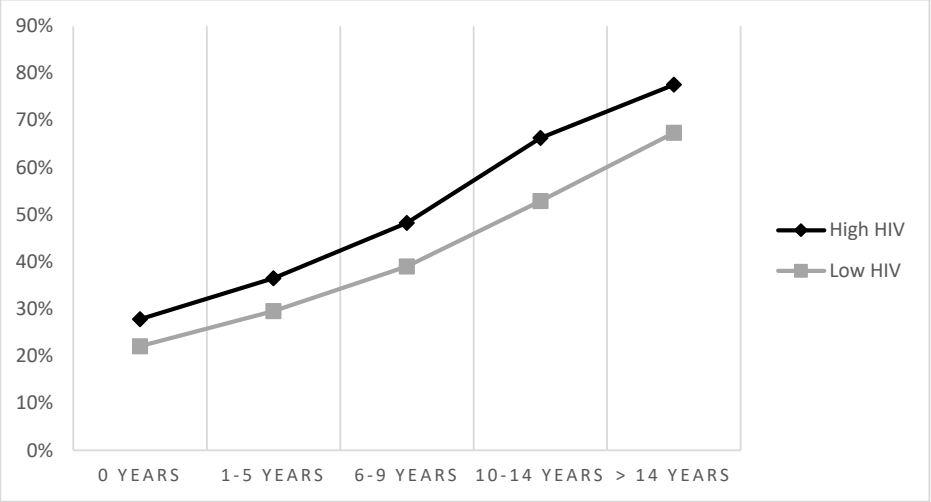

Figure 7.5: HIV knowledge and education - HIV subsamples

Later in the multivariate analysis, we analyse whether the effect of education is different for low and high HIV prevalence countries. The Figure 7.5 shows the HIV knowledge proportions based on the education level for the low and high HIV prevalence subsamples. It appears that the knowledge about the HIV is slightly higher in countries with higher HIV prevalence. Nevertheless, the effect of education appears to be very similar for both subsamples. Whether the effect of education would be statistically different will be evaluated further in this chapter.

The Table 7.2 in the second column presents the results (AMEs) of logit estimation with HIV knowledge variable as a dependent variable. The pooled dataset consists of 252,108 observations. We use the same independent variables as for the regression concerning HIV status. Including dummies regarding marital status as used for example by De Walque (2009) yield no significant effect and we thus did not include them into the regression. The McFadden's pseudo R^2 is relatively low as it equals 0.1074.

Age group dummies are significant only for group 15-25 when the coefficient is negative. We thus observe that being in the lowest age group is associated with the lower knowledge about the HIV infection compared to the rest of the dataset.

We do not observe any differences regarding the HIV knowledge that would be driven by the sex of the individuals as the coefficient at women dummy is insignificant. It is somehow against the generally accepted opinion that women in Africa have fewer possibilities regarding access to the information. On the other hand, both living in the city and wealth increase the probability of having comprehensive knowledge about the HIV (people living in the city have 5.5% higher chances of having HIV knowledge compared to the people living in the rural areas. Also, an increase in on asset used for wealth measure in associated with 1.2% increase in probability of having comprehensive HIV knowledge. Moreover, having all 8 assets relates to the 10% higher chances of having the HIV knowledge compared to people with no assets). These results appear to be consistent with the assumption that wealthier people can have better access to information as they can afford to have e.g. internet connection. Also, living in the city would certainly improve the access to the information e.g. through libraries or public events. Regarding the effect of religion, the no religion dummy and Christian dummy are positive and significant. On the other hand, being Muslim does not imply any effect on the HIV knowledge. The religion dummies are jointly significant when $\chi^2 = 27.10$ with p-value < 0.001. Also, the regional dummies are jointly significant as can be seen from the table.

The coefficient at education is statistically significant and positive which supports our hypothesis. Adjusting the results for the average marginal effects, the one year increase in schooling increase the probability that the individual would possess comprehensive knowledge about the HIV by 2.6%. 9 years of schooling which roughly corresponds with finished primary school is associated with 23% higher chances of having comprehensive knowledge about the HIV

compared to people with no education. Thus, it appears that education is very important determinant of the HIV knowledge.

The Table A.5 in the Appendix presents the results of the regression with HIV knowledge as independent variable for two subsamples, one for countries with HIV prevalence below 1.8% and one for countries with HIV prevalence above 1.8%. The coefficients at education are positive and statistically significant for both regressions which is in line with our previous regression using whole sample.

To test the statistical difference, we again include the high HIV prevalence country dummy and interaction term into the regression with pooled data. The interaction term is not statistically significant implying that there are no differences between the effect of education on HIV knowledge for low and high HIV prevalence countries. This finding is against the hypothesis where we anticipated differences in the effect of education (see Chapter 4). Also, the coefficient at high HIV prevalence country dummy is not statistically significant meaning that overall HIV prevalence has not effect on the HIV knowledge in the country.

7.3 Sexual behaviour and education

The two-way statistics for education and proxy for sexual behaviour, i.e. whether respondent used condom during the last sexual intercourse during the past 12 months, is presented in Table 7.4. There is a large difference between men and women regarding the use of condom during the last sexual intercourse which can be explained e.g. by the fact that men tend to have more random sexual partners compared to women. Nevertheless, there is an overall trend that within groups of people with higher education, the share of people that used condom during the last sexual intercourse increases. The differences are relatively profound when the people with highest education used condom during the last sexual intercourse in 25% and 31% cases for women and men, respectively, whereas for people with no education, the use of condom is only in 2.3% cases for women and 7.2% cases for men.

The Figure 7.6 presents average number of sexual partners for each group of education for both men and women. There is an upward trend regarding number of sexual partners for men with increasing number of schooling years when men with no schooling have on average slightly above 4 sexual partners, but with 9 years of schooling, the number almost doubles. On the other hand,

Table 7.4: Education and use of condom two-way statistics

	Total number	Condom use	% Condom use
1. WOMEN			
0 years	34,705	808	2.3%
1-5 years	18,577	1,578	7.8%
6-9 years	24,277	4,766	16.4%
10-14 years	10,842	3,583	24.8%
> 15 years	1,459	481	24.8%
2. MEN			
0 years	18,949	1,480	7.2%
1-5 years	13,688	2,460	15.2%
6-9 years	20,490	6,921	25.2%
10-14 years	13,154	6,230	32.1%
> 15 years	2,749	1,212	30.6%

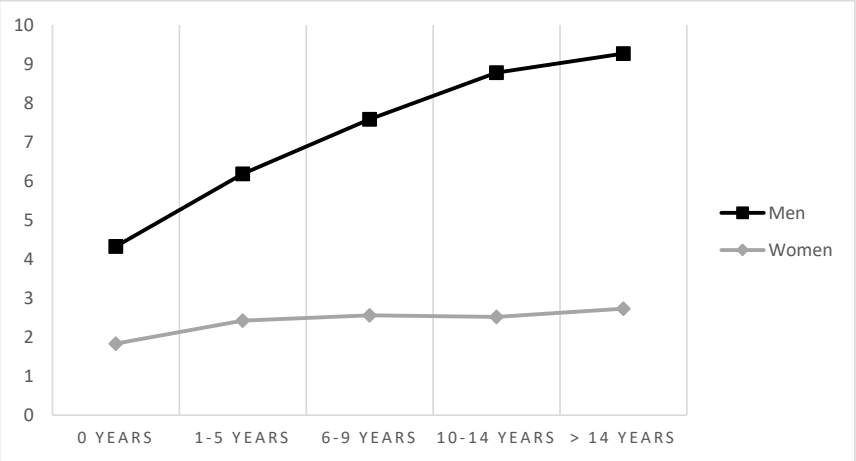

Figure 7.6: Education and sexual partners

for women the number of sexual partners seems to remain more or less constant for different levels of education at 2 sexual partners on average.

The Table 7.2 in the third and fourth column provides the results of the last two regressions concerning relation of education and sexual behaviour. As the dependent variables, we use number of sexual partners and binomial variable whether respondent used condom during the last sexual intercourse. We use the same independent variables as in the case of HIV knowledge regression when using number of sexual partners as a dependent variable. When using use of condom as a dependent variable, we control also for the nature of the last sexual intercourse, i.e. we whether the last sexual partner was a stable partner, a random partner, a sex worker or other (omitted in the regression).

Moreover, we also control for whether the respondent is HIV positive which certainly could affect the use of condom to protect the partner. There are 180,261 observations for the pooled data for condom use regression and 203,447 observations for number of sexual partners regression. As the condom use is a binomial variable, we use again logistic estimation. The number of sexual partners is nevertheless in range from 1 to 95. For the estimation, we use ordinary least square regression.

Regarding the use of condom, the age group dummies are all significant. Based on the coefficients, it appears that the use of condom is more common for younger groups of people. This can be explained by the fact that younger people also get involved into e.g. random sex compared to older groups where people are usually married. Based on AME, the probability of using condom is 25% higher for people from the youngest age group compared to the base group of oldest respondents. The coefficient at women dummy is negative and significant, based on AME, the women were 9% less likely to use condom during the last sexual intercourse compared to men. It somehow does not correspond with the fact that the use of condom should be similar for women and men. Nevertheless, the difference could be explained by the general larger promiscuity among men compared to women (e.g. in the dataset, the number of men that reported last sexual intercourse with sexual worker is 405 whereas 60 for women). Living in the city is positive and significant with AME 0.043. The wealth index is insignificant which can be possibly explained by the good availability of condoms in Africa as they are often handed for free by e.g. volunteers or health organizations. The religion dummies are jointly significant when $\chi^2 = 14.46$ with p-value 0.0018. Also, the regional dummies are jointly significant as can be seen from the table.

The dummies for last sexual partner are jointly significant when $\chi^2 = 47.43$ with p-value < 0.001. The use of condom is insignificant when the last sexual partner was a stable one. On the other hand, random partner significantly increases the probability of using condom as well as sexual worker (AME 0.12 and 0.19 respectively compared to the base group).

Education is positively associated with the use of condom. Adjusting the results for the average marginal effects, the one year increase in schooling increase the probability that the individual used condom during the last sexual intercourse by 0.9%. This supports the hypothesis of education being a protective measure regarding the sexual behaviour.

As for the number of sexual partners, the effect of age dummies is as ex-

pected, i.e. increasing age is associated with the increasing number of sexual partners. All age dummies are significant at 0.01 significance level. Also the coefficient at women dummy is significant and negative. Thus, being a woman is, ceteris paribus, associated with having four less sexual partners compared to men. This can be explained similarly to the differences for the use of condom above, i.e. by the higher promiscuity of men. Also, unfortunately, the real number of sexual partners could have been exaggerated by men and underrated by women. Living in the city and wealth seems to have no effect on the number of sexual partners as both coefficients are insignificant. The religion dummies are jointly significant $F = 17.52$ with p-value < 0.001. Nevertheless, the effect of religion is small as e.g. being a Christian implies 0.8 less sexual partners compared to the base group (other religion). Also, the regional dummies are jointly significant as can be seen from the table.

The effect of education is positive and significant. However, not very strong. Each additional year of schooling increases the number of sexual partners by 0.139 (10 years of schooling thus implies 1.4 more sexual partner compared to people with no education). Nevertheless, the higher education seems to bring some additional risks as more sexual partners certainly imply higher exposure to the sexually transmitted diseases.

Chapter 8

Robustness test

In this chapter we conduct several robustness tests to support our findings from the previous chapter. Specifically, we use alternative measures for education in the regressions and we estimate the equations using different estimation techniques. For brevity, we comment and present only the main and important results of the robustness test. The results of the main regressions can be then found in the Appendix.

Firstly, we focus on the education measurement. In the previous regressions, we used as a measure for education number of completed schooling years. The DHS dataset offers additional variable stating highest educational attainment. It specifies whether respondent has no schooling, finished primary school, finished secondary school or finished some higher education.

We would expect the correlation between number of schooling years and highest attained education be very close to one as the variables are in their nature very similar. In fact, the correlation is 0.94. We substitute the highest attained education into the regression to check the consistency of the results. We include the highest education level into the regression in a form of three dummies for primary, secondary and higher education with omitted variable no education. The results with pooled data are presented in Table A.6.

The first column presents the logistic regression with HIV status as a dependent variable (again, AMEs are presented). The coefficients for women, urban and wealth are similar in terms of significance and magnitude as for the base regression. The dummies for religion, age and region are not reported (reported is only joint significance). The education dummies are jointly significant when $\chi^2 = 54.57$ with p-value lower than 0.001. For the individual dummies, only primary school education and secondary school education is statistically

significant. The dummy for higher education that corresponds with the most educated part of the population appears insignificant. The primary and secondary school dummies are positive which is in line with the previous analysis. The results thus correspondes to the results from the base regression with non-linear relationship.

The second column of the Table A.6 presents the output of regression with the HIV knowledge as dependent variable. The linear relation between education and HIV knowledge is apparent similarly to the base regression. All education level coefficients are significant and positive with increasing magnitude for higher educational levels. Also, other variables remained with the same significance and direction as for the base regression.

The third and fourth column of the Table A.6 present the regression output with last condom use and number of sexual partner as dependent variables, respectively. The relation between the use of condom and education appears to be consistent with the previous regression as well as relation between number of sexual partners and education. For all education dummies, the coefficients are significant, positive and increasing in magnitude with increasing level of education. The other coefficients at urban, women and wealth are consistent with the base regressions regarding significance and direction.

To further check for robustness of education measure, we use also literacy as a proxy for education (results shown in the Table A.7). The first column shows the results of logit regression (AMEs are again presented) regarding the correlation between literacy and HIV status. The coefficient at literacy is significant and positive which somehow validates the earlier results when the education is generally positively correlated with HIV status. All other coefficients are also in line with previous findings.

Regarding the HIV knowledge, the results are also generally consistent with the previous results. Also for the use of condom during the last sexual intercourse, the coefficient is significant and positive. The literacy has, even though significant, only very little impact on number of sexual partners when being literate is associated with 0.7 more sexual partners compared to illiterate people.

We also run the regressions with interaction terms to find whether the differences between low and high HIV prevalence and low and high GNI countries are still insignificant. For brevity, the results are not reported. All interaction terms for all three regressions remained insignificant, meaning that with the use of different education measure (both highest educational attainment and

literacy), the correlation between education and HIV status and education and HIV knowledge remain the same across the subsamples.

To show that the results are robust to the using different estimation technique, i.e. probit instead of logit, we re-estimated all regressions using probit. For brevity, the results are not reported, however, we did not observe any major differences between the two set of results.

The differences between men and women appear to be relatively large as can be seen from the basic statistics presented in the previous chapter. Thus, we also run regressions for subsamples of only men and only women. Consequently, we tested whether the association between education and the variables of our interest vary between men and women. As this analysis is not the main topic for the thesis, the results are presented in Appendix only. The Tables A.8, A.9, A.10 and A.11 present the results from the regressions with dependent variables HIV status, HIV knowledge, use of condom and number of sexual partners, respectively. The results from logit regressions are presented in a form of AME.

From the results, we see that the direction of the education AMEs is same for men and women in all regressions. Moreover, the education AMEs are also significant for both subsamples. Nevertheless, the magnitude of AME differs as e.g. for women, the correlation between education and HIV status is somehow larger for women than for men. Also the effect of education on condom use is stronger for women than for men when e.g. having 10 years of schooling implies 7.8% larger chance of using condom during the past sexual intercourse for men and almost 11% for women compared to respondents with no education. The education has a very different effect also on number of sexual partners when for women, the 10 years of education is associated only with 0.27 more sexual partners whether for men, it is associated with 2.1 more sexual partners.

Based on the presented robustness test, we can conclude that generally, our results are consistent when using different educational measures both on the pooled data level and on the testing for differences between the subsamples of the high and low GNI per capita and HIV prevalence countries. Moreover, the results are also robust for using different estimation technique, i.e. probit.

Chapter 9

Discussion of the results

The presented analysis, similarly to other analyses that use cross-sectional data, is certainly not perfect. The dependent variables could be influenced by variety of observed and unobserved effects we were not able to control for due to the dataset limitations. The analysis should be interpreted with caution and with the limitations in mind. It can, however shed some light onto the drivers and factors that relate to the HIV/AIDS infection in Sub-Saharan Africa.

Regarding the results, we found a quadratic relation regarding the correlation between the education and the HIV status (similarly to e.g. Fortson (2008)). For people with no education, the correlation is smaller than for those with up to 16 years of schooling. More than 16 years of schooling on the other hand appear to negatively correlated with the HIV status. Thus, we believe that more attention should be paid to the groups of people at the highest risk, i.e. people with approximately primary education as the peak of the quadratic relation is around 9 years of schooling. The help should be targeted to those group especially to prevent the further spread of the disease and to help the families of the affected individuals. Also, the higher education such as college or university, should be further promoted to fight the HIV infection. The findings about the correlations are further supported by the robustness tests.

What causes the correlation and education to be overall positively correlated? Possible explanation can be e.g. that people with some education are living in the generally different environment compared to people with no education and that this environment is more prone to the HIV transmission. On the other hand, people with more education tend to have somehow safer sexual behaviour as education is a predictor for using condom. Or, as stated earlier, HIV positive people with better education could live longer as they usually

have had better jobs and overall better social position before they became HIV positive. Thus, they can afford the treatment that can significantly prolong the life expectancy of the HIV positive patients. If we accept the fact that people with no education would have more difficulties to obtain such treatment, they would have generally higher mortality due to the HIV. This would in turn decrease the number of respondents within the dataset that are HIV positive and have no education.

The results regarding HIV status and education appear to be consistent among countries with different HIV prevalence and different GNI. Thus, the highest risk groups in respect to education are similar for different countries in Sub-Saharan Africa.

The education appears to be the strong predictor of the HIV knowledge. More education translates into the higher chances of the individuals to possess the comprehensive knowledge about the HIV which is in line with the current literature. It is thus desired to further promote education in the Sub-Saharan Africa by which the HIV knowledge would spread into the population. The focus should be especially targeted to the people that have little possibilities to attain some school education as the HIV knowledge in this social group is very small (the share of people with comprehensive knowledge about the HIV is 50 percentage points smaller for people with no education compared to people with highest education). The knowledge about the HIV is important for further development in Sub-Saharan Africa in many ways as it increases the understanding about the disease that could help to erase stigma carried by the HIV positive people that is very common in the region, could lead to better sexual behaviour and could increase probability that people would be interested in their HIV status and go to testing places.

Lastly, we evaluated the effect of education on the sexual behaviour, using proxy of condom use and number of lifetime sexual partners. Here, we found rather contradicting results when education has positive effect on both use of condom and number of sexual partners. Nevertheless, the effect of education on number of sexual partners is rather small (the difference between no education and 10 years of education is only 1.4 sexual partners). The effect of condom use is, on the other hand, rather large and we believe that more important. People with 10 years of schooling used 10% more likely condom during the last sexual intercourse compared to people with no education. Also, looking at the simple two-way statistics, only 2.3% of women and 7.2% of men with no education used condom during the last sexual intercourse compared to 24% of

women and 30% of men with more than 15 years of schooling. As the use of condom is not connected with the wealth of the individuals, we believe that the access to condoms is relatively easy. Thus, we believe that further promotion of education could have a positive effect on the use of condom that would in turn prevent the HIV/AIDS to spread.

The analysis could be improved by adding some variables that were not available in the DHS dataset. Firstly, we believe that the accuracy of the analysis could be improved by including the variables describing ethnicity. Some ethnics could be more prone to the HIV infection and possess different behavioural patterns regarding sexual behaviour than others due to e.g. different social status of the ethnics in the society or by different culture. Although the ethnicity variable is available in the datasets, the information is missing for many countries (in the case of countries used in the thesis, the ethnicity variable is not available for 8 countries out of 21). Therefore, we suggest to focus on the ethnicity variable in the further DHS data collection.

Moreover, it would be beneficial to collect information about the education towards the sexual behaviour and HIV/AIDS in general. It could be interesting to control in the regressions also for variable that would describe whether the respondent had e.g. sexual education during his school years or whether they received some education from e.g. volunteer workers or state organizations regarding sexual protection and HIV. This would then allow to see more precise effect of education as sexual education does not necessarily need to be a part of a formal education. Thus, we would suggest to add the variables concerning above described topic into the next DHS data collection as it could bring additional insight into the HIV/AIDS problem.

Lastly, it would be beneficial for further research to implement also questions regarding the ART treatment, e.g. to implement questions whether respondent knows their HIV status and if positive, whether they have access to the ART. These variables could be then used to control for in the regressions and improve the results of the analysis. Also, we would be able to determine which groups of population lack access to the ART and then target the help directly to those groups.

Chapter 10

Conclusion

In the thesis we evaluated the relationship between education and several factors connected with the HIV infection, namely HIV status, HIV knowledge and sexual behaviour measured by proxies such as the number of sexual partners and the use of condom during the last sexual intercourse. For the analysis we used large dataset from the Demographic and Health Survey with data from 21 countries in Sub-Saharan Africa.

The study was motivated by the heterogeneity of the previous results on the similar topics when especially the correlation between HIV status and education appears to differ significantly using different datasets and estimation techniques. Also, as the new data collection took place in the years 2008-2014, we were able to take advantage of the new dataset.

Using the econometric analysis, we found non-linear relationship between HIV status and education, when primary and secondary schooling is associated with higher risk of being HIV positive compared to no schooling. Special focus should be therefore targeted to those groups to prevent further spread of the HIV and to help the families of the infected individuals. The higher education (i.e. university) appears to be negatively correlated. Moreover, we found that the effect of education is consistent across countries with different GNI per capita and different overall HIV prevalence.

Additionally, the thesis shows that education is an important predictor of the comprehensive HIV knowledge. Also, the education seems to increase the probability of condom use. At the same time, more schooling years are associated with the higher number of sexual partners, however the effect is rather small as 10 additional years of schooling increase ceteris paribus the number of sexual partners by only 1.4.

The analysis is subject to several limitations that could have significant impact on the results. However, we believe that the presented results can at least serve as an indicator of which social groups are in the greatest threat of the HIV infection and that the help can be directly and efficiently targeted to those groups. The education appears to be important in the fight against the HIV infection as it is positively related to both condom use and HIV knowledge. Thus, further promotion of education in Sub-Saharan Africa is highly recommended to help to prevent the further spread of the HIV.

We believe that additional research should be done periodically using more and more accurate datasets to monitor the changes in the HIV and socio-economic correlates patterns with keeping in mind all limitations that are present when evaluating this relationship. Even though the ART is being more available and affordable even for poor people, the fight against HIV should still be done also through understanding the socio-economic mechanism that have influence on the HIV status and factors connected to the epidemic.

Bibliography

Agüero, J. M. & P. Bharadwaj (2014): "Do the more educated know more about health? Evidence from schooling and HIV knowledge in Zimbabwe." *Economic Development and Cultural Change* **62(3)**: pp. 489–517.

AVERT (2016): "Women and HIV/AIDS." `http://www.avert.org/professionals/hiv-social-issues/key-affected-populations/women`. Accessed: 10-12-2016.

Bachmann, M. O. & F. L. Booysen (2003): "Health and economic impact of HIV/AIDS on South African households: a cohort study." *BMC Public Health* **3(1)**: p. 1.

Baker, D. P., J. Leon, & J. M. Collins (2011): "Facts, attitudes, and health reasoning about HIV and AIDS: explaining the education effect on condom use among adults in sub-Saharan Africa." *AIDS and Behavior* **15(7)**: pp. 1319–1327.

Béchu, N. (1998): "The impact of AIDS on the economy of families in Côte d Ivoire: Changes in consumption among AIDS-affected households." *Confronting AIDS: Evidence from the developing world* pp. 241–253.

Case, A., C. Paxson, & J. Ableidinger (2004): "Orphans in Africa: parental death, poverty, and school enrollment." *Demography* **41(3)**: pp. 483–508.

Corno, L. & D. De Walque (2007): "The determinants of HIV infection and related sexual behaviors: Evidence from Lesotho." *World Bank Policy Research Working Paper, No. 4421* .

De Walque, D. (2009): "Does education affect HIV status? Evidence from five African countries." *The World Bank Economic Review* **23(2)**: pp. 209–233.

Dixon, S., S. McDonald, & J. Roberts (2001): "AIDS and economic growth in Africa: a panel data analysis." *Journal of International Development* **13(4)**: pp. 411–426.

Fako, T. T., L. W. Kangara, & N. Forcheh (2010): "Predictors of knowledge about HIV/AIDS among young people: Lessons from Botswana." *Journal of AIDS and HIV Research* **2**: pp. 116–130.

Fortson, J. G. (2008): "The gradient in Sub-Saharan Africa: Socioeconomic status and HIV/AIDS." *Demography* **45(2)**: pp. 303–322.

Fox, A. M. (2010): "The Social Determinants of HIV Serostatus in Sub-Saharan Africa: An Inverse Relationship between Poverty and HIV?" *Public Health Reports* **125(4 suppl)**: pp. 16–24.

Fylkesnes, K., R. M. Musonda, M. Sichone, Z. Ndhlovu, F. Tembo, & M. Monze (2001): "Declining HIV prevalence and risk behaviours in Zambia: evidence from surveillance and population-based surveys." *Aids* **15(7)**: pp. 907–916.

Glynn, J. R., M. Carael, A. Buve, S. Anagonou, L. Zekeng, M. Kahindo, & R. Musonda (2004): "Does increased general schooling protect against HIV infection? A study in four African cities." *Tropical Medicine and International Health* **9(1)**: pp. 4–14.

Greener, R. (2002): "AIDS and macroeconomic impact." `http://www.policyproject.com/pubs/other/SOTAecon.pdf`. Special series State of the Art: AIDS and Economics prepared for the International AIDS Economics Network, Accessed: 24-10-2016.

Hargreaves, J., L. Morison, J. Kim, C. Bonell, J. Porter, C. Watts, J. Busza, G. Phetla, & P. Pronyk (2008): "The association between school attendance, HIV infection and sexual behaviour among young people in rural South Africa." *Journal of epidemiology and community health* **62(2)**: pp. 113–119.

Igwike, R. S. & M. E. Hussain (2012): "Examining the impact of HIV-Prevalence on economic growth in Sub-Saharan Africa: a panel data analysis." *Available at SSRN 2087874* .

Kirigia, J. M., L. G. Sambo, T. Okorosobo, & G. M. Mwabu (2001): "Impact of HIV/AIDS on Gross Domestic Product (GDP) in the WHO Africa Region." *African journal of health sciences* **9(1)**: pp. 27–39.

Lagarde, E., M. Caraël, J. R. Glynn, L. Kanhonou, S.-C. Abega, M. Kahindo, R. Musonda, B. Auvert, A. Buvé, S. G. on the Heterogeneity of HIV Epidemics in African Cities *et al.* (2001): "Educational level is associated with condom use within non-spousal partnerships in four cities of sub-Saharan Africa." *Aids* **15(11)**: pp. 1399–1408.

Long, J. S. & J. Freese (2006): *Regression models for categorical dependent variables using Stata*. Stata press.

Mahajan, A. P., J. N. Sayles, V. A. Patel, R. H. Remien, D. Ortiz, G. Szekeres, & T. J. Coates (2008): "Stigma in the HIV/AIDS epidemic: a review of the literature and recommendations for the way forward." *AIDS (London, England)* **22(Suppl 2)**: p. S67.

Mwakalobo, A. B. S. (2007): "Implications of HIV/AIDS for Rural Livelihoods in Tanzania: The Example of Rungwe District." *African Studies Review* **50(3)**: pp. 51–73.

Mwamwenda, T. S. & P. W. Kariuki (2014): "University of Nairobi Students Level of HIV/AIDS Knowledge." *Mediterranean Journal of Social Sciences* **5(27 P1)**: pp. 500–505.

Nagot, N., A. Ouangré, A. Ouedraogo, M. Cartoux, P. Huygens, M. C. Defer, T. Zékiba, N. Meda, & P. Van de Perre (2002): "Spectrum of commercial sex activity in Burkina Faso: classification model and risk of exposure to HIV." *Journal of acquired immune deficiency syndromes (1999)* **29(5)**: pp. 517–521.

Nasir, E. F., A. N. Åstrøm, J. David, & R. W. Ali (2008): "HIV and AIDS related knowledge, sources of information, and reported need for further education among dental students in Sudan-a cross sectional study." *BMC Public Health* **8(1)**: p. 1.

Negin, J., B. Nemser, R. Cumming, E. Lelerai, Y. B. Amor, & P. Pronyk (2012): "HIV attitudes, awareness and testing among older adults in Africa." *AIDS and Behavior* **16(1)**: pp. 63–68.

Rutstein, S. O. & G. Rojas (2006): "Guide to DHS statistics." *Calverton, MD: ORC Macro* .

Shelton, J. D., M. M. Cassell, & J. Adetunji (2005): "Is poverty or wealth at the root of HIV?" *The Lancet* **366(9491)**: pp. 1057–1058.

Smith, J., F. Nalagoda, M. J. Wawer, D. Serwadda, N. Sewankambo, J. Konde-Lule, T. Lutalo, C. Li, & R. H. Gray (1999): "Education attainment as a predictor of HIV risk in rural Uganda: results from a population-based study." *International Journal of Std & Aids* **10(7)**: pp. 452–459.

Smith, W., D. Salinas, & D. P. Baker (2012): "Multiple effects of education on disease: The intriguing case of HIV/AIDS in Sub-Saharan Africa." *The Impact of HIV/AIDS on Education Worldwide (International Perspectives on Education and Society). Emerald Group Publishing Limited* pp. 79–104.

StataCorp. (2013): *Stata Statistical Software: Release 13.* College Station, TX: StataCorp LP.

UNAIDS (2012): "HIV treatment now reaching more than 6 million people in sub-Saharan Africa." `http://www.unaids.org/sites/default/files/web_story/20120706_PR_africatreatment_en_0.pdf`. Accessed: 14-12-2016.

UNAIDS (2016a): "Fact sheet November 2016." `http://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf`. Accessed: 02-01-2017.

UNAIDS (2016b): "Fast-track commitments to end AIDS by 2030." `http://www.unaids.org/sites/default/files/media_asset/fast-track-commitments_en.pdf`. Accessed: 02-01-2017.

Wooldridge, J. M. (2015): *Introductory econometrics: A modern approach.* Nelson Education.

World Bank (2017): "World Bank Country and Lending Groups." `https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups`. Accessed: 02-01-2017.

Appendix A

Additional tables

Table A.1: HIV testing response rates

Country	Female				Male			
	No educ.	Primary	Secondary	Higher	No educ	Primary	Secondary	Higher
Burkina Faso	96.7%	96.6%	95.8%	89.7%	94.7%	93.7%	93.6%	83.0%
Burundi	94.0%	91.0%	85.6%	83.7%	90.2%	87.2%	82.2%	85.0%
Cameroon	93.4%	95.5%	95.1%	89.7%	90.5%	94.7%	94.1%	87.9%
Congo (Dem. Rep.)	97.5%	97.0%	95.2%	87.9%	92.1%	95.0%	93.8%	87.6%
Ethiopia	90.5%	90.1%	84.3%	80.4%	82.6%	84.2%	74.9%	74.7%
Gabon	97.5%	97.9%	96.7%	89.0%	91.6%	97.4%	95.4%	89.3%
Gambia	85.2%	88.3%	84.3%	72.0%	66.6%	76.2%	74.6%	67.5%
Ghana	95.0%	94.7%	93.0%	88.7%	88.9%	90.1%	90.0%	81.4%
Guinea	97.4%	97.5%	96.7%	97.1%	96.4%	96.1%	95.1%	94.1%
Kenya	83.5%	88.7%	87.8%	83.2%	69.0%	82.0%	82.2%	77.3%
Lesotho	84.2%	95.3%	94.1%	87.7%	88.6%	89.0%	87.5%	83.7%
Liberia	90.5%	91.9%	89.4%	76.3%	84.8%	89.1%	86.3%	83.7%
Malawi	91.1%	91.7%	90.1%	87.0%	80.5%	85.1%	84.7%	77.0%
Mali	91.9%	92.4%	91.1%	79.5%	79.1%	78.5%	75.5%	68.6%
Namibia	91.4%	91.5%	87.8%	73.8%	85.6%	84.2%	78.7%	66.1%
Rwanda	99.8%	99.7%	99.4%	96.7%	99.4%	99.6%	98.9%	94.5%
Senegal	83.8%	87.2%	90.1%	73.3%	74.3%	80.4%	83.9%	82.7%
Sierra Leone	94.4%	94.1%	94.9%	91.2%	90.7%	91.3%	90.7%	88.9%
Togo	95.2%	96.1%	95.7%	86.8%	91.4%	93.1%	93.5%	85.4%
Zambia	92.4%	95.0%	93.2%	87.2%	90.0%	93.6%	91.6%	83.4%
Zimbabwe	83.1%	85.0%	80.6%	75.2%	72.7%	77.3%	69.3%	62.1%

Source: DHS dataset, author's calculations

Table A.2: GNI per capita and HIV prevalence

Country	GNI per capita (US dollars)	HIV prevalence
Burkina Faso	640	0.8%
Burundi	260	1.0%
Cameroon	1,320	4.5%
Congo (Dem. Rep.)	410	0.8%
Ethiopia	590	1.2%
Gabon	9,200	3.8%
Gambia	460	1.8%
Ghana	1,480	1.6%
Guinea	470	1.6%
Kenya	1,340	5.9%
Lesotho	1,280	22.7%
Liberia	380	1.1%
Malawi	340	9.1%
Mali	760	1.3%
Namibia	5,190	13.3%
Rwanda	700	2.9%
Senegal	980	0.5%
Sierra Leone	620	1.3%
Togo	540	2.4%
Zambia	1,490	12.9%
Zimbabwe	860	16.7%

Data about GNI per capita (by Atlas method) from World Bank database for year 2015

Data about HIV prevalence from UNAIDS database for year 2015

Table A.3: HIV status and education - HIV subsamples

	HIV status		
	High HIV prevalence	Low HIV prevalence	Whole dataset
Schooling years	0.00800***	0.00129*	0.00538*
	(0.000868)	(0.000653)	(0.00244)
Schooling years squared	-0.000509***	-0.0000991*	-0.000293*
	(0.0000564)	(0.0000445)	(0.000138)
Age 15-25	-0.115***	-0.0182***	-0.0656***
	(0.0113)	(0.00424)	(0.0126)
Age 26-35	-0.00958	-0.00153	-0.00417
	(0.00733)	(0.00517)	(0.00749)
Age 36-45	0.0209**	0.00210	0.0125
	(0.00706)	(0.00517)	(0.00776)
Age46-55	0.0109	-0.00241	0.00528
	(0.00786)	(0.00474)	(0.00810)
Women	0.0375***	0.00839***	0.0259***
	(0.00382)	(0.00142)	(0.00239)
Urban	0.0551***	0.0142***	0.0284**
	(0.00730)	(0.00317)	(0.00945)
Wealth	-0.0563***	-0.00328	-0.0447**
	(0.0163)	(0.00443)	(0.0149)
No religion	0.0289**	-0.000975	0.0263**
	(0.00983)	(0.00453)	(0.00956)
Christian	0.0154	-0.00189	0.0223*
	(0.0125)	(0.00409)	(0.0108)
Islamic	0.00861	-0.00613	0.00389
	(0.0113)	(0.00475)	(0.00972)
High HIV country			0.0761***
			(0.0229)
Schooling years*High HIV country			0.00329
			(0.00276)
Schooling years sq.*High HIV country			-0.000161
			(0.000145)
Regional dummies χ^2-test	3.1e+12	1.5e+10	
Prob > χ^2	0.0000	0.0000	
N	123,945	128,112	252,057
R squared	0.1465	0.0823	0.1519

Robust and clustered standard errors in parentheses
** p<0.05, ** p<0.01, *** p<0.001*
Base group - men, more than 55 years old, living in rural area, other religion

Table A.4: HIV status and education - GNI subsamples

	HIV status		
	High GNI countries	Low GNI countries	Whole dataset
Schooling years	0.00727***	0.00327***	0.0133**
	(0.00118)	(0.000730)	(0.00483)
Schooling years squared	-0.000490***	-0.000209***	-0.000666**
	(0.0000511)	(0.0000483)	(0.000223)
Age 15-25	-0.111***	-0.0391***	-0.0611***
	(0.0120)	(0.00976)	(0.0136)
Age 26-35	-0.0107	0.00193	0.00279
	(0.00727)	(0.0105)	(0.00826)
Age 36-45	0.0154*	0.0142	0.0182
	(0.00666)	(0.0106)	(0.00953)
Age46-55	0.00427	0.00922	0.0113
	(0.00719)	(0.0108)	(0.00944)
Women	0.0362***	0.0166***	0.0270***
	(0.00533)	(0.00162)	(0.00496)
Urban	0.0485***	0.0276***	0.0212*
	(0.0119)	(0.00298)	(0.00912)
Wealth	-0.0524*	-0.0186***	-0.0387*
	(0.0219)	(0.00479)	(0.0178)
No religion	0.0168	0.00782*	0.0204
	(0.0164)	(0.00375)	(0.0137)
Christian	0.0197	-0.000475	0.0110
	(0.0160)	(0.00401)	(0.0104)
Islamic	0.00639	-0.00525	-0.0328
	(0.0182)	(0.00553)	(0.0211)
Rich country			0.0642**
			(0.0229)
Schooling years*Rich country			-0.00597
			(0.00408)
Schooling years sq.*Rich country			0.000215
			(0.000189)
Regional dummies χ^2-test	4.8e+11	6.6e+09	
Prob > χ^2	0.0000	0.0000	
N	83,254	168,803	252,057
R squared	0.1466	0.2125	0.1168

Robust and clustered standard errors in parentheses
** p<0.05, ** p<0.01, *** p<0.001*
Base group - men, more than 55 years old, living in rural area, other religion

Table A.5: HIV knowledge and education - HIV subsamples

	HIV knowledge		
	High HIV prevalence	Low HIV prevalence	Whole dataset
Schooling years	0.0303***	0.0222***	0.0215***
	(0.00190)	(0.00139)	(0.00250)
Age 15-25	-0.00112	-0.0565*	-0.0241
	(0.0206)	(0.0221)	(0.0159)
Age 26-35	0.0586***	-0.00530	0.0279
	(0.0141)	(0.0177)	(0.0145)
Age 36-45	0.0653***	-0.00333	0.0283*
	(0.00732)	(0.0155)	(0.0139)
Age46-55	0.0627***	-0.0155	0.0222
	(0.00842)	(0.0152)	(0.0144)
Women	0.0126	-0.0614**	-0.0245
	(0.0108)	(0.0212)	(0.0155)
Urban	0.0521***	0.0595***	0.0309*
	(0.00618)	(0.0122)	(0.0123)
Wealth	0.0969***	0.108***	0.108***
	(0.0198)	(0.0188)	(0.0229)
No religion	0.0183	-0.0506	-0.0347
	(0.0116)	(0.0410)	(0.0197)
Christian	0.0453**	0.0793***	0.0542*
	(0.0156)	(0.0209)	(0.0221)
Islamic	0.0337*	0.0548*	0.00727
	(0.0166)	(0.0252)	(0.0382)
High HIV country			0.0380
			(0.0623)
Schooling years*High HIV country			0.00514
			(0.00481)
Regional dummies χ^2-test	3.2e+11	7.5e+10	
Prob > χ^2	0.0000	0.0000	
N	123,958	128,150	252,108
R squared	0.0837	0.0972	0.0777

Robust and clustered standard errors in parentheses
** p<0.05, ** p<0.01, *** p<0.001*
Base group - men, more than 55 years old, living in rural area, other religion

Table A.6: Highest attained level of education as education measure

	Dependent variables			
	HIV status	HIV knowledge	Use of condom	Number of sex partners
Primary education	0.00985**	0.105***	0.0641***	0.421**
	(0.00356)	(0.0193)	(0.00995)	(0.147)
Secondary education	0.0121***	0.233***	0.112***	1.117**
	(0.00343)	(0.0154)	(0.0112)	(0.307)
Higher education	-0.00550	0.350***	0.138***	1.721*
	(0.00445)	(0.0164)	(0.0150)	(0.738)
Women	0.0227***	-0.0284	-0.0957***	-4.016***
	(0.00211)	(0.0161)	(0.00572)	(0.652)
Urban	0.0346***	0.0634***	0.0440***	0.189
	(0.00405)	(0.00768)	(0.00733)	(0.118)
Wealth	-0.0301***	0.128***	0.0379***	-0.160
	(0.00874)	(0.0147)	(0.0111)	(0.323)
Stable partner			-0.0114	
			(0.0137)	
Random partner			0.115***	
			(0.0227)	
Sexual worker			0.186***	
			(0.0466)	
HIV status			0.0966***	
			(0.0105)	
Intercept				10.55***
				(1.124)
Age dummies χ^2 - test	509.82	158.96	722.91	7.56
Prob > χ^2	0.0000	0.0000	0.0000	0.0007
Religion dummies χ^2 - test	12.69	34.13	11.35	16.38
Prob > χ^2	0.0054	0.0000	0.0105	0.0000
Region dummies χ^2 - test	6.5e+09	2.2e+09	6.2e+11	28,557
Prob > χ^2	0.0000	0.0000	0.0000	0.0000
N	250,057	252,108	180,261	203,447
R squared	0.2084	0.1074	0.2389	0.1526

Robust and clustered standard errors in parentheses
** p<0.05, ** p<0.01, *** p<0.001*
Base group - men, more than 55 years old, living in rural area, other religion

Table A.7: Literacy as education measure

	Dependent variables			
	HIV status	HIV knowledge	Use of condom	Number of sex partners
Literacy	0.00968***	0.177***	0.0712***	0.572**
	(0.00159)	(0.0117)	(0.0113)	(0.198)
Women	0.0236***	-0.0304*	-0.0986***	-4.089***
	(0.00219)	(0.0140)	(0.00583)	(0.669)
Urban	0.0343***	0.0777***	0.0500***	0.289*
	(0.00421)	(0.00802)	(0.00722)	(0.112)
Wealth	-0.0360***	0.181***	0.0624***	0.234
	(0.00962)	(0.0145)	(0.0134)	(0.317)
Stable partner			-0.0118	
			(0.0138)	
Random partner			0.114***	
			(0.0230)	
Sexual worker			0.185***	
			(0.0445)	
HIV status			0.0954***	
			(0.0107)	
Intercept				10.59***
				(1.113)
Age dummies χ^2 - test	465.95	210.71	786.50	7.75
Prob > χ^2	0.0000	0.0000	0.0000	0.0006
Religion dummies χ^2 - test	13.67	59.02	12.76	14.91
Prob > χ^2	0.0037	0.0000	0.0058	0.0000
Region dummies χ^2 - test	1.6e+10	3.4e+09	3.1e+11	1.5e+05
Prob > χ^2	0.0000	0.0000	0.0000	0.0000
N	249,949	249,949	178,876	201,823
R squared	0.2068	0.0954	0.2338	0.1513

Robust and clustered standard errors in parentheses
** p<0.05, ** p<0.01, *** p<0.001*
Base group - men, more than 55 years old, living in rural area, other religion

Table A.8: HIV status and education - men and women subsamples

	HIV status		
	Men	Women	Whole dataset
Schooling years	0.00305***	0.00618***	0.00327***
	(0.000699)	(0.000963)	(0.000881)
Schooling years squared	-0.000211***	-0.000410***	-0.000226***
	(0.0000469)	(0.0000763)	(0.0000502)
Age 15-25	-0.0636***	-0.0627***	-0.0649***
	(0.00718)	(0.00473)	(0.00565)
Age 26-35	-0.00679	0.00279	-0.00463
	(0.00397)	(0.00442)	(0.00440)
Age 36-45	0.0164***	0.0145***	0.0121**
	(0.00368)	(0.00265)	(0.00416)
Age 46-55	0.0110**	0	0.00501
	(0.00355)	(.)	(0.00432)
Urban	0.0236***	0.0437***	0.0344***
	(0.00310)	(0.00495)	(0.00387)
Wealth	-0.0138	-0.0430***	-0.0298***
	(0.00872)	(0.00836)	(0.00825)
No religion	0.0156**	0.0122	0.0151**
	(0.00474)	(0.00860)	(0.00555)
Christian	0.00968	0.00487	0.00696
	(0.00696)	(0.00908)	(0.00682)
Muslim	0.00299	-0.00388	-0.00103
	(0.00798)	(0.00887)	(0.00728)
Women			0.0157***
			(0.00357)
Women*education			0.00225**
			(0.000871)
Women*education squared			-0.000137*
			(0.0000581)
Number of observations	112,567	137,919	252,057
R squared 0.2137	0.2059	0.2085	

Robust and clustered standard errors in parentheses
** p<0.05, ** p<0.01, *** p<0.001*
Base group - men, more than 55 years old, living in rural area, other religion

Table A.9: HIV knowledge and education - men and women subsamples

	HIV knowledge		
	Men	Women	Whole dataset
Schooling years	0.0266***	0.0249***	0.0232***
	(0.00142)	(0.00123)	(0.00172)
Age 15-25	-0.0324	-0.0594***	-0.0325*
	(0.0166)	(0.0112)	(0.0158)
Age 26-35	0.0323*	-0.00955	0.0254
	(0.0132)	(0.00697)	(0.0134)
Age 36-45	0.0377**	-0.00874	0.0307*
	(0.0130)	(0.00674)	(0.0124)
Age 46-55	0.0134	0	0.0216
	(0.0127)	(.)	(0.0128)
Urban	0.0398***	0.0692***	0.0546***
	(0.0107)	(0.00886)	(0.00708)
Wealth	0.115***	0.0993***	0.101***
	(0.0160)	(0.0171)	(0.0134)
No religion	0.0118	0.0137	0.00950
	(0.0121)	(0.0250)	(0.0132)
Christian	0.0421**	0.0795***	0.0599***
	(0.0153)	(0.0152)	(0.0125)
Muslim	0.0214	0.0618***	0.0409**
	(0.0184)	(0.0185)	(0.0153)
Women			-0.0591*
			(0.0279)
Women*education			0.00570*
			(0.00245)
Number of observations	114,168	137,940	252,108
R squared 0.0895	0.1299	0.1079	

Robust and clustered standard errors in parentheses
** p<0.05, ** p<0.01, *** p<0.001*
Base group - men, more than 55 years old, living in rural area, other religion

Table A.10: Condom use and education - men and women subsamples

	Condom use		
	Men	Women	Whole dataset
Schooling years	0.0100***	0.00884***	0.00779***
	(0.00123)	(0.000697)	(0.000992)
Age 15-25	0.336***	0.0950***	0.258***
	(0.0136)	(0.0148)	(0.0114)
Age 26-35	0.196***	0.0366***	0.164***
	(0.0151)	(0.0110)	(0.0111)
Age 36-45	0.114***	0.0182**	0.114***
	(0.0144)	(0.00694)	(0.0116)
Age 46-55	0.0661***	0	0.0685***
	(0.0174)	(.)	(0.0144)
Urban	0.0515***	0.0350***	0.0430***
	(0.0104)	(0.00569)	(0.00765)
Wealth	0.0473***	0.00636	0.0263*
	(0.0121)	(0.0114)	(0.0109)
No religion	0.0314*	0.0388***	0.0332***
	(0.0145)	(0.00841)	(0.00888)
Christian	0.0269	0.0408***	0.0333**
	(0.0170)	(0.00912)	(0.0108)
Muslim	0.0360	0.00255	0.0182
	(0.0210)	(0.0119)	(0.0113)
Stable partner	0.000878	-0.0552***	-0.0133
	(0.0119)	(0.0156)	(0.0131)
Random partner	0.143***	0.0821***	0.114***
	(0.0248)	(0.0213)	(0.0220)
Sexual worker	0.225***	0.267**	0.183***
	(0.0429)	(0.0868)	(0.0473)
HIV status	0.0774***	0.1251***	0.0974***
	(0.0097)	(0.0162)	(0.0102)
Women			-0.123***
			(0.0112)
Women*education			0.00363***
Number of observations	82,034	98,008	180,223
R squared	0.2457	0.2327	0.2482

Robust and clustered standard errors in parentheses
** p<0.05, ** p<0.01, *** p<0.001*
Base group - men, more than 55 years old, living in rural area, other religion, HIV negative, other partner

Table A.11: Number of sexual partners and education - men and women subsamples

	HIV knowledge		
	Men	Women	Whole dataset
Schooling years	0.213**	0.0276**	0.298**
	(0.0637)	(0.00935)	(0.0782)
Age 15-25	-6.462***	-0.683***	-4.949***
	(1.175)	(0.121)	(0.993)
Age 26-35	-3.189***	-0.185**	-3.388***
	(0.626)	(0.0546)	(0.726)
Age 36-45	-1.756***	0	-2.675***
	(0.422)	(.)	(0.630)
Age 46-55	-0.331	-0.123	-1.575***
	(0.207)	(0.0861)	(0.396)
Urban	0.278	0.197*	0.168
	(0.247)	(0.0848)	(0.122)
Wealth	0.0599	-0.384*	-0.337
	(0.563)	(0.158)	(0.344)
No religion	0.149	-0.0480	0.304
	(0.581)	(0.174)	(0.364)
Christian	-1.409**	-0.0565	-0.758*
	(0.471)	(0.149)	(0.301)
Muslim	-2.121**	-0.338	-1.069*
	(0.714)	(0.204)	(0.377)
Women			-2.160**
			(0.590)
Women*education			-0.308**
			(0.0873)
Constant	11.13***	2.887***	9.611***
	(0.723)	-0.193	-0.94
Number of observations	88,157	115,290	203,447
R squared	0.1577	0.0602	0.1581

Robust and clustered standard errors in parentheses
** p<0.05, ** p<0.01, *** p<0.001*
Base group - men, more than 55 years old, living in rural area, other religion