UiT The Arctic University of Norway

School of Business and Economics

The Relationship Between Education and Health in Russia

Does Younger School Enrollment Improve Health Outcomes in Adulthood?

Marius Johan Jensen

Master's Thesis in Economics, SOK-3091, May 2020

ACKNOWLEDGEMENTS

I must express my gratitude to my supervisor Associate Professor Mikko Moilanen. His feedback and guidance have been invaluable to me, and I appreciate all our interesting conversations. Moilanen allowed this thesis to be my own work, but candidly shared his suggestions along the way. His enjoyment of research has made it appealing to write a thesis.

I acknowledge PhD Candidate Emre Sari for ideas and discussions in the beginning of this project. His experience and creative mind made reviewing literature and developing research questions exciting, witty, and lively.

I thank the Russian Longitudinal Monitoring Survey, RLMS-HSE, conducted by the National Research University Higher School of Economics and ZAO "Demoscope" together with Carolina Population Center, University of North Carolina at Chapel Hill and the Federal Center of Theoretical and Applied Sociology of the Russian Academy of Sciences for making these data available (*http://www.hse.ru/rlms*, *http://www.cpc.unc.edu/projects/rlms*).

I would also like to thank Academic Librarian Erlend Hagan of the University Library for assisting me in the search for literature about the Soviet school system and its reforms.

Last, thank you to friends and family for working sessions, feedback, and encouragement.

Abstract

This thesis investigates the causal relationship between schooling and health outcomes in adulthood. It attempts to clarify if younger enrollment in the Russian primary school improved health later in life. Research on causal inferences in social processes and public services is important because it can give policy makers information on how to allocate resources more efficiently and more equitably. The fact that such research only indicate local treatment effects makes it relevant to add evidence from as many settings as possible. My motivation for conducting research in health and education economics is the distinctive feature of causality. Instead of describing how the worlds looks, causality explains how the world works. Only by fully understanding interactions are we able to create a society that ensures public welfare and individual well-being. To answer the research question, I used a school reform as a natural experiment to resemble randomized treatment conditions. The data was collected in the "Russian Longitudinal Monitoring Survey", and the estimated results were derived in a sharp regression discontinuity design. The analysis shows that lowering school entry from age 7 to age 6 leads to better self-reported health and healthier body mass index in adulthood. However, the policy change is also shown to increase the probability of acquiring chronic health conditions, which is a deterioration in health. These causal inferences are almost the same as the correlations estimated in a regular OLS regression. The implication of this research is that formal education at age 6 does improve health later in life. Even if these effects are only local, we might suspect individuals in similar settings to acquire such results.

Keywords: Health, Education, Human capital, Causality, Russia, School reform

Table of Content

List of Tables and Figures

1 Introduction

It is not easy for policy makers to create fair and efficient health services. Treating medical issues requires resources not everyone has access to. There is no secret that most places in this world is characterized by significant inequalities in health services. If we only knew how to allocate public resources better to prevent health inequalities and improve people's quality of life. The field of *health and education economics* addresses how assets like education and health status increase productivity and well-being. Countless theories and empirical studies seek to enlighten the complementarities between education and health. Is it possible that investments in childhood education cause improved health later in life? Can public spending on schooling purposely make us healthier and avoid the large costs of medical care? These are central questions of the research field, and of this thesis.

The research topic of this thesis is how school entry age in Russia impacts health outcomes in adulthood. Health inequalities and mortality rates are large problems in Russia compared to more developed Western countries. This thesis serves as a contribution to research literature on the relationship between education and health. By providing new evidence to the field, we can increase our ability to create more efficient and profitable policies. We have long believed that education leads to better health. However, we must be cautious before we invest in education as a mean to improve public and individual health. Empirical evidence confirming our predictions and hypotheses is vital to justify such policy changes. Evidence on *cause and effect* is not easy to provide as it often requires challenging research designs to show us something past the well-known correlation that we are already familiar with.

It was established decades ago that education and health are correlated and yet, it is a highly researched field today. Social sciences, medical sciences and economics are among multiple academic disciplines interested in the correlation between education and health. When two variables are correlated, it means there has been measured a dependence between their data values. They are statistically related to each other through their data values. The problem is that correlation does not necessarily mean that change in one variable *causes* change in the other. Three possibilities can explain the correlation.

First, the cause of the correlation may be *unidirectional* from education to health, or vice versa. Second, it may be *bidirectional* between them, both causing an effect in the other. Third, both may depend on other factors outside their correlation.

Defining the cause and effect of such a dependence is what we call *causality analysis*. Although the concept of causality has strong roots in the works of philosophers like Immanuel Kant and David Hume, only recent decades have social sciences sharpened their empirical attention on causality rather than correlation. Simply pointing out correlations has its limitation on moving science forward. To *describe* how the world *is* does contribute to science, but our goal should be to *explain* how the world *works*.

Science is not only about the satisfaction of solving complex problems, but about improving how we live. In other words, science is not only interesting, it is important. Basing our policies on simple correlations may waste more resources than they return. Awareness of the direction and size of causal interactions between education and health suggest how we can structure society *efficiently* and *equitable*.

The issue of distinguishing between correlation and causation, and detecting true causal effects, is a matter of data analysis and empirical methods. Experimental data is great when seeking to estimate cause and effect. However, most economic research address observational data. The solution to detect causal inference in observational data is to *resemble* experiments in the estimation strategy. Natural experiments occur when situations outside the researcher's control seem to randomly assign treatment conditions on a variable. Causality studies on education and health often apply school reforms and other policy changes as natural experiments to estimate causal inferences.

To fully trust what we have found to be causal interactions, large amounts of research evidence is needed. The effects estimated in an analysis do not guarantee any effect outside the given setting. The research field is therefore driven by continuously adding knowledge on nuances previously unaddressed, such as data from different places, measuring different mechanisms of education and different outcomes of health. The purpose of this thesis is precisely that, to contribute with information on an unaddressed aspect of the relationship between education and health.

This thesis attempts to estimate causal effects of younger entry in the Russian primary school on several health outcomes in adulthood. The outcomes are *self-reported health*, *body mass index* and *chronic health conditions*. To resemble randomization, I consider a USSR school reform from 1984 that lowered the school entry age from age 7 to age 6. The data set I use is from the "Russian Longitudinal Monitoring Survey" *(RLMS-HSE, 2020)*.

The research field is relatively young and entered economics in recent decades, but quickly became populated with researchers and studies. Hundreds of articles have been published on the relationship between education and health. Many of them seek to detect causal inferences. In the beginning, mortality rates was a popular case to represent health outcome, but more recent evidence aims towards health measures and health behaviors such as smoking, alcohol consumption, physical activity, and mental health, along with those presented in the previous paragraph. Most studies so far analyzed data from the U.S. and Western Europe using school reforms as natural experiments. Therefore, the field still has many small and large gaps to fill. New contributions might investigate unstudied areas, or unstudied pathways of influence from education to health.

This thesis will assist in filling three gaps. First, it estimates the effect of lowering school entry age, which is mostly untouched in existing literature. Second, it uses data from Russia which has not been done before in this setting. Russia is a transitioning economy with higher inequalities than Western countries. Evidence on Russia might suggest how other similar countries react to this kind of policy change. Third, previous research warrants more evidence on the health outcomes used here. Results in the current research field is mixed and mostly inconclusive, though suggestive, on the existence of causal mechanisms from education to health. Based on theoretical predictions and empirical evidence, I hypothesize that younger entry in primary school improves health outcomes in adulthood because children get more time to accumulate human capital and develop health production skills.

Chapter 2 explains how the research topic is rooted in economic theory and puts forth theoretical predictions to be tested in empirical analysis. Chapter 3 presents and discusses existing evidence on the relationship between education and health. Chapter 4 addresses the USSR school system and the educational reform approved in 1984. The empirical methods and estimation strategies are explained in chapter 5. Information about the data and its variables is given in chapter 6, along with some descriptive statistics. The empirical results are presented in chapter 7. Last, chapter 8 sums up the thesis and discusses important implications from the results, before it considers ideas for further research.

2 Theory

Education and health are important in economic development and public finances. Their relationship raises many interesting topics, such as

1. What determines an individual's educational attainment, and what determines the individual's health?
2. How does policies and public spending affect these outcomes?
3. Can we use this knowledge to explain economic growth and improve the individual's utility?

Before presenting my empirical analysis, I here introduce existing economic ideas my work is rooted in. The purpose of this chapter is to place education and health within what we call human capital, and then derive a prediction on how they might affect each other.

2.1 What is Human Capital?

Paraphrasing Oxford Learner's Dictionaries (2020), human capital is

> "the **skills, knowledge and experience** of **a person, group of people or labor force**, regarded as a resource or an asset. That is, these resources are reckoned as something **valuable** that an **organization, company, country or economy** can make us of."

Whereas physical capital comprises assets such as machinery, buildings, land and stock shares, human capital includes assets like education, job-training, and health. Despite human capital entering our terminology in the 20th century, Adam Smith discussed the same idea (Goldin, 2016). He said that the costs with acquiring skills are what constitutes the capital *in* a person. However, even if acquiring education and health are costly, the expenses would be repaid with a profit. In addition, skills attached to an individual is a fortune to both himself and society. It improves the individual's well-being **and** the welfare of the economy. When we facilitate for someone to become more educated and healthier, we make investments in this person and assume the investment to increase his productivity (Goldin, 2016).

Becker (1993, p. 16) distinguishes human and physical capital in their attachment to the employer and the employee:

> "we cannot separate a person from his or her knowledge, skills, health, or values the way it is possible to move financial or physical capital assets while the owner stays put."

For example, assume a company owner wants to move production to another city or country. Machinery and equipment can be moved while the owner and his administration stay put. Buildings and land can be sold, and a new factory can be bought or built in the new location. The problem is to maintain the skills and experience of the production site labor force. The employees' identities and families prevent many from migrating. One solution is to hire workers at the new site, but they are most likely not as productive and profitable as the original labor force. Even if their education and health status are the same, they lack knowledge on the company's production and organizational culture. It requires time and money to increase productivity to its previous level through practice and training.

Human capital is also important to economic growth and has been shown to reduce the growth residual (Barro & Sala-i-Martin, 2004; Jones & Vollrath, 2013; Mankiw, Romer, & Weil, 1992). Whereas Mankiw, Romer and Weil show that rich countries have high investment rates in human capital, Jones and Vollrath demonstrate that rich countries spend more time acquiring skills. Despite different assumptions, both models tell us that investing in education and health contributes to production, wealth, and welfare.

2.2 Education and Health

Gary S. Becker is one of the most cited economists on health and education, known for several theoretical and empirical works on the relationship between these two human capital assets. He refers to the relationship as economic *complementarities*. His theoretical models promote predictions for what we should expect to find in empirical analysis. In the following model, Becker (2007, pp. 389-390) predicts that "an increase in survivorship at later ages raises the returns from investments in education because educational costs come at earlier ages and returns at later ages." I wish to present this model and thereafter see if its predictions hold. If empirical analysis shows that schooling and health has complementarities, then Becker's predictions can be confirmed.

Becker's (2007) model is a two-period example, at time t_0 and t_1. Assume the cost of education (E) occurs at t_0 and that the individual survives this period. If the return of education at t_1 is represented by a higher wage rate, we know that $\partial w_1(E)/\partial E > 0$. This prediction is well known and heavily supported by evidence. The argument that education increases wage, which increases spending on health services, is often referred to as the indirect, or monetary, effects. Becker also states that an increase in education raises survival

rates directly, or non-monetary. Indeed, evidence suggest that more educated individuals manage their health better, even with medical expenses fixed at a given level. Such individuals are thought to have healthier habits and lifestyle in many ways. For example, they visit better doctors, consume healthier diets, and take their medications as prescribed. That means, if S_1 is the probability of surviving to t_1 and h is the expenditure on health, then $\partial S_1(h,E)/\partial E > 0$. In other words, schooling raises life expectancy. This statement is the foundation of Becker's model. However, it should not be taken for granted, but rather be treated like a theoretical prediction which we can test empirically and then offer additional support.

In Becker's model, the individual's utility function is

$$V = u_0 + BS_1(h,E)u_1 \tag{2.1}$$

B is the discount rate and u_i is the utility at the respective age, or time period. Note that u_i is the utility that depends on goods and leisure, $u_i(x_i, l_i)$. The budget constraint with annuity for both periods is

$$x_0 + \frac{S_1 x_1}{1+r} + E + g(h) = w_0(1-l_0) + \frac{S_1 w_1(E)(1-l_1)}{1+r} \tag{2.2}$$

$g(h)$ is the convex function of health expenditure and r is the interest rate the individual faces. The left-hand side (LHS) represents the consumption of goods and the investments in education and health that the individual can afford given his income, which is shown on the right-hand side (RHS). In this two-period model, Becker assumes education and health costs occur at t_0. Maximizing the utility function with respect to x_i, l_i and h, subject to the budget constraint, gives the first order conditions (FOC) for goods and leisure

$$u_{0x} = B(1+r)u_{1x} \qquad \frac{u_{0l}}{u_{0x}} = w_0 \qquad \frac{u_{1l}}{u_{1x}} = w_1 \tag{2.3}$$

and for health expenditure

$$\frac{dlogS_1}{dh} BS_1 u_1 = u_{0x}\{g'(h) + \frac{1}{1+r}\frac{dS_1}{dh}(x_1 - w_1(1-l_1))\} \tag{2.4}$$

In the FOC with respect to h, the LHS represents the marginal benefit of increased spending on health. It depends on the effect of health expenditure on survivorship, the discount rate and the survivorship and utility in the future. The RHS depends on, among other things, the opportunity cost of health expenditure, represented by u_{0x}. Continuing, Becker derives the optimal investment in education given by

$$\frac{1}{1+r}S_1 w'_1(E)(1-l_1) + \frac{1}{1+r}B\frac{\partial S_1}{\partial E}\frac{u_1}{u_{1x}} = 1 + \frac{1}{1+r}\frac{\partial S_1}{\partial E}[x_1 - w_1(1-l_1)] \qquad (2.5)$$

The first term on the LHS shows discounted earnings caused by a higher wage rate from an increase in education. The second term is the increased utility caused by higher survivor rate from an increase in education. Becker refers to these terms as the *market* effect and the *psychic* effect. You may regard them as synonyms to the indirect and direct effects. The former is the higher earnings, while the latter is the increased value of a life with a higher probability of surviving. Together, they give the total benefit from increased education expenditure. The RHS gives the cost.

The optimal education investment equation has two important implications for the complementarities between schooling and health. First, increased spending in education raises wealth due to a higher wage rate. With increased wealth, the individual can increase its life expectancy by increased spending on health. Second, education increases life expectancy directly through making the individual more productive in health investments, and through inducing a healthier lifestyle.

Becker predicts that investing in education raises survivorship. In other words, he predicts that there are complementarities in the relationship between education and health. This thesis sets out to empirically test Becker's prediction that schooling causes change in health-related outcomes and health-behavior.

3 Literature

Few ideas are as established as the correlation between education and health. Abundant evidence supports a statistically significant relationship between schooling and health outcomes. However, results from *ordinary least squares* regressions (OLS), only pose a *dependence* between two variables and do not explain the *cause* of such complementarities. Increasing schooling may lead to better health. Poor health may lead to lower educational attainment. Or, seeing that both schooling and health are endogenous, a third possibility arise. Unobserved, omitted variables may cause changes in both schooling and health, thus explaining the strong correlation. Such third factors may be social background, genetics, and time preferences. Evidence on all three possibilities are important. Cutler and Lleras-Muney (2006, p. 10) put it wisely in that policy makers must "understand how much of the observed correlation between education and health can be explained by each of these explanations". For example, public spending on schooling will only improve health if education causes health.

In the 2000s, a desire to accurately define causal interactions emerged, and the research field exploded with studies addressing causal effects between education and health. Research have applied numerous identifications, models, and outcomes. Due to the narrow scope of this thesis and the large size of the research field, the following literature review only addresses some of the most relevant articles. The selected articles here have three things in common with my research topic.

1. They attempt to measure causal effects from education to health. Effects of health on education are excluded.
2. One or several *school reforms* are used as natural experiments resemble random assignment to control and treatment groups. The treatment condition affects *the number of years in school*.
3. Only effects of *own* education on health outcomes in *adulthood* are considered. Impacts from parent to child or vice versa are excluded.

For broader information of the research field, I recommend the literature reviews by Cutler and Lleras-Muney (2006), Eide and Showalter (2011) and Grossman (2015).

One of the earliest and most cited evidence in the field is from Lleras-Muney (2005). She tests effects of educational attainment on mortality rates in the U.S. Using compulsory education laws from the first half of the 20^{th} century, she suggests that education has a causal impact on

mortality. First, she uses a *regression discontinuity* (RD) *design* to estimate these effects by comparing mortality rates of cohorts right before and right after a change in legislation. To make the control and treatment groups as similar as possible, she included only 7 cohorts. 3 cohorts before, 3 cohorts after and the cohort of the change. Her results show that mortality rates drop for the cohort of the change and remain low for the following three cohorts. However, the analysis consisted of several samples from multiple states, which made each sample small. Therefore, estimates were not significant, but "they do provide suggestive evidence that compulsory laws lowered mortality" (Lleras-Muney, 2005, p. 198).

In addition to an RD approach, she conducts an analysis with *instrumental variables* (IV), also known as *two-stage least squares* (2SLS). The IV estimations show that an additional year of education lowers the probability of dying in the next 10 years by approximately 3.6%. Also, one more year of compulsory schooling decreased mortality after age 35 by 3%. She reminds us that these IV estimates and the OLS were not statistically different. That means she finds no evidence that education is endogenous in the mortality equation, in contrast to what we expect.

Despite these seemingly great results, Lleras-Muney requests us to be wary of making policies based on them. First, we do not know about the specific mechanisms by which education affects health. The pathways of influence need more attention from research and should be identified before we make educational investments meant to improve health. Second, the compulsory schooling laws addressed here were implemented in the first half of the 20^{th} century and concern low initial levels of education. Effects of policies today might not give the same effect for every additional year. However, some present time developing countries have average levels of education like those in the U.S. early 20^{th} century. Her results therefore imply that such countries can increase life expectancy in adulthood by more aggressive education policies.

Lleras-Muney suggest one of the possible direct mechanisms of a causal link may be that schooling gives individuals critical thinking skills, which is useful in health production. That is consistent with hypotheses raised by Becker and Grossman. More educated people comply with their treatments and manage chronic conditions better, especially when treatments are complex and requires learning by doing. In that, she warrants research on how different stages of schooling develop unique, productive thinking skills.

At this point in time, most research evidence used American data. Arendt (2005) adds one of the early European evidence to the field. He uses a two-period panel of almost 3,500 Danish workers, observed in 1990 and 1995. Two Danish school reforms from 1958 and 1975 are used as instrumental variables to estimate effects of education. Before 1958, pupils had to pass a test to continue from 7^{th} form to 8^{th} through 10^{th} form, which was necessary to attend higher education. Few schools in rural areas offered 8^{th} through 10^{th} form, which made educational attainment between people in urban and rural areas unequal. From 1958, every pupil could attend 8^{th} through 10^{th} form and more secondary schools were built in rural areas. However, only 7 years were compulsory. The reform of 1975 increased the minimum school leaving age, leading to an increase in compulsory schooling years from 7 to 9 years. Most children already passed 9^{th} form, so this reform should have limited effects on mean education. Nonetheless, this reform removed the distinction between two tracks of secondary school, making 8^{th} through 10^{th} form equal for all pupils. It is interesting and clever that Arendt studies two different school reforms, because different types of reforms cannot be expected to yield equal effects.

Arendt analyzes newer data and slightly younger adults than Lleras-Muney, thus making mortality either unreasonable or unavailable to study. He uses *self-reported health* (SRH) as the main health outcome and *body mass index* (BMI) as one of the supplemental outcomes. These types of health outcomes served as important contributions to the field. According to Arendt, SRH as a health measure has both advantages and disadvantages. SRH is thought to be a useful summary measure of general health. The reason is that it may capture health aspects which is difficult to obtain in more objective and clinical measures. On the other side, SRH might give errors if people "justify" their deprived situations by reporting their health to be poorer than it is. Therefore, recessions and other damaging changes in people's lives may cause error in using SRH. Arendt tries to control for such bias using a time dummy in all his estimations.

From the first-stage regression, he shows that the 1975 reform does not have statistically significant effects on educational attainment, just as he suspected. In contrast, the 1958 reform lead to significantly higher educational attainment, showing a jump both for men and women. Ordered logit models show that for both genders, longer education is associated with better SRH. However, because of IV standard errors, he cannot reject that education is exogenous to SRH, and he cannot reject that there is no effect of education. Similar results were obtained for the supplementary outcome BMI. Hausman tests for weak instruments did not detect

problems with his estimation strategy. Therefore, his results remain inconclusive about the effect of education on health.

Suspecting that mechanisms between education and health vary across countries, researchers continued to add evidence from countries that were previously not studied. Albouy and Lequien (2009) provide the first evidence on causality between education and health using French data. With panel data on 1 % of France's population, they seek to test if individuals provided with a higher level of education obtains better health. To identify causal links, two school reforms are used. One raised the minimum school leaving age from 13 to 14 years. Three decades later, the next reform raised it to 16 years. Health outcome is measured in mortality. The earliest reform allowed them to check mortality at age 80, and the second mortality at age 50. Albouy and Lequien did not consider age 50 as too young to reveal significant returns to education on survival. They referred to Lleras-Muney (2005) who suggested a causal effect on mortality at similar ages.

Both reforms were shown to increase mean school leaving age of the cohorts. Only cohorts at a maximum of three years before or after interventions were included. The samples represent some of the largest in the field at approximately 36,000 and 47,000 persons. Despite jumps in educational attainment and declines in mortality, neither an RD design nor a 2SLS regression resulted in a significant causal effect of more years in school on mortality. However, insignificant estimations do not exclude further implications.

The authors emphasize that the results only indicated that mortality at age 50 and 80 were not affected by schooling at age 13-16. Schooling in this age interval may affect mortality at other ages, and it may affect health in other dimensions than survival. Also, it is possible that schooling at ages below 13 and above 16 causes lower mortality rates. Hypothetically, each year in school should add knowledge and skills that improves health production differently from other years. Albouy and Lequien also states that the student's motivation may influence his acquisition of skills and therefore the size of these effects. In conclusion, they cannot provide evidence on a causal link and note that perhaps school leaving age is not the educational mechanism that improves health production the most.

Since then, the 2010s has frequently seen new evidence published. Research papers appear similar, but important differences characterize their contribution to the field. There is still a continuous inclination to add new evidence from different geographical areas, with larger

sample sizes, using new strategies of identification and estimations, with new or relatively unstudied explanatory and outcome variables.

Among studies done on UK and Great Britain data, many fall inside the three criteria presented in the introduction to this chapter. Braakmann (2011) finds neither an effect of education on various health-related measures nor on health-related behaviors such as smoking, drinking and nutrition. His natural experiment is that British compulsory schooling laws allowed January-born students to leave school earlier than February-born students in the cohorts from 1957 to 1970. Being February-born is therefore an instrument for education. His results show no causal effect, at least for this intervention. However, Braakmann emphasizes that his results do not rule out causal links between other forms of education and health.

Clark and Royer (2013) exploit British school reforms in an RD design. The school reform of 1947 increased the minimum school leaving age from 14 to 15 and the reform of 1972 increased it one more year up to age 16. Despite the reforms leading to dramatically different educational attainment for individuals born just days apart, they find that the reforms did not affect mortality or other health outcomes with statistical significance. They suspect that such interventions have small causal effects on health and that the economic models we base our hypotheses upon may need to be rethought. There may be other factors of education with significantly large effects on health. Nonetheless, both reforms increased educational attainment and wages, which suggests indirect effects on health. Clark and Royer recommend caution in basing health policies on educational investments.

Jürges, Kruk, and Reinhold (2013) also exploit these two school reforms, but measures health as SRH and biological stress markers. Like with previous research, causal effects of compulsory schooling on health remain ambiguous and statistically insignificant.

Silles (2015) examines a possible causal link between schooling and smoking. She compares the effect of additional schooling in cohorts who were teenagers before and after the health consequences of smoking were widely known. In accordance with previous work of Clark and Royer, her results conclude that there was no causal effect in the Great Britain data. However, for Northern Irish men, schooling had a causal effect on health. Before the dissemination of health implications of smoking, individuals with more schooling achieved better health, most likely due to being better informed about the consequences of smoking. After these consequences became widely known to everyone, schooling differences did not remain.

Not all education policies address the number of compulsory years in school. Ma, Nolan, and Smith (2018) consider a UK policy change made in the 1960s that eliminated public secondary school fees and test its effects on clinically measured health outcomes. They find that an additional year of schooling decreases the probability of having hypertension and diabetes by approximately 3% and 1% respectively. Their analysis also show that additional schooling increases physical activity and decreases smoking. These results imply that number of years in school is not the only causal mechanism. Making school accessible to individuals who otherwise cannot enroll has positive effects on health outcomes.

Even after tens of UK studies on the relationship between education and health, new evidence is still presented today. Most new evidence still use compulsory schooling laws as identification, but they contribute with new ideas by estimating educational effects on less studied health outcomes. Janke, Johnston, Propper, and Shields (2020) study the causal impact of education on *chronic health conditions* (CC). The survey sample they use is the largest in the UK. Causal effects are estimated from two different school reforms. The first raised the minimum school leaving age. The second combined several policy changes that broadly affected educational attainment distribution. Tests across both reforms do not show statistically significant causal impact of additional schooling on most chronic health conditions. The only exception, both reforms reduces the probability of having diabetes.

Despite their results showing considerably smaller effects than the associated OLS regressions, the authors are unable to statistically rule out educational effects for many of the health conditions. The causal estimates are too unprecise, and results remain inconclusive. They recommend to further examine how education causally affects diabetes and other highly lifestyle-related conditions. The results from this paper is to some degree consistent with theoretical predictions that say education has direct effects on health through lifestyle, habits and thinking.

An even more recent UK paper from Avendano, de Coulon, and Nafilyan (2020) examines whether increased minimum school leaving age has a causal effect on mental health. From existing literature, it is obvious that mental health lacks attention. Using the 1972 reform in an RD design, they find that the reform did not improve mental health. The authors believe increased leaving age may worsen mental health directly and indirectly. Whilst compulsory schooling laws can raise educational attainment, improve labor market outcomes and specific aspects of health, it may have negative effects for some individuals. Coercion may not always

be welfare-enhancing if the returns to increased education is overrun by the psychological and emotional costs of extra schooling. Avendano, Coulon and Nafilyan ask to bear this in mind when designing compulsory schooling laws. Coercive schooling might not have optimal effects on health outcomes. Research on other mechanisms is needed.

Another country which is exposed to thorough analysis over the years is Germany. Jürges, Reinhold, and Salm (2011) use variation in upper secondary education across states and time to estimate causal effects on health behavior. During the postwar-period, German states introduced educational changes at different times, creating exogenous variation. Their results show robust negative effects on women's smoking, and large but not robust negative effects on men's smoking. They find no causal effect of education on BMI.

In another study using the same educational reforms, Kemptner, Jürges, and Reinhold (2011) find evidence for a significant causal effect of years in school on long-term illness for men. Despite the results in their other paper, the estimations done here show no causal effect of education on smoking behavior. Even if results are mixed, they conclude that there are significant non-monetary returns to education on health measures, but not necessarily on health-related behavior. Further research needs to explore possible non-monetary mechanisms.

In a recent study on the same topic, Jürges and Meyer (2020) states "the effectiveness of education policy to combat smoking" as limited. The reason is that educational differences in smoking develop before the school leaving age. Increasing compulsory schooling further and further will not achieve desired effects. Maybe other factors in already compulsory schooling can combat smoking behavior.

The German cross-state studies are not the only ones to exploit variation across geographical areas and time. Brunello, Fabbri, and Fort (2013) use data from nine European countries to estimate causal effects of education on BMI. School reforms from the 1960s and 1970s that increased the minimum school leaving age in these countries create exogenous variation. They find that additional schooling has a causal protective effect on BMI only for women and suggest the largest effect for individuals with low initial levels of schooling. That is consistent with what other researchers have stated, such as Lleras-Muney and Jürges and Meyer. The implication is that lifestyle and thinking skills related to combat overweight and obesity are acquired in earlier years of education, and not at leaving age. The countries included comprise a selection from Scandinavia and Western, Central and Southern Europe. The lack of research

on Eastern European countries remain. The authors themselves question whether their results hold more generally for the rest of Europe. However, they do believe so due to the broad group of countries analyzed here.

In the most recent cross-country analysis, Fonseca, Michaud, and Zheng (2020) aim to estimate the causal effect of education on health by combining three surveys with nationally representative samples from fourteen OECD countries. The observed individuals are aged 50 and older. They use differences in compulsory schooling laws across time and countries as IV. In addition to its unique cross-country variation, the paper contributes to the literature by estimating effects on a wide range of health outcomes, including SRH, difficulties in *activities of daily living* (ADL) and CC. Their hypothesis is that different compulsory schooling laws affect educational attainment differently across birth cohorts and countries. Eight of the fourteen countries implemented nationwide compulsory schooling laws that affected cohorts in the survey. The remaining six countries either had no legislation change or legislation changes varied geographically within the country.

The IV estimates reported that an increase in compulsory schooling leads to lower probabilities of poor SRH and lower probabilities of difficulties in ADL. Increasing the number of years also has a significant causal effect on CC such as heart diseases and diabetes. However, education has no significant effect on cancer, stroke, and psychiatric illness. The size of the significant impacts is larger than those in most existing literature. One additional compulsory year in school reduces the probability of poor SRH, difficulties in ADL and chronic illness by 6.85%, 3.8% and 4.4% respectively. Seeing that the compulsory schooling laws here were introduced in the first half of the 20^{th} century, when intital levels of education were low, the authors conclude that countries with weaker compulsory schooling laws can obtain health benefits by intervening in education. These results are to some extent consistent with Lleras-Muney (2005), but show higher and more significant impacts. Even though Fonseca, Michaud and Zheng addressed the compulsory number of years in school, they advise interventions aimed at improving both quantity and quality of education. However, to make the most efficient interventions, further research should address the quality of education.

It should be noticeably clear that much evidence identifies educational attainment in the minimum school leaving age, but some studies consider the effect of higher education on health. In an RD design, Zhong (2015) finds that a higher education expansion in China do

not cause better SRH, BMI or smoking and drinking behavior. However, college education may significantly reduce the probability of hypertension. The cause of hypertension is often considered as genetical, but research have found that it also relates to physical activity and diet. Zhong suggest that higher education may affect health-related behavior positively in these two factors. He emphasizes that the causal effects of "education on health might be heterogeneous across levels of education, age distribution, different health measures, different social settings, etc." (Zhong, 2015, p. 651). That is, treatment effects in existing evidence is only estimated locally. According to Zhong, we should accumulate local effects from many different settings.

Zhong (2016) has further contributions to the literature. He points out that existing evidence address education at a specific level. In this analysis however, Zhong estimates the average effect of education across all levels in the aftermath of the Chinese Cultural Revolution. The results show that higher average education reduces the probability of poor SRH. Both of his papers serve as important contributions to the literature as they investigate Chinese data.

The research field has been continuously active the last two decades with additional evidence being published frequently. Yet, evidence is mostly inconclusive about the effects of education on health. New contributions consider areas, data, strategies, outcomes, and mechanisms that was previously inconclusive or unaddressed. It is narrow gaps like these that this thesis will attempt to fill. This thesis contributes to the field in at least three ways.

1. It addresses data from Russia. As clear from the review, most evidence today are based on data from Northern America and Western Europe, and some from Asia. Russia does not have the same culture or economy as for example UK, Germany and Sweden. This thesis may provide new insight on transitioning economies that is slightly less developed and has higher inequalities. The fact that Russia's population has been literate long before the 1984 school reform inclines us to believe that a new reform cannot impact health that much. Nonetheless, Russia does still struggle with preventable diseases that Western countries have mostly succeeded to combat. Therefore, it is reasonable to hypothesize that a school reform in the 1980s may have improved the health of Russians.

2. It identifies education in the entry age at which primary school was started. As thoroughly discussed above, most evidence approach the top year, or minimum school leaving age. We should believe the bottom year, or school entry age, to have different effects on health outcome. The reason is that six-year-old children do not learn the same

health production skills in their grade as what 16-year-old children do. Obviously, the curriculum is different, but even with equal syllabus, children of different ages process information differently. I believe that those aged six improve future health production through education distinctly from those aged 16, and vice versa.

3. The cohorts of this analysis are younger than those observed in other studies. Not only are mortality rates unknown in the data, it would also be unreasonable to believe mortality changing so far from life expectancy. Therefore, health outcomes will be analyzed in variables for SRH, BMI and CC. Recent research finds such variables more interesting than mortality because they give more information about the individual's quality of life and health *as* he lives. It is commonly believed that these variables should be changeable through public interventions and policy changes. It makes the research more relevant and rewarding that we might improve individuals' own perception of their health and satisfaction with life. These variables are to some extent investigated, but economists have warranted further research.

The literature reviewed above also warrant evidence on more untouched aspects of the relationship between education and health. Some of the larger gaps in the field relate to highly unknown factors such as quality of education, voluntary schooling, and education in developing countries. This thesis will leave these gaps unanswered, but I will discuss them in a section on further research.

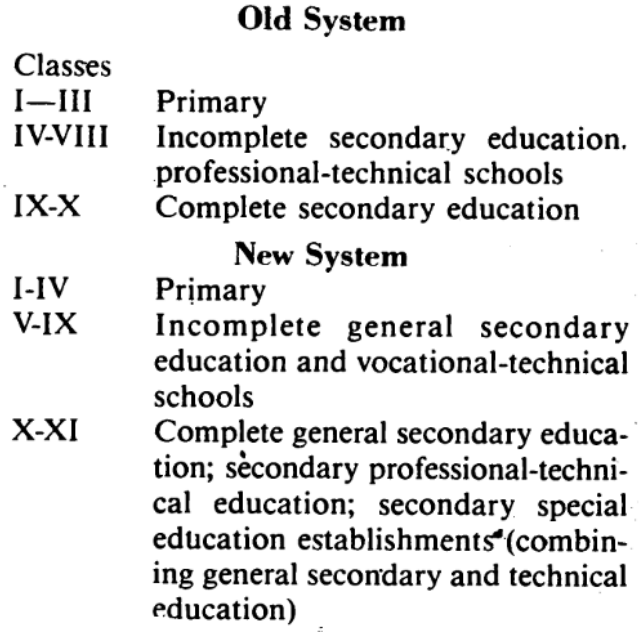
4 USSR Educational Reform 1984 – A Natural Experiment

"Identifying causal relationships is important, but not always possible, and often exceedingly difficult" (Finseraas & Kotsadam, 2013). This is especially true for non-experimental, or observational, data. Contrary to true experiments, investigators of observational data do not have the privilege of randomizing individuals to treatment conditions. Observational data is often collected for multiple uses unknown at the time of collection, and the investigator must analyze data on what is seen and heard without having control over the variables. Such data is often meant to measure effects of a wide range of policy changes, but ironically the goal is sometimes counteracted in its purpose.

Most data in social sciences is observational because it may be unreasonable to place a country-wide population under experimental design, and ethical considerations arise when giving people unequal access to treatment conditions like public welfare services. Due to these issues, researchers strive to use estimation strategies aimed to *resemble* randomization of the explanatory variable. Some of these strategies use *natural experiments*, in which individuals are assigned to treatment and control groups outside the researcher's control. It is assigned by nature. This resemblance of randomization can help us estimate causal inference, but it has both pros and cons. Whereas true experiments has fewer threats to internal validity, its external validity often fall short compared to the generalizability of quasi-experiments and natural experiments. In the RLMS-HSE (2020) data, we can use knowledge on school reforms to conduct a natural experiment. This chapter presents an important, historical context about the educational system in the USSR and how it changed with the Educational Reform 1984. The more methodological considerations of the reform are considered in chapter 6: Method.

In the mid-20th century, compulsory schooling in the USSR consisted of four years in primary school and four years in lower secondary school. Eventually, one year was moved from primary to lower secondary school, thus creating variation among students leaving primary school. Pressure was put on secondary school to close the gap between students and on primary school to ensure high quality even though it had one less year to do so. These issues would come to strengthen the argument for increasing the number of compulsory schooling years in the coming reform. After eight years of compulsory schooling, students chose to either work, enroll in technical/vocational schools, or study two years in upper secondary school to qualify for higher education in universities. The school system in the USSR reflected the Communist doctrine and aimed to educate people in a collective manner to meet the needs of the society. Standardized materials were to be memorized, thus

discouraging individual development and creativity. The positive consequence of such policies was that schools were made free and compulsory for everyone so that the Soviet regime could achieve ideological goals. Even though the completion rate of upper secondary school differed greatly between urban and rural areas, the USSR is known to have had one of the highest literacy rates in the world at the time. In 1979, 99.8% of the population aged 9-49 knew how to read and write (Mironov, 1991, p. 243). Yet, the school system underwent several major changes the succeeding decade.



Old System	
Classes	
I—III	Primary
IV-VIII	Incomplete secondary education. professional-technical schools
IX-X	Complete secondary education
New System	
I-IV	Primary
V-IX	Incomplete general secondary education and vocational-technical schools
X-XI	Complete general secondary education; secondary professional-technical education; secondary special education establishments (combining general secondary and technical education)

Figure 1 Soviet school system, Reform 1984

Reprinted from Gidadhubli (1984, p. 1737).

Conducted by the USSR leaders Chernenko and Gorbachev, the new educational reform was originally approved in 1984. Hence, its nickname “Reform 1984”. It comprised several changes of the Soviet school system. Most visually was the structural change of compulsory schooling years. Contrary to many of the world’s school reforms during the 20th century, Reform 1984 increased compulsory schooling by adding a year at the bottom rather than at the top. Since primary school only comprised three years versus five years in lower secondary school, policy makers found the most reasonable solution to the challenges in school to lower the entry age rather than increasing the leaving age. Instead of attending school from age 7-15, children were expected to begin their education at age 6. The old Brezhnev pattern 3-5-(2) was replaced by the new Gorbachev pattern 4-5-(2) (Gidadhubli, 1984).

The number of compulsory years and entry age were not the only changes of the reform. Further goals were set, such as decreasing teacher-pupil ratio, providing schools with better facilities, improving vocational and technical schools, increasing labor content in school by periodic placements in firms, making pupils more committed to Marxist-Leninist ideology, and improving education management. New curriculum and textbooks had to be made; the teaching body was going to be subjected to peer and supervisory reviews; and new schools for millions of pupils were planned to be built in the Twelfth USSR Five-Year Plan, the years 1986-90 (Gidadhubli, 1984). The goals were overwhelming, and implementation turned out to be slow. Many changes were delayed, including admission of six-year-old children. Politicians and sociologists were excited and nervous to observe effects of Reform 1984. Moreover, I am interested in investigating the socioeconomic effect. This thesis attempts to use the new, lowered school entry age to estimate causal inference of schooling on health outcomes in adulthood.

5 Method

This thesis estimates causal inference in an RD design, but to do that it is important to know the cutoff value and how treatment was enforced. We must ask ourselves two questions: "When was the new entry age implemented? Was treatment strictly enforced, or voluntary?"

In 1984, the First Deputy Prime Minister argued that admission of children at age 6 will commence from 1986 (Gidadhubli, 1984). However, the size of what was supposed to be the first cohort was underestimated. Szekely (1986) believed implementation to take longer than the Five-Year Plan unless they damaged the quality of education. In fact, implementation was done over multiple years to avoid overcrowding. According to the USSR government, 1.7 million six-year-old pupils were enrolled in 1986, but they admitted the challenges of unsatisfactory teaching facilities. The construction of new schools delayed many changes of the reform, but they assured that the goals would be met by 1992 (Sutherland, 1999). Without data or statistics to confirm the *treatment cutoff* for when the new entry age started, I accept the government's statement and assume that true and significant enrollment started in 1986.

In addition to confusion about the cutoff year, admission of children aged 6 depended on what the parents desired and on the child's development and health. Some parents held their child back until aged 7. Moreover, Szekely (1986) said the new enrollment age did not pose as a startling innovation. Prior to 1986, many children aged 6 already enrolled in class *zero*, a pre-school preparatory class not too different to what the new class *one* would be. In 1984, over one million children attended such classes (Gidadhubli, 1984). Even after years, parents still chose to enroll their children at age 7. Only 20% of enrolled pupils in 1991 were aged 6 (Eklof & Dneprov, 1993). It is fair to say that treatment conditions were most likely mixed at both sides of the cutoff. Treatment was not strictly enforced, or sharply cut, between the control and treatment groups. With better data, the *treatment probability* would be possible to derive, but it remains unknown for this analysis.

5.1 Multiple Linear Regression

A multiple linear regression is estimated so that we have regular correlations to compare with output from the RD design. Seeing that comprehensive data like this often have two or more variables related to the dependent variable, it is reasonable to include control variables. The model contains six independent variables: Five control variables Z_i and the binary treatment variable X_i. The treatment is defined by which side of the cutoff the individual is born. If the

individual is born in 1979, he is in the control group. If he is born in 1980, he turned 6 years old in 1986 and is assumed to have enrolled in school that year. In the results tables, I refer to the treatment variable as "Implementation 1986".

$$X_i = \begin{cases} 1 & BIRTH_i \geq 1980 \\ 0 & BIRTH_i < 1980 \end{cases} \tag{5.1}$$

To validate the comparison between the OLS and RD estimates, the OLS includes only the same six cohorts as in the RD design. Cohorts 1977-1979 are in the control group and cohorts 1980-1982 are in the treatment group. The model is

$$Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 Z_{4i} + \beta_5 Z_{5i} + \beta_6 X_i + e_i \tag{5.2}$$

The *local average treatment effect* (LATE) is β_6. In a level-level regression, we interpret the coefficient as a unit change. If you change X_i from zero to one, we expect the dependent variable Y_i to change by β_6. Note that this change is a correlation, not a causal effect.

5.2 Sharp Regression Discontinuity

The optimal model when the treatment probability is not equal to one is *fuzzy* RD design (Finseraas & Kotsadam, 2013, p. 16; Hill, Griffiths, & Lim, 2018, p. 350). However, since the treatment probability is unknown, the RD design conducted here will be *sharp*. Even if estimates may be unprecise in size, they should indicate effects like those the fuzzy design would estimate. A sharp RD design involves running two separate OLS regressions, one for the control group and one for the treatment group. The value of the two models will be compared at the cutoff, showing a *jump*, or *discontinuity*, in the dependent variable if treatment has an effect. The running variable, *birth year*, allocates treatment just as in the multiple regression model (equation 5.1). The regression is the same as in equation (5.2), except that is it done separately for the two groups. It contains the same control variables.

In RD designs, the researcher must weigh the pros and cons of wide and narrow *bandwidths*. Wider bandwidths give more observations, but it will most likely make the control and treatment group more different on other relevant variables, thus damaging the resemblance of an experiment (Finseraas & Kotsadam, 2013). I choose to use the same bandwidth as Lleras-Muney (2005) and Albouy and Lequien (2009). With three cohorts on both sides of the cutoff

the estimation has sufficient observations without making the control and treatment groups too different to each other in observed and unobserved variables.

Three techniques will be used to check the robustness of the estimated LATE. First, the main model should be run with different bandwidths. If the model specification is correct, then we can expect the LATE to be stable across bandwidths. Second, to check that the groups are equal in other observed variables, RD models can be run on control variables to confirm that there are no discontinuities. We assume that the control variables are not affected by treatment. Last, placebo analyses on other cutoffs will indicate if the running variable is fit to allocate treatment at what we believe to be the cutoff. Discontinuities at placebo cutoffs could mean that the changes in the dependent variable is caused by unobserved trends or other interventions than the one we sought to test (Finseraas & Kotsadam, 2013).

6 Data

The Russian Longitudinal Monitoring Survey (RLMS-HSE, 2020) is a panel data set of interviews with Russian nationals. It was designed to monitor effects of Russian reforms on health and economic welfare and has been used in a variety of research projects. The survey has been conducted annually since 1994, except 1997 and 1999, and is organized in one data set for households and one for individuals. Some individuals are only interviewed once, others have participated almost every year. Therefore, the panel is unbalanced.

In this thesis I use individual data. Since we are interested in health outcomes in adulthood, all observations of the individuals when they were younger than 25 years old are excluded. This way we remove extreme data points from children and young adults that have "not yet been treated". Remember that returns on education come at later ages.

Estimations will be done on three different dependent variables: Self-reported health, body mass index and chronic health conditions. SRH is coded in five categories, from 'very bad' $=$ 1 to 'very good' $= 5$. Therefore, a positive effect on SRH indicates an improvement in health. Missing values and "Does not answer" are removed to avoid spikes and bias in the data.

BMI is calculated from self-reported weight and self-reported height as $\frac{weight(kg)}{height(m)^2} = kg/m^2$, and is categorized as follows (WHO, 2020):

- Underweight (unhealthy): BMI $= [0,18.5\rangle$
- Normal (healthy) weight: BMI $= [18.5, 25\rangle$
- Overweight (unhealthy): BMI $= [25, 30\rangle$
- Obese (unhealthy): BMI $= [30, \rightarrow\rangle$

Like Arendt (2005) does, I recode BMI as a binary variable. Unhealthy weight $= 0$ and healthy weight $= 1$. Therefore, a positive effect on BMI indicates an improvement in health. Biased and extreme values are removed.

The data offers information on a wide range of chronic health conditions related to heart, lung, liver, kidney, stomach, spinal, endocrine (diabetes) and hypertension. The dependent variable is not an index. It is coded binary so that $CC = 0$ if the individual reports one or more chronic illnesses, and $CC = 1$ if the individual reports no chronic illness. Like the variables for SRH and BMI, an increase in CC indicates an improvement in health. Individuals with a CC acquired as child or with a congenital or hereditary cause are removed since we only have

interest in outcomes that may be affected by education and therefore acquired in adulthood. Observations from individuals not answering are also removed.

With a bandwidth equal to three cohorts on each of the cutoff, we are left with approximately 4,300 individuals and 23,000 observations. To control for demographic variables that may affect health outcomes, I include birth year, age, gender, region, and settlement-type as control variables in all models, both for the multiple regression and sharp RD regressions. Much of existing literature use the same control variables. The data contains information on the individual's completed stage of education. However, in the Soviet and Russian school system some people completed lower secondary school after 8 years and some after 9 years in school. Therefore, the data does not tell us how many years in school the individual has. The only way to check an individual's treatment for Reform 1984 is in the birth year cutoff, as described in the previous chapter.

Table 1 below shows descriptive statistics. The three samples are close to equal in size, and the independent variables are approximately equal across the samples.

Table 1 Descriptive statistics

	Observations	Mean	Std. dev.	Min	Median	Max
Dependent variables						
SRH	24,220	3.52	0.61	1	4	5
BMI	23,308	24.92	4.55	12.17	24.22	54.88
BMI, binary	23,308	0.54	0.44	0	1	1
CC	22,896	0.73	0.50	0	1	1
Explanatory variable						
Treatment condition		0.46[a]	0.50[a]	0	0	1
Control variables						
Age		32.26[a]	4.18[a]	25	32	41
Gender		1.53[a]	0.50[a]	1	2	2
Settlement-type		2.14[a]	1.20[a]	1	2	4

a Different samples for each dependent variable gives three unique descriptive statistics for independent variables. Mean and SD vary little to nothing, but these numbers are the average of the three.
b The statistics here include cohorts [1980,1982] in treatment group and [1977,1979] in control group.
c Gender: 1 = Male, 2 = Female. Settlement-type: 1 = Oblast center, 2 = Town, 3 = Urban-type, 4 = Rural
d SRH: 1 = Very bad, 5 = Very good. BMI: 0 = Unhealthy, 1 = Healthy. CC: 0 = Diagnosed, 1 = Healthy.

7 Results

7.1 Multiple Linear Regression, OLS

First, I present the OLS results. These results were estimated to show how the causal effects in the RD design compares to the statistical correlations. Table 2 below contains OLS coefficients for the treatment variable and the control variables. Most control variables significantly correlate to the outcome. As expected, higher ages are correlated with worse SRH, unhealthier BMI and higher probabilities of acquiring chronic health conditions. The sign of the other controls varies somewhat across the different dependent variables.

The treatment group correlates significantly with all outcomes. As expected, the signs indicate that being born after the treatment cutoff relates to better SRH at 5% significance and healthier BMI at 1% significance. However, the correlation with CC is negative at 10% significance, indicating that the treatment group experiences more chronic health conditions than the control group.

Table 2 Multiple linear regression (OLS)

	Self-reported health	Body mass index	Chronic health conditions
Constant	-6.053 (9.631)	37.17*** (7.907)	-11.86* (7.116)
Implementation 1986	0.038** (0.016)	0.046*** (0.013)	-0.024* (0.012)
Birth year	0.005 (0.005)	-0.018*** (0.004)	0.007* (0.004)
Age	-0.006*** (0.001)	-0.016*** (0.0008)	-0.014*** (0.0007)
Gender	-0.106*** (0.008)	0.085*** (0.006)	-0.039*** (0.006)
Region	0.00007 (0.00009)	-0.0006*** (0.00007)	-0.00004 (0.00007)
Settlement-type	0.019*** (0.003)	-0.030*** (0.003)	0.020*** (0.002)
Observations Men Women	24,220 2,172 2,204	23,308 2,134 2,188	22,896 2,113 2,128

a *Significant at 10%; **Significant at 5%; ***Significant at 1%. Standard errors in parentheses.
b Bandwidth = 3. Cohorts [1980,1982] in treatment group and [1977,1979] in control group.
c Gender: 1 = Male, 2 = Female. Settlement-type: 1 = Oblast center, 2 = Town, 3 = Urban-type, 4 = Rural
d SRH: 1 = Very bad, 5 = Very good. BMI: 0 = Unhealthy, 1 = Healthy. CC: 0 = Diagnosed, 1 = Healthy.

7.2 Sharp Regression Discontinuity Design

Table 3 on the next page contains estimates from all the sharp RD regressions. As you can see, there are no coefficients for control variables. The software function for RD does not print these coefficients. However, their inclusion does affect the LATE.

The main results in the top row show that being assigned to treatment conditions does improve SRH at 10% significance and BMI at 1% significance, just as hypothesized. In contrast, treatment worsen the probability of acquiring CC at 10% significance. All three LATEs are consistent with the OLS coefficients with the same sign and almost equal size. The RD conducted here explain most of the correlation observed in a regular OLS regression.

The first test of robustness is checking bandwidths of 2 and 4 cohorts on each side of the cutoff. The LATE stays stable for BMI at both bandwidths, but only at bandwidth 4 for SRH and CC. Nonetheless, the insignificant treatment effects are remarkably close to its correspondence in the main model. My interpretation is that we have reason to believe the model specification is correct.

The second test of robustness is checking that the control and treatment groups are equal in observed variables. There are no discontinuities at the cutoff for age, gender or region, which is as expected. However, there is a strong significant discontinuity for settlement-type on all three outcomes. The sign is negative, indicating that the treatment group settles in urban areas more than the control group. It might be caused by an underlying trend. Something happened that made the 2-3 years younger generation attract towards cities, towns and urban-type areas, rather than rural areas. Though unexpected, I would not interpret it as damaging for the LATE of Reform 1986.

The last test of robustness is checking placebo cutoffs. Some of the LATEs on placebo cutoffs are insignificant, but some are significant. Usually, discontinuities at other cutoffs indicate that changes in the dependent variables may be caused by unobserved trends or other happenings than the one we assumed would create discontinuity. However, have in mind that the treatment conditions are mixed around the cutoff and that the treatment probability is below one. The placebo discontinuities are not necessarily caused by other interventions than the school reform. It might be the school reform and school enrollment at age 6, because we know the cutoff is not sharp. Some pupils enrolled at age 6 in the years before 1980, and the proportion of six-year-old pupils enrolling after 1980 may have increased from year to year, resembling an intervention.

Table 3 Sharp regression discontinuity design

	Self-reported health	Body mass index	Chronic health conditions
Treatment ≥ 1980			
Implementation 1986[b]	0.034* (0.017)	0.054*** (0.014)	-0.025* (0.013)
Implementation 1986 bw = 2	0.029 (0.022)	0.067*** (0.018)	-0.025 (0.016)
Implementation 1986 bw = 4	0.036** (0.015)	0.035*** (0.012)	-0.020* (0.011)
Age[b]	-0.114 (0.114)	-0.131 (0.116)	-0.113 (0.117)
Gender[b]	-0.001 (0.014)	0.0008 (0.014)	-0.001 (0.015)
Region[b]	0.003 (1.224)	-0.336 (1.248)	0.123 (1.261)
Settlement-type[b]	-0.099*** (0.034)	-0.099*** (0.035)	-0.135*** (0.035)
Observations at c ≥ 1980	24,220	23,308	22,896
Placebo cutoffs			
Placebo #1 (Treatment ≥ 1976)	0.037** (0.016)	-0.025* (0.013)	0.019 (0.012)
Placebo #2 (Treatment ≥ 1978)	-0.019 (0.017)	-0.030** (0.014)	-0.004 (0.012)
Placebo #3 (Treatment ≥ 1982)	-0.027 (0.018)	0.019 (0.015)	0.028** (0.013)
Placebo #4 (Treatment ≥ 1984)	0.033* (0.018)	-0.031** (0.015)	-0.009 (0.013)

a *Significant at 10%; **Significant at 5%; ***Significant at 1%. Standard errors in parentheses.
b Bandwidth = 3, cutoff ≥ 1980. Cohorts [1980,1982] in treatment group and [1977,1979] in control group.
c Gender: 1 = Male, 2 = Female. Settlement-type: 1 = Oblast center, 2 = Town, 3 = Urban-type, 4 = Rural
d SRH: 1 = Very bad, 5 = Very good. BMI: 0 = Unhealthy, 1 = Healthy. CC: 0 = Diagnosed, 1 = Healthy.

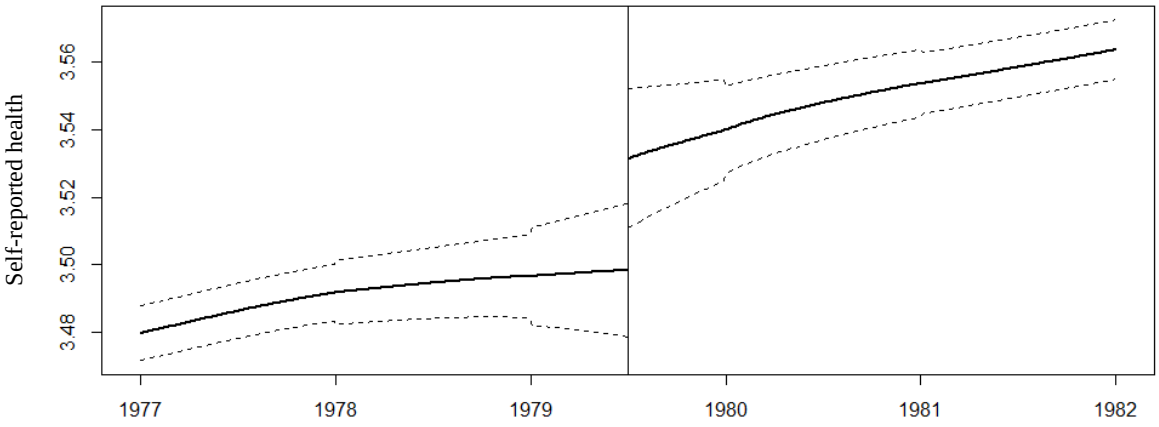


Figure 2 Sharp RD on self-reported health

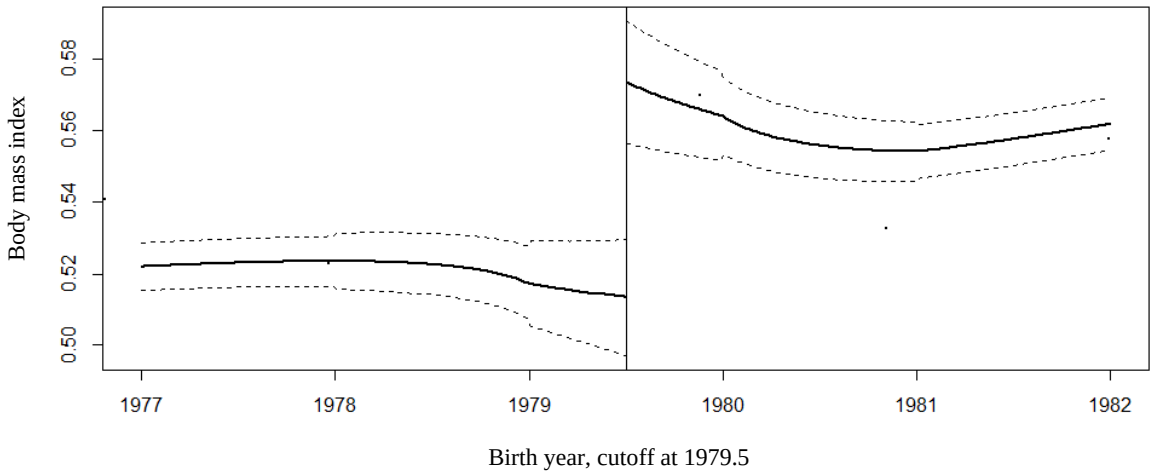


Figure 3 Sharp RD on body mass index

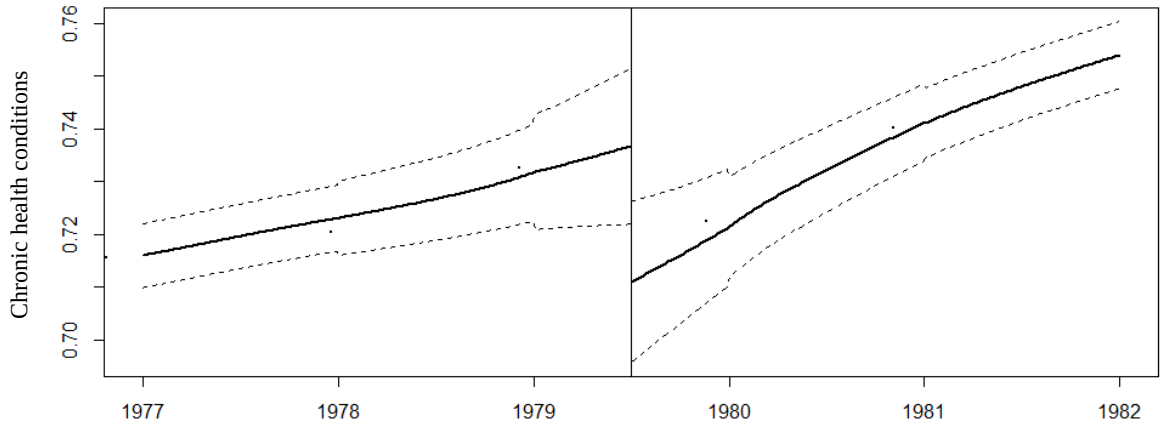


Figure 4 Sharp RD on chronic health conditions

Figures for the three main models in the RD design is seen on the previous page. First, look at figure 3. Since treatment improves BMI at 1% significance, the visual *jump* in BMI is large. Relative to the LATE, the small standard error puts the confidence interval of the two OLS regressions far from each other with no overlap. Graphically, it looks like the linear model is a correct specification. The common trend assumption for the control and treatment group seem to hold.

In figure 2 and 4, we notice that the discontinuity is smaller and that the confidence intervals overlap slightly. Remember that these models are significant only at a 10% level. For SRH and CC, the standard error is closer to the LATE than in the BMI model. Nonetheless, the control and treatment groups in both models seem to follow a similar linear trend.

8 Discussion

This thesis set out to estimate the effect of earlier entry age in the Russian primary school on a list of health outcomes in adulthood. The theoretical foundation of my research topic is human capital theory and Becker's (2007) prediction that education investments have both direct and indirect effects on health outcomes. Reviewing literature showed mixed results, indicating that the field can still enjoy further research on SRH, BMI and CC. Especially Russian data and school entry age as the treatment would be welcomed contributions previously unaddressed in the setting of causality from education to health.

The method used here to estimate causal inference is the sharp RD design. Despite the unknown treatment probability proposing the fuzzy RD, checking robustness in placebo bandwidths, placebo outcomes, placebo cutoffs and visual plots, shows that the sharp RD performed well. However, the levels of the LATEs might be unprecise because of the non-compliance problem and possible errors in the implementation year. One important implication to have in mind when conducting RD approaches is that the treatment effect cannot be generalized. The LATE is valid only at the cutoff. It implies that the USSR Educational Reform of 1984 did improve the health of cohorts 1980-1982 compared to cohorts 1977-1979, but we do not know the effect it would have had on other cohorts.

Consistent with Arendt's (2005) results, I found that more educated are more likely to have healthier BMI. Our research differs in two ways. Arendt estimated the treatment effect in a random effects logistic regression, and his treatment on education was increased school leaving age, which was not a characteristic of the school reform I used. The implication is that these results combined suggest education might affect BMI in multiple pathways. It may be that the health production skills that improves BMI can be acquired in several mechanisms of education. For SRH, both Arendt and I show that more education gives better SRH. It is difficult to compare the level of our treatment effects since our methods print different types of output.

Also Fonseca et al. (2020) estimated that compulsory schooling gives better SRH. However, they found that compulsory schooling reduced the chance of acquiring CC such as heart diseases and diabetes, which is not consistent with my results. The RD conduction here show that education cause higher probabilities for CC. Do keep in mind that the dependent variable used in this thesis measures several CC collectively. We do not know the results if we estimated for heart diseases and diabetes individually. If it is true that the treatment group

here acquire more CC, it is not necessarily because of education. There might be unobserved trends such as genetics. Or it might be that CC are not affected by education to the same degree as SRH and BMI. Janke et al. (2020) found no significant impacts of education on CC, except for education reducing the probability of having diabetes.

It is difficult to be certain that no unobserved trends contribute to the statistically significant results found in this thesis. Reform 1984 comprised several changes to the school system. We cannot defend it for being *ceteris paribus* when lowering school entry age, but as seen in the previous chapter, the RD approach performed well in many ways.

In the introduction of this thesis, I questioned whether investments in childhood education cause improved health later in life. The results from an RD analysis do imply so and show us that even the youngest children in school acquire skills that contribute to their health production. For policy makers, these results should be interesting when preparing new education and health policies. However, I recommend more research on the effects of early education stages before basing any policy solely on these results. The evidence presented here would gain from addition and more precise estimations done on better data.

Future research should also investigate other education-related mechanisms at young ages. For example, do we know anything about the effect of more formal educational programs versus more social and playful programs at pre-schools and kindergarten ages? And what curriculum best ensures children's well-being both short- and long-term? An important thing to remember is that adulthood health is not the only outcome that matter. Children should feel healthy and happy during their childhood years and adolescence.

Russia is a unique society that differs from other European countries. Not only because of its transitioning economy, social inequalities, and public health challenges, but also in how the authorities require children to study. As I mentioned in chapter 4, the educational system in the USSR strictly enforced the Communist doctrine from young ages, and the enrolling of children aged 6 has received a wide range of negative critiques from social scientists and psychologists. Some people warned against replacing young children's playful activities with more formal education. The stereotype that Russia enforces strict and collective education routines that lacks individual development still exists today. On this matter, Russia differs both from developed and developing countries, making it an interesting and fruitful subject of research.

This thesis has shown that education leads to better self-reported health and healthier BMI, which is consistent with Becker's (2007) predictions. We know that healthy SRH and BMI is related to better employability, and thus increasing production. It is fair to say that improvements in such health outcomes increase the returns on human capital and educational investments.

Further Research

The literature review in chapter 3 implied several gaps in the research field which is outside the topic of this thesis. These gaps are mostly untouched and should not be left unmentioned even if they are outside the scope of this project. As repeated several times now, most existing evidence address the same identification: Compulsory number of years in school increased by an increase in the minimum school leaving age. However, there are many more characteristics and mechanisms of education that we suspect of having causal inferences on health, without having been properly studied.

First, what are the effects from other stages in school than leaving age? This thesis attempted to contribute to fill this gap, but I still warrant more research. I believe that pre-school, primary school, secondary school and higher education affect health differently. Pre-school pupils and university students do not process health information and manage health behaviors the same way.

Second, what are the effects of voluntary educational attainment? Existing research address strictly enforced compulsory schooling, but Albouy and Lequien (2009) mentions that the student's motivation might influence the effect of schooling. Avendano et al. (2020) suggest that coercive education can worsen mental health and consequently worsen the individual's labor-related outcomes. We need research on how voluntary and motivated education affect health compared to strictly coerced education.

Third, existing research address old reforms for low initial levels of education. Researchers often write that evidence should be collected on developing countries since they might not have had the effect Western countries got from their school reforms the last century. Both Lleras-Muney (2005) and Fonseca et al. (2020) imply that countries with low levels of education should be more aggressive in policy making.

The last gap concerns the quality of education rather than the number of years in school. What are the effects of teacher-student ratio, curriculums, test scores, and self-reported satisfaction with education, on health? All these questions deserve attention. We must think differently and collect data in a way that lets us investigate mechanisms of education we so far have found too difficult to measure. We can plan and design observational data better. Research should seek to not only analyze what has already happened, but also decide what to analyze in the future before it happens.

References

Albouy, V., & Lequien, L. (2009). Does compulsory education lower mortality? *Journal of Health Economics, 28*(1), 155-168. doi:https://doi.org/10.1016/j.jhealeco.2008.09.003

Arendt, J. N. (2005). Does education cause better health? A panel data analysis using school reforms for identification. *Economics of Education Review, 24*(2), 149-160. doi:https://doi.org/10.1016/j.econedurev.2004.04.008

Avendano, M., de Coulon, A., & Nafilyan, V. (2020). Does longer compulsory schooling affect mental health? Evidence from a British reform. *Journal of Public Economics, 183*, 104137. doi:https://doi.org/10.1016/j.jpubeco.2020.104137

Barro, R. J., & Sala-i-Martin, X. (2004). *Economic Growth* (Second ed.). Cambridge: MIT Press.

Becker, G. S. (1993). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education* (Third ed.). Chicago: The University of Chicago Press.

Becker, G. S. (2007). Health as human capital: synthesis and extensions1. *Oxford Economic Papers, 59*(3), 379-410. doi:10.1093/oep/gpm020

Braakmann, N. (2011). The causal relationship between education, health and health related behaviour: Evidence from a natural experiment in England. *Journal of Health Economics, 30*(4), 753-763. doi:https://doi.org/10.1016/j.jhealeco.2011.05.015

Brunello, G., Fabbri, D., & Fort, M. (2013). The Causal Effect of Education on Body Mass: Evidence from Europe. *Journal of Labor Economics, 31*(1), 195-223. doi:10.1086/667236

Clark, D., & Royer, H. (2013). The Effect of Education on Adult Mortality and Health: Evidence from Britain. *American Economic Review, 103*(6), 2087-2120. doi:10.1257/aer.103.6.2087

Cutler, D. M., & Lleras-Muney, A. (2006). Education and Health: Evaluating Theories and Evidence. *National Bureau of Economic Research Working Paper Series, No. 12352*. doi:10.3386/w12352

Eide, E. R., & Showalter, M. H. (2011). Estimating the relation between health and education: What do we know and what do we need to know? *Economics of Education Review, 30*(5), 778-791. doi:https://doi.org/10.1016/j.econedurev.2011.03.009

Eklof, B., & Dneprov, E. (1993). *Democracy in the Russian School: The Reform Movement in Education Since 1984* (B. Eklof & E. Dneprov Eds.). Boulder: Westview Press.

Finseraas, H., & Kotsadam, A. (2013). Hvordan identifisere årsakssammenhenger i ikke-eksperimentelle data? En ikke-teknisk introduksjon. *Tidsskrift for samfunnsforskning, 54*(3), 371-387. Retrieved from http://www.idunn.no/tfs/2013/03/hvordan_identifisere_aarsakssammenhenger_i_ikke-eksperimente

Fonseca, R., Michaud, P.-C., & Zheng, Y. (2020). The effect of education on health: evidence from national compulsory schooling reforms. *SERIEs, 11*(1), 83-103. doi:10.1007/s13209-019-0201-0

Gidadhubli, R. G. (1984). Reform of School Education. *Economic and Political Weekly, 19*(40), 1737-1741. Retrieved from https://www.jstor.org/stable/4373646

Goldin, C. (2016). Human Capital. In *Handbook of Cliometrics*. Heidelberg, Germany: Springer Verlag.

Grossman, M. (2015). The Relationship between Health and Schooling: What's New? *National Bureau of Economic Research Working Paper Series, No. 21609*. doi:10.3386/w21609

Hill, R. C., Griffiths, W. E., & Lim, G. C. (2018). *Principles of Econometrics* (5th ed.). New Jersey: John Wiley & Sons.

Janke, K., Johnston, D. W., Propper, C., & Shields, M. A. (2020). The causal effect of education on chronic health conditions in the UK. *Journal of Health Economics, 70*, 102252. doi:https://doi.org/10.1016/j.jhealeco.2019.102252

Jones, C. I., & Vollrath, D. (2013). *Introduction to Economic Growth* (Third ed.). New York: W. W. Norton & Company.

Jürges, H., Kruk, E., & Reinhold, S. (2013). The effect of compulsory schooling on health—evidence from biomarkers. *Journal of Population Economics, 26*(2), 645-672. doi:10.1007/s00148-012-0409-9

Jürges, H., & Meyer, S.-C. (2020). Educational Differences in Smoking: Selection Versus Causation. *Jahrbücher für Nationalökonomie und Statistik, 240*(4), 467-492. doi:https://doi.org/10.1515/jbnst-2019-0004

Jürges, H., Reinhold, S., & Salm, M. (2011). Does schooling affect health behavior? Evidence from the educational expansion in Western Germany. *Economics of Education Review, 30*(5), 862-872. doi:https://doi.org/10.1016/j.econedurev.2011.04.002

Kemptner, D., Jürges, H., & Reinhold, S. (2011). Changes in compulsory schooling and the causal effect of education on health: Evidence from Germany. *Journal of Health Economics, 30*(2), 340-354. doi:https://doi.org/10.1016/j.jhealeco.2011.01.004

Lleras-Muney, A. (2005). The Relationship Between Education and Adult Mortality in the United States. *The Review of Economic Studies, 72*(1), 189-221. doi:10.1111/0034-6527.00329

Ma, Y., Nolan, A., & Smith, J. P. (2018). The value of education to health: Evidence from Ireland. *Economics & Human Biology, 31*, 14-25. doi:https://doi.org/10.1016/j.ehb.2018.07.006

Mankiw, N. G., Romer, D., & Weil, D. N. (1992). A Contribution to the Empirics of Economic Growth*. *The Quarterly Journal of Economics, 107*(2), 407-437. doi:10.2307/2118477

Mironov, B. N. (1991). The Development of Literacy in Russia and the USSR from the Tenth to the Twentieth Centuries. *History of Education Quarterly, 31*(2), 229-252. doi:10.2307/368437

Oxford Learner's Dictionaries. (2020). human capital. *Oxford University Press*. Retrieved from https://www.oxfordlearnersdictionaries.com/definition/english/human-capital

RLMS-HSE. (2020). "Russia Longitudinal Monitoring survey, RLMS-HSE", conducted by the National Research University Higher School of Economics and ZAO "Demoscope" together with Carolina Population Center, University of North Carolina at Chapel Hill and the Institute of Sociology RAS. Retrieved from https://www.cpc.unc.edu/projects/rlms-hse

Silles, M. (2015). The causal effect of schooling on smoking behavior. *Economics of Education Review, 48*, 102-116. doi:https://doi.org/10.1016/j.econedurev.2015.06.004

Sutherland, J. (1999). *Schooling in the new Russia : innovation and change, 1984-95*. New York: St. Martin's Press.

Szekely, B. B. (1986). The New Soviet Educational Reform. *Comparative Education Review, 30*(3), 321-343. Retrieved from https://www.jstor.org/stable/1188037

WHO. (2020). Mean Body Mass Index (BMI). Retrieved from https://www.who.int/gho/ncd/risk_factors/bmi_text/en/

Zhong, H. (2015). Does a college education cause better health and health behaviours? *Applied Economics, 47*(7), 639-653. doi:10.1080/00036846.2014.978074

Zhong, H. (2016). Effects of quantity of education on health: A regression discontinuity design approach based on the Chinese Cultural Revolution. *China Economic Review, 41*, 62-74. doi:https://doi.org/10.1016/j.chieco.2016.08.011

Appendix

The appendix covers all R codes necessary to replicate my research. The RLMS-HSE (2020) data can be found on their website.

Self-Reported Health

```
# Preparatory Work ####

# Clean environment.
rm(list=ls())

# What folder is RStudio reading from?
getwd()
setwd("C:/Users/Bruker/Desktop/RLMS")

# List files.
dir()

# Load data frame.
load("RLMS_IND.RData")

# Rename data frame.
rlms <- USER_RLMS_HSE_IND_1994_2018_v2_eng
remove(USER_RLMS_HSE_IND_1994_2018_v2_eng)

# Load packages.
library(tidyverse)
library(mosaic)
library(dplyr)
library(car)

# Reduce size of data frame by removing irrelevant cohorts.
# Keep 10 cohorts on each side of the cutoff to have enough data for place
bo tests.
rlms <- rlms[rlms$H6 < 1990,]
rlms <- rlms[rlms$H6 > 1969,]

# Generate age variable.
rlms$age <- rlms$YEAR - rlms$H6

# We are only interesting in health outcomes in adulthood and do not want
# information from childhood to make extremes in the data.
# Like Arendt (2005), exclude observations where age < 25.
rlms <- rlms[rlms$age > 24,]

# Create treatment variables for implementation in 1986.
# Cohort 1980 was the first treated.
rlms$x <- as.numeric(rlms$H6 >= 1980)

# Create treatment variables for implementation in 1990.
```

```r
# Cohort 1984 was the first treated.
rlms$y <- as.numeric(rlms$H6 >= 1984)

# Data frame is now smaller and prepared for analysis.
setwd("C:/Users/Bruker/Desktop/master/data-frames")
save(rlms, file = "RLMS_PREP.RData")

# Select variables for analysis.
rlms_srh <- select(rlms,
                   ID_W, IDIND, YEAR, REGION, STATUS, H5, H6, M3, age, x,
y)
remove(rlms)

# Dependent variable is health outcome measured as self-reported health.
# Remove observations with no answer.
rlms_srh <- rlms_srh[rlms_srh$M3 < 6,]

# Flip the direction of the variable so that 1 is poor SRH and 5 is great
SRH.
table(rlms_srh$M3)
rlms_srh$srh <- recode(rlms_srh$M3, '1=5; 2=4; 4=2; 5=1')
table(rlms_srh$srh)
rlms_srh$M3 <- NULL

# Data frame is now smaller and prepared for analysis.
setwd("C:/Users/Bruker/Desktop/master/data-frames")
save(rlms_srh, file = "RLMS_SRH.RData")

# Sharp Regression Discontinuity Design ####

# Clean environment.
rm(list=ls())

# Load packages.
library(plyr)
library(tidyverse)
library(mosaic)
library(dplyr)

# Load data frame.
setwd("C:/Users/Bruker/Desktop/master/data-frames")
load("RLMS_SRH.RData")

# Load the package 'rdd'.
library(rdd)

# Count the number of men and women in the sample of +/- 3 cohorts on each
side of the cutoff.
# Copy data frame.
last <- rlms_srh
# Remove all but the last observation for every individual to make it cros
s-sectional.
last <- ddply(last, .(IDIND), function(X) X[which.max(X$YEAR), ])
# Bandwidth 3.
```

```r
last <- last[last$H6 < 1983 & last$H6 > 1976, ]
table(last$H5)
# There are 2172 men and 2204 women in the sample.
remove(last)

## Simple Linear Regression, Reform 1986 ####

# Find the LATE of enrolling at age 6 on SRH in adulthood using a simple l
inear regression.
# Use a bandwidth of three cohorts in control and treatment group.
rlms_srh1 <- rlms_srh[rlms_srh$H6 < 1983 & rlms_srh$H6 > 1976, ]
lm <- lm(srh ~ x + H6 + age + H5 + REGION + STATUS, data = rlms_srh1)
summary(lm)
#' A simple linear regression indicate that treatment has a significant ef
fect leading to better SRH.
#' LATE estimate = 0.038** with standard error = 0.016.
#' In addition, three out of our control variables are significantly corre
lated with SRH.
#' Women are correlated with worse SRH than men.
#' Higher ages are correlated with worse SRH.
#' A more rural settlement-type is correlated with better SRH than cities
and towns.
#' People born in 1980-1982 are suspected to acquire better self-reported
health in adulthood than those born in 1977-1979.

### Sharp RD, Reform 1986 ####

# Bandwidth 3.
rd <- RDestimate(formula = srh ~ H6 | age + H5 + REGION + STATUS,
                 data = rlms_srh, cutpoint = 1979.5, bw = 3)
summary(rd)
#' LATE estimate = 0.034* with standard error = 0.017.
#' Treatment significantly improves SRH at 10%.

# Bandwidth 2.
rd2 <- RDestimate(formula = srh ~ H6 | age + H5 + REGION + STATUS,
                 data = rlms_srh, cutpoint = 1979.5, bw = 2)
summary(rd2)
#' LATE estimate = 0.029 with standard error = 0.022.
#' Treatment does not affect SRH.

# Bandwidth 4.
rd4 <- RDestimate(formula = srh ~ H6 | age + H5 + REGION + STATUS,
                  data = rlms_srh, cutpoint = 1979.5, bw = 4)
summary(rd4)
#' LATE estimate = 0.036** with standard error = 0.015.
#' Treatment significantly improves SRH at 5%.

# Plot discontinuity
plot(rd.st, range = c(1977,1982))
abline(v = 1979.5)

### Sharp RD, Placebo Outcomes ####
```

```r
# Age
rd.age <- RDestimate(formula = age ~ H6 | H5 + REGION + STATUS,
                         data = rlms_srh, cutpoint = 1979.5, bw = 3)
summary(rd.age)
#' No discontinuity on age at cutoff.

# Gender
rd.gender <- RDestimate(formula = H5 ~ H6 | age + REGION + STATUS,
                  data = rlms_srh, cutpoint = 1979.5, bw = 3)
summary(rd.gender)
#' No discontinuity on gender at cutoff.

# Region
rd.region <- RDestimate(formula = REGION ~ H6 | age + H5 + STATUS,
                    data = rlms_srh, cutpoint = 1979.5, bw = 3)
summary(rd.region)
#' No discontinuity on gender at cutoff.

# Settlement-type
rd.st <- RDestimate(formula = STATUS ~ H6 | age + H5 + REGION,
                     data = rlms_srh, cutpoint = 1979.5, bw = 3)
summary(rd.st)
#' Treatment significantly affects settlement-type.

### Sharp RD, Placebo Cutoffs ####

# Perform placebo tests to check for discontinuities at other cutoffs.
# Four years prior.
placebo1 <- RDestimate(formula = srh ~ H6 | age + H5 + REGION + STATUS,
                       data = rlms_srh, cutpoint = 1975.5, bw = 3)
summary(placebo1)
# Significant improvement in SRH at 5%.

# Two year prior.
placebo2 <- RDestimate(formula = srh ~ H6 | age + H5 + REGION + STATUS,
                       data = rlms_srh, cutpoint = 1977.5, bw = 3)
summary(placebo2)
# No discontinuity or effects.

# Two years after.
placebo3 <- RDestimate(formula = srh ~ H6 | age + H5 + REGION + STATUS,
                       data = rlms_srh, cutpoint = 1981.5, bw = 3)
summary(placebo3)
# No discontinuity or effects.

# Four years after.
placebo4 <- RDestimate(formula = srh ~ H6 | age + H5 + REGION + STATUS,
                       data = rlms_srh, cutpoint = 1983.5, bw = 3)
summary(placebo4)
# Significant improvement in SRH at 10%.
```

Body Mass Index

```
# Preparatory Work ####

# Clean environment.
rm(list=ls())

# Load packages.
library(tidyverse)
library(mosaic)
library(dplyr)
library(car)

# What folder is RStudio reading from?
getwd()
setwd("C:/Users/Bruker/Desktop/master/data-frames")

# List files.
dir()

# Load data frame.
load("RLMS_PREP.RData")

# Select variables for analysis.
rlms_bmi <- select(rlms,
                   ID_W, IDIND, YEAR, REGION, STATUS, H5, H6,
                   M1, M2, age, x, y)
remove(rlms)

# Generate outcome variable for body mass index, kg/m^2.
rlms_bmi$bmi <- rlms_bmi$M1/(rlms_bmi$M2/100)^2

# After inspecting the data frame, remove unrealistic low and high BMI
# caused by "Does not answer".
rlms_bmi <- rlms_bmi[rlms_bmi$bmi > 10,]
rlms_bmi <- rlms_bmi[rlms_bmi$bmi < 100,]

# Generate categorical variable for BMI.
# 1 = Underweight
# 2 = Healthy weight
# 3 = Overweight
# 4 = Obese
rlms_bmi$bmiint <- cut(rlms_bmi$bmi, c(0,18.5,25,30,100), right=FALSE)
rlms_bmi$bmicat <- cut(rlms_bmi$bmi, c(0,18.5,25,30,100), right=FALSE, lab
els=c(1:4))

table(rlms_bmi$bmicat)
rlms_bmi$bmidep <- recode(rlms_bmi$bmicat, '1=0; 2=1; 3=0; 4=0')
table(rlms_bmi$bmidep)

# Change from factor to numeric.
bmidep.factor <- factor(rlms_bmi$bmidep)
rlms_bmi$bmidep <- as.numeric(bmidep.factor)
rlms_bmi$bmidep <- recode(rlms_bmi$bmidep, '1=0; 2=1')
```

```r
table(rlms_bmi$bmidep)

# Data frame is now smaller and prepared for analysis.
setwd("C:/Users/Bruker/Desktop/master/data-frames")
save(rlms_bmi, file = "RLMS_BMI.RData")

# Sharp Regression Discontinuity Design ####

# Clean environment.
rm(list=ls())

# Load packages.
library(plyr)
library(tidyverse)
library(mosaic)
library(dplyr)

# Load data frame.
setwd("C:/Users/Bruker/Desktop/master/data-frames")
load("RLMS_BMI.RData")

# Load the package 'rdd'.
library(rdd)

# Count the number of men and women in the sample of +/- 3 cohorts on each
side of the cutoff.
# Copy data frame.
last <- rlms_bmi
# Remove all but the last observation for every individual to make it cros
s-sectional.
last <- ddply(last, .(IDIND), function(X) X[which.max(X$YEAR), ])
# Bandwidth 3.
last <- last[last$H6 < 1983 & last$H6 > 1976, ]
table(last$H5)
# There are 2134 men and 2188 women in the sample.
remove(last)

## Simple Linear Regression, Reform 1986 ####

# Find the LATE of enrolling at age 6 on BMI in adulthood using a simple l
inear regression.
# Use a bandwidth of three cohorts in control and treatment group.
rlms_bmi1 <- rlms_bmi[rlms_bmi$H6 < 1983 & rlms_bmi$H6 > 1976, ]
lm <- lm(bmidep ~ x + H6 + age + H5 + REGION + STATUS, data = rlms_bmi1)
summary(lm)
#' A simple linear regression indicate that treatment has a significant ef
fect leading to healthier BMI.
#' LATE estimate = 0.046*** with standard error = 0.013.
#' In addition, all control variables are significantly correlated with BM
I.
#' Women are correlated with healthier BMI.
#' Higher ages are correlated with unhealthy BMI.
#' A more rural settlement-type is correlated with unhealthy BMI compared
to cities and towns.
```

```r
#' People born in 1980-1982 are suspected to acquire healthier BMI in adul
thood than those born in 1977-1979.

### Sharp RD, Reform 1986 ####

# Bandwidth 3.
rd <- RDestimate(formula = bmidep ~ H6 | age + H5 + REGION + STATUS,
                 data = rlms_bmi, cutpoint = 1979.5, bw = 3)
summary(rd)
#' LATE estimate = 0.054*** with standard error = 0.014.
#' Treatment significantly improves BMI at 1%.

# Bandwidth 2.
rd2 <- RDestimate(formula = bmidep ~ H6 | age + H5 + REGION + STATUS,
                 data = rlms_bmi, cutpoint = 1979.5, bw = 2)
summary(rd2)
#' LATE estimate = 0.067*** with standard error = 0.018.
#' Treatment significantly improves BMI at 1%.

# Bandwidth 4.
rd4 <- RDestimate(formula = bmidep ~ H6 | age + H5 + REGION + STATUS,
                 data = rlms_bmi, cutpoint = 1979.5, bw = 4)
summary(rd4)
#' LATE estimate = 0.035*** with standard error = 0.012.
#' Treatment significantly improves BMI at 1%.

# Plot discontinuity
plot(rd, range = c(1977,1982))
abline(v = 1979.5)

### Sharp RD, Placebo Outcomes ####

# Age
rd.age <- RDestimate(formula = age ~ H6 | H5 + REGION + STATUS,
                     data = rlms_bmi, cutpoint = 1979.5, bw = 3)
summary(rd.age)
#' No discontinuity on age at cutoff.

# Gender
rd.gender <- RDestimate(formula = H5 ~ H6 | age + REGION + STATUS,
                        data = rlms_bmi, cutpoint = 1979.5, bw = 3)
summary(rd.gender)
#' No discontinuity on gender at cutoff.

# Region
rd.region <- RDestimate(formula = REGION ~ H6 | age + H5 + STATUS,
                        data = rlms_bmi, cutpoint = 1979.5, bw = 3)
summary(rd.region)
#' No discontinuity on gender at cutoff.

# Settlement-type
rd.st <- RDestimate(formula = STATUS ~ H6 | age + H5 + REGION,
                    data = rlms_bmi, cutpoint = 1979.5, bw = 3)
summary(rd.st)
```

```r
#' Treatment significantly affects settlement-type.

### Sharp RD, Placebo Cutoffs ####

# Perform placebo tests to check for discontinuities at other cutoffs.
# Four years prior.
placebo1 <- RDestimate(formula = bmidep ~ H6 | age + H5 + REGION + STATUS,
                       data = rlms_bmi, cutpoint = 1975.5, bw = 3)
summary(placebo1)
# Significant deterioration in BMI at 10%.

# Two year prior.
placebo2 <- RDestimate(formula = bmidep ~ H6 | age + H5 + REGION + STATUS,
                       data = rlms_bmi, cutpoint = 1977.5, bw = 3)
summary(placebo2)
# Significant deterioration in BMI at 5%.

# Two years after.
placebo3 <- RDestimate(formula = bmidep ~ H6 | age + H5 + REGION + STATUS,
                       data = rlms_bmi, cutpoint = 1981.5, bw = 3)
summary(placebo3)
# No discontinuity or effects.

# Four years after.
placebo4 <- RDestimate(formula = bmidep ~ H6 | age + H5 + REGION + STATUS,
                       data = rlms_bmi, cutpoint = 1983.5, bw = 3)
summary(placebo4)
# Significant deterioartion in BMI at 5%.
```

Chronic Health Conditions

```r
# Preparatory Work ####

# Clean environment.
rm(list=ls())

# Load packages.
library(tidyverse)
library(mosaic)
library(dplyr)
library(car)

# What folder is RStudio reading from?
getwd()
setwd("C:/Users/Bruker/Desktop/master/data-frames")

# List files.
dir()

# Load data frame.
load("RLMS_PREP.RData")
```

```r
# Select variables for analysis.
rlms_cc <- select(rlms,
                  ID_W, IDIND, YEAR, REGION, STATUS, H5, H6, age, x, y,
                  M20.61, M20.61C, M20.62, M20.62C, M20.63, M20.63C, M20.6
4,
                  M20.64C, M20.65, M20.65C, M20.66, M20.66C, M20.620, M20.
69)
remove(rlms)

# Dependent variable is health outcome measured as having a chronic condit
ion with one or more the following organs:
# Heart, lung, liver, kidney, stomach, spinal, endocritine system (diabete
s), hypertension.
# The other outcome is having no chronic condition with any of the organs.
# Remove observations with biased or irrelevant cause of disease.
# That is, remove all conditions acquired as child or with a hereditary or
congenital cause.

# Chronic heart disease
# Filter out individuals with hereditary and congenital heart diseases, an
d those acquired for children.
heart <- rlms_cc %>% filter(M20.61C == 1 | M20.61C == 2 | M20.61C == 5)
# Note that the question of cause was not asked all years they were asked
about a diagnose.
# Therefore, we do not remove only those observations, but all observation
s from those individuals.
# Remove those IDs found in data frame 'heart' from main data frame.
rlms_cc <- rlms_cc[!(rlms_cc$IDIND %in% heart$IDIND),]
# We also have to remove NAs where there is no value.
rlms_cc <- rlms_cc[!is.na(rlms_cc$M20.61),]
# And observations where the respondent did not answer the question.
rlms_cc <- rlms_cc[rlms_cc$M20.61 < 3,]
remove(heart)

# Chronic lung disease
# Filter out individuals with hereditary and congenital lung diseases, and
those acquired for children.
lung <- rlms_cc %>% filter(M20.62C == 1 | M20.62C == 2 | M20.62C == 5)
# Note that the question of cause was not asked all years they were asked
about a diagnose.
# Therefore, we do not remove only those observations, but all observation
s from those individuals.
# Remove those IDs found in data frame 'lung' from main data frame.
rlms_cc <- rlms_cc[!(rlms_cc$IDIND %in% lung$IDIND),]
# We also have to remove NAs where there is no value.
rlms_cc <- rlms_cc[!is.na(rlms_cc$M20.62),]
# And observations where the respondent did not answer the question.
rlms_cc <- rlms_cc[rlms_cc$M20.62 < 3,]
remove(lung)

# Chronic liver disease
# Filter out individuals with hereditary and congenital liver diseases, an
d those acquired for children.
liver <- rlms_cc %>% filter(M20.63C == 1 | M20.63C == 2 | M20.63C == 5)
```

```r
# Note that the question of cause was not asked all years they were asked
about a diagnose.
# Therefore, we do not remove only those observations, but all observation
s from those individuals.
# Remove those IDs found in data frame 'liver' from main data frame.
rlms_cc <- rlms_cc[!(rlms_cc$IDIND %in% liver$IDIND),]
# We also have to remove NAs where there is no value.
rlms_cc <- rlms_cc[!is.na(rlms_cc$M20.63),]
# And observations where the respondent did not answer the question.
rlms_cc <- rlms_cc[rlms_cc$M20.63 < 3,]
remove(liver)

# Chronic kidney disease
# Filter out individuals with hereditary and congenital kidney diseases, a
nd those acquired for children.
kidney <- rlms_cc %>% filter(M20.64C == 1 | M20.64C == 2 | M20.64C == 5)
# Note that the question of cause was not asked all years they were asked
about a diagnose.
# Therefore, we do not remove only those observations, but all observation
s from those individuals.
# Remove those IDs found in data frame 'kidney' from main data frame.
rlms_cc <- rlms_cc[!(rlms_cc$IDIND %in% kidney$IDIND),]
# We also have to remove NAs where there is no value.
rlms_cc <- rlms_cc[!is.na(rlms_cc$M20.64),]
# And observations where the respondent did not answer the question.
rlms_cc <- rlms_cc[rlms_cc$M20.64 < 3,]
remove(kidney)

# Chronic stomach disease
# Filter out individuals with hereditary and congenital stomach diseases,
and those acquired for children.
stomach <- rlms_cc %>% filter(M20.65C == 1 | M20.65C == 2 | M20.65C == 5)
# Note that the question of cause was not asked all years they were asked
about a diagnose.
# Therefore, we do not remove only those observations, but all observation
s from those individuals.
# Remove those IDs found in data frame 'stomach' from main data frame.
rlms_cc <- rlms_cc[!(rlms_cc$IDIND %in% stomach$IDIND),]
# We also have to remove NAs where there is no value.
rlms_cc <- rlms_cc[!is.na(rlms_cc$M20.65),]
# And observations where the respondent did not answer the question.
rlms_cc <- rlms_cc[rlms_cc$M20.65 < 3,]
remove(stomach)

# Chronic spinal disease
# Filter out individuals with hereditary and congenital spinal diseases, a
nd those acquired for children.
spinal <- rlms_cc %>% filter(M20.66C == 1 | M20.66C == 2 | M20.66C == 5)
# Note that the question of cause was not asked all years they were asked
about a diagnose.
# Therefore, we do not remove only those observations, but all observation
s from those individuals.
# Remove those IDs found in data frame 'spinal' from main data frame.
rlms_cc <- rlms_cc[!(rlms_cc$IDIND %in% spinal$IDIND),]
```

```r
# We also have to remove NAs where there is no value.
rlms_cc <- rlms_cc[!is.na(rlms_cc$M20.66),]
# And observations where the respondent did not answer the question.
rlms_cc <- rlms_cc[rlms_cc$M20.66 < 3,]
remove(spinal)

# Endocrine Disease
rlms_cc$M20.620[ is.na(rlms_cc$M20.620) ] <- 0
rlms_cc <- rlms_cc[rlms_cc$M20.620 < 3,]

# Hypertension
rlms_cc$M20.69[ is.na(rlms_cc$M20.69) ] <- 0
rlms_cc <- rlms_cc[rlms_cc$M20.69 < 3,]

# Recode binary variables from Yes/No = 1/2 to = 1/0
rlms_cc$M20.61[ rlms_cc$M20.61 == 2 ] <- 0
rlms_cc$M20.62[ rlms_cc$M20.62 == 2 ] <- 0
rlms_cc$M20.63[ rlms_cc$M20.63 == 2 ] <- 0
rlms_cc$M20.64[ rlms_cc$M20.64 == 2 ] <- 0
rlms_cc$M20.65[ rlms_cc$M20.65 == 2 ] <- 0
rlms_cc$M20.66[ rlms_cc$M20.66 == 2 ] <- 0
rlms_cc$M20.620[ rlms_cc$M20.620 == 2 ] <- 0
rlms_cc$M20.69[ rlms_cc$M20.69 == 2 ] <- 0

# Calculate new variable = 0 if no conditions diagnosed and > 0 if one or
more conditions diagnosed.
rlms_cc$totalcc <- rlms_cc$M20.61 + rlms_cc$M20.62 + rlms_cc$M20.63 + rlms
_cc$M20.64 + rlms_cc$M20.65 +
  rlms_cc$M20.66 + rlms_cc$M20.620 + rlms_cc$M20.69

# Change all values > 0 to equal 1.
rlms_cc$totalcc[ rlms_cc$totalcc > 1 ] <- 1
# Dependent variable = 1 if diagnosed with one or more conditions, or = 0
if diagnosed with no conditions.
rlms_cc$cc <- rlms_cc$totalcc

# Recode so that 0 = unhealthy and 1 = healthy.
table(rlms_cc$cc)
rlms_cc$cc <- recode(rlms_cc$cc, '0=1; 1=0')

# Data frame is now smaller and prepared for analysis.
setwd("C:/Users/Bruker/Desktop/master/data-frames")
save(rlms_cc, file = "RLMS_CC.RData")

# Sharp Regression Discontinuity Design ####

# Clean environment.
rm(list=ls())

# Load packages.
library(plyr)
library(tidyverse)
library(mosaic)
library(dplyr)
```

```r
# Load data frame.
setwd("C:/Users/Bruker/Desktop/master/data-frames")
load("RLMS_CC.RData")

# Load the package 'rdd'.
library(rdd)

# Count the number of men and women in the sample of +/- 3 cohorts on each
side of the cutoff.
# Copy data frame.
last <- rlms_cc
# Remove all but the last observation for every individual to make it cros
s-sectional.
last <- ddply(last, .(IDIND), function(X) X[which.max(X$YEAR), ])
# Bandwidth 3.
last <- last[last$H6 < 1983 & last$H6 > 1976, ]
table(last$H5)
# There are 2134 men and 2128 women in the sample.
remove(last)

## Simple Linear Regression, Reform 1986 ####

# Find the LATE of enrolling at age 6 on CC in adulthood using a simple li
near regression.
# Use a bandwidth of three cohorts in control and treatment group.
rlms_cc1 <- rlms_cc[rlms_cc$H6 < 1983 & rlms_cc$H6 > 1976, ]
lm <- lm(cc ~ x + H6 + age + H5 + REGION + STATUS, data = rlms_cc1)
summary(lm)
#' A simple linear regression indicate that treatment has a significant ef
fect leading to
#' higher probability of acquiring chronic conditions.
#' LATE estimate = -0.024* with standard error = 0.012.
#' In addition, some control variables are significantly correlated with c
hronic conditions.
#' Women are correlated with higher probability of CC.
#' Higher ages are correlated with higher probability of CC.
#' A more rural settlement-type is correlated with lower probability of CC
.
#' People born in 1980-1982 are suspected to acquire more chronic conditio
ns in
#' adulthood than those born in 1977-1979.

### Sharp RD, Reform 1986 ####

# Bandwidth 3.
rd <- RDestimate(formula = cc ~ H6 | age + H5 + REGION + STATUS,
                 data = rlms_cc, cutpoint = 1979.5, bw = 3)
summary(rd)
#' LATE estimate = -0.025* with standard error = 0.013.
#' Treatment significantly increases the probability of acquiring CC at 10
%.

# Bandwidth 2.
```

```r
rd2 <- RDestimate(formula = cc ~ H6 | age + H5 + REGION + STATUS,
                  data = rlms_cc, cutpoint = 1979.5, bw = 2)
summary(rd2)
#' LATE estimate = -0.025 with standard error = 0.016.
#' No discontinuity.

# Bandwidth 4.
rd4 <- RDestimate(formula = cc ~ H6 | age + H5 + REGION + STATUS,
                  data = rlms_cc, cutpoint = 1979.5, bw = 4)
summary(rd4)
#' LATE estimate = -0.020* with standard error = 0.011.
#' Treatment significantly increases the probability of acquiring CC at 10
%.

# Plot discontinuity
plot(rd, range = c(1977,1982))
abline(v = 1979.5)

### Sharp RD, Placebo Outcomes ####

# Age
rd.age <- RDestimate(formula = age ~ H6 | H5 + REGION + STATUS,
                     data = rlms_cc, cutpoint = 1979.5, bw = 3)
summary(rd.age)
#' No discontinuity on age at cutoff.

# Gender
rd.gender <- RDestimate(formula = H5 ~ H6 | age + REGION + STATUS,
                        data = rlms_cc, cutpoint = 1979.5, bw = 3)
summary(rd.gender)
#' No discontinuity on gender at cutoff.

# Region
rd.region <- RDestimate(formula = REGION ~ H6 | age + H5 + STATUS,
                        data = rlms_cc, cutpoint = 1979.5, bw = 3)
summary(rd.region)
#' No discontinuity on gender at cutoff.

# Settlement-type
rd.st <- RDestimate(formula = STATUS ~ H6 | age + H5 + REGION,
                    data = rlms_cc, cutpoint = 1979.5, bw = 3)
summary(rd.st)
#' Treatment significantly affects settlement-type.

### Sharp RD, Placebo Cutoffs ####

# Perform placebo tests to check for discontinuities at other cutoffs.
# Four years prior.
placebo1 <- RDestimate(formula = cc ~ H6 | age + H5 + REGION + STATUS,
                       data = rlms_cc, cutpoint = 1975.5, bw = 3)
summary(placebo1)
# No discontinuity or effects.

# Two year prior.
```

```
placebo2 <- RDestimate(formula = cc ~ H6 | age + H5 + REGION + STATUS,
                       data = rlms_cc, cutpoint = 1977.5, bw = 3)
summary(placebo2)
# No discontinuity or effects.

# Two years after.
placebo3 <- RDestimate(formula = cc ~ H6 | age + H5 + REGION + STATUS,
                       data = rlms_cc, cutpoint = 1981.5, bw = 3)
summary(placebo3)
# Significant improvement in health at 5%.

# Four years after.
placebo4 <- RDestimate(formula = cc ~ H6 | age + H5 + REGION + STATUS,
                       data = rlms_cc, cutpoint = 1983.5, bw = 3)
summary(placebo4)
# No discontinuity or effects.
```

Descriptive Statistics

```
# Descriptive statistics

# Clean environment.
rm(list=ls())

# Load packages.

library(plyr)
library(tidyverse)
library(mosaic)
library(dplyr)

# Load data frame.
setwd("C:/Users/Bruker/Desktop/master/data-frames")
load("RLMS_BMI.RData")
load("RLMS_CC.RData")
load("RLMS_SRH.RData")

# Isolate cohorts 1977-1982
rlms_srh <- rlms_srh[rlms_srh$H6 < 1983 & rlms_srh$H6 > 1976, ]
rlms_bmi <- rlms_bmi[rlms_bmi$H6 < 1983 & rlms_bmi$H6 > 1976, ]
rlms_cc <- rlms_cc[rlms_cc$H6 < 1983 & rlms_cc$H6 > 1976, ]

# SRH
favstats(rlms_srh$srh)
# BMI
favstats(rlms_bmi$bmi)
favstats(rlms_bmi$bmidep)
# CC
favstats(rlms_cc$cc)

# Treatment condition
favstats(rlms_cc$x)
```

```r
favstats(rlms_bmi$x)
favstats(rlms_srh$x)
# Average mean
(mean(rlms_cc$x) + mean(rlms_bmi$x) + mean(rlms_srh$x))/3
# Average SD
(sd(rlms_cc$x) + sd(rlms_bmi$x) + sd(rlms_srh$x))/3

# Age
favstats(rlms_cc$age)
favstats(rlms_bmi$age)
favstats(rlms_srh$age)
# Average mean
(mean(rlms_cc$age) + mean(rlms_bmi$age) + mean(rlms_srh$age))/3
# Average SD
(sd(rlms_cc$age) + sd(rlms_bmi$age) + sd(rlms_srh$age))/3

# Gender
favstats(rlms_cc$H5)
favstats(rlms_bmi$H5)
favstats(rlms_srh$H5)
# Average mean
(mean(rlms_cc$H5) + mean(rlms_bmi$H5) + mean(rlms_srh$H5))/3
# Average SD
(sd(rlms_cc$H5) + sd(rlms_bmi$H5) + sd(rlms_srh$H5))/3

# Settlement-type
favstats(rlms_cc$STATUS)
favstats(rlms_bmi$STATUS)
favstats(rlms_srh$STATUS)
# Average mean
(mean(rlms_cc$STATUS) + mean(rlms_bmi$STATUS) + mean(rlms_srh$STATUS))/3
# Average SD
(sd(rlms_cc$STATUS) + sd(rlms_bmi$STATUS) + sd(rlms_srh$STATUS))/3
```