The effect of education on health perception and health related behavior: evidence from the Russian Federation

Ilia Gorshkov[a], Anastasia Zharinova[a]

[a]*Moscow State University*

Abstract

We use individual data from SAGE and RLMS surveys to infer whether education has an effect on the health of Russian citizens. Both direct and indirect ways of impact are estimated by means of the variety of econometrics methods. IV regression is used to assess the former effect which implies profound changes in understanding individual's health and how health care system works. The latter way means changes in the health-related behaviour, estimates for which are deduced by OLS. Different statistical methods like Lasso regression and Generalized Additive Model with splines are used to go far beyond OLS results. The obtained results confirm the existence and significance of both ways of impact: better educated people understand their health and bodies better and also switch to a much healthier way of life.

Keywords: education, health, health-related behavior, self-reported health, direct effect, indirect effect, Lasso regression, splines, IV
JEL: C21, C26, I10, I12, I20

1. Introduction

Education is considered to be one of the elements that influence citizens' health. The effect of education has been described in many papers. The differences in life expectancy between people with different levels of education exist in many countries throughout the world: the National Vital Statistics Reports show that mortality rates for 25-64-year-old

Email addresses: ilygor84@gmail.com (Ilia Gorshkov), nastya-zharin@mail.ru (Anastasia Zharinova)

We would like to address special thanks to our associate professor *N.M. Kalmykova* for great help, patience and mentorship during the whole project

people with school education were twice as high than for those with higher degrees of education. The same results were presented in Kitagawa and Hauser's (1), Elo and Preston's (2), Kunst and Mackenbach's (3) papers. However, there are few works that econometrically assess to what extent education influences health. One of the most extensive research has been conducted by Cutler and Lleras-Muney (4). The paper shows that education changes health-related behavior (e.g. smoking, fruit consumption, physical activity), which then directly influences health. Research also shows evidence for the existence of direct effect. Snowdon (5) shows in his work that nuns with higher levels of education live longer and are less prone to Alzheimer disease, although all nuns are put under the same conditions.

There have not been written extensive papers that statistically examine the effect in the Russian Federation. The primary aim of our analysis is to find out whether the educational gradient exists in Russia. Using SAGE and RLMS databases we econometrically estimate the two ways of education influence on health, namely **direct** and **indirect**. The aforementioned Cutler and Lleras-Muney's research (4) basically assesses the **indirect** effect, which is changing the health-related behavior. We mainly base our statistical inference of the **indirect** effect on the methodology of these researchers, but we want to go beyond this and also estimate **direct** effect of health on education. We consider this to be a qualitative change in health perception and better understanding health care system. To do this, we use Self-Reported Health (SRH) as a dependent variable. Our research also utilizes sophisticated statistical methods to show the non-linearity of influence and infer how important education is for the explanation of the effect.

2. Data

2.1. Datasets

To examine the casual effect of education on Self-Reported Health (SRH) and health related behavior among Russian people we extensively use two datasets with self-reported individual data: SAGE and RLMS.

SAGE is a longitudinal study on global ageing and global health conducted by World Health Organization. This study collects data on people over 50 for various countries including Russia. For the research we use individual data from Wave 1 that started in 2007 and ended in 2010. The sample provided is nationally representative. The total number of observations in the survey is 4947, but the actual number of observations for different regressions will vary due to some respondents, who did not provide answers for the related questions. In addition to individual dataset we used household data to get additional information about an individual's financial position and a respondent's location (urban or rural). To get the information, datasets were matched using provided households IDs. This combined dataset is used to infer the effect of education on health perception (SRH) and a variety of health-related activities.

We additionally use datasets provided by the RLMS study (Russian Longitudinal Monitoring Survey) which is conducted by the National Research University Higher School of Economics, ZAO Demoscope, the Carolina Population Center, the University of North Carolina at Chapel Hill and the Institute of Sociology RAS. This nationally representative research is a series of surveys monitoring effects of reforms on health and welfare of the Russian society. We take Wave 23 of the survey which was conducted in 2014 and has a total number of 18372 observations. As with SAGE, the actual number of observations will vary due to refusals to answer some questions. We filter out respondents 45 and under to get a comparable sample. The dataset is used to infer the effect of education on health behavior and indicators such as visiting hospital for preventive reasons and check-ups, alcohol consumption and regular intakes of vitamins, BMI and obesity.

2.2. Key variables

The main independent variable in the study is *Education*, which measures the whole number of completed years of education including school, college and university studies.

For the dependent variable we use an extensive list of variables depicting health related behavior and Self-Reported Health (*SRH*). *SRH* is measured as an integer number ranging from 1 to 5 as an answer to a question "How would you rate your health today?", where 1 is "Very Bad" and 5 is "Very Good". For the health related behavior, we use variables indicating range of activities from fruit intakes to check-ups. SAGE and RLMS provide us with 17 variables describing a respondent's activities that influence health and health indicators.

For control variables we use an available list of demographic and socio-economic variables. We split these variables into three groups:

- Demographics and social controls
- Income
- Economic controls

The first group of variables includes: *Age*, *Sex*, *Depression state* and *Memory* over the last 30 days, dummy variables for *Ethnic group* and *Religious denomination* and *Urban* location. The second is a respondent's income. Unfortunately, the reported quantitative information about households income was inaccurate and turned out to be statistically insignificant. In search of a good proxy, we tried to use food expenditure. With SAGE database we came across a problem: the sample shrank significantly. For SAGE database we used a household answer to a question "Would you say your household financial situation is" with answers ranging from 1 (Very Good) to 5 (Very Bad). We create 4 dummy variables, representing

people with various financial positions. The set of dummies turned out to be statistically significant at the 1% level, giving far greater explanatory power than self-reported income. To add more confidence and explanatory power, we include more economic variables in the third group. We also add a variable indicating respondent's quintile in an income distribution of the whole sample. This variable is calculated by WHO using 21 assets owned by respondents to estimate a pure random effect model and Bayesian post-estimation. With the RLMS database we included the same set of dummies (though there it is ranged from 1 to 9, so we did four groups of people with answers: 1-3, 4-6, 7 and 8). We also used a qualitative income variable, though it didn't provide explanatory power.

For the third group we added some economic variables: *Marital status*, *Working hours* and *Working days* on a typical week, the amount of *people in the household*, *number of dependent* people, dummies indicating whether a respondent gets *health care benefits* as a part of a job, working on *more than one job* and employment for the *public sector*. For SAGE, as it has already been stated, we add *Quintile* variable. For RLMS we are not able to retrieve information about number of people in a household, the amount of working days and employment for the public sector.

We use IV method to estimate **direct** effect of education on health, so we have to find suitable instrumental variables for our research. The first two variables are *Mother education* and *Father education*, which were extensively used in academic literature on the topic. In our dataset these two indicate the highest completed level of education of parents, ranges from 0 to 6 where 0 stands for "No Formal Education" and 6 is "Completed Postgraduate Degree". The instruments are relevant because education of parents are correlated with education of their children. As for the exogeneity we believe that parents education is not connected with a child's health perception. While this position can be doubted in terms of genetics, we do not consider genetics has a strong influence on health perception.

The third instrumental variable is called *Freedom* and it is a respondent's answer to the question "How free do you think you are to express yourself with no fear of government reprisal?" The answer is numerical and ranges from 1 to 5, where 1 is "Completely Free" and 5 is "Not Free at All". We presume that higher educated people report that they are less free in terms of expressing their thoughts, so the instrument is relevant. Obviously, this freedom is not connected with health perception, so the instrument is exogenous.

3. Methodology

The primary goal of our work is to assess the impact of education on people's health. Essentially, there are **two** ways through which education influences peoples health: *direct* and *indirect*. The **direct** influence is a qualitative change in health perception and better understanding of available public health services. This means that education provides people with better cognitive abilities: they can better understand and interpret their own body

signals, use medicines and follow prescriptions properly and utilize a wide range of obtainable public health services. The **indirect** way concerns a fundamental change of a person's way of life. That implies the whole range of activities from switching to healthier diet to refraining from bad habits. We try to test a hypothesis that more educated people are more likely to engage in healthy activities on a permanent basis.

To assess the two ways of influence we need two different approaches. To estimate the **direct** effect of education on health we use a regression of *SRH* on *Education*:

$$SRH = const + \gamma * Education + \alpha * Control + \epsilon, \tag{1}$$

where *Control* is a vector of control variables. For this regression we use one set of variables. It comprises the first two aforementioned sets (demographics and income, excluding *Memory*), *Marital Status*, dummies indicating a person who consumed alcohol or tobacco at least once, dummy for a heavy drinker and dummy for a person who does moderate sports, we also use a polynomial for age.

There are two key points at this stage. Firstly, the only coefficient we are interested in is γ because with this regression we want to assess only the **direct** effect of education on health. Other ways through which education influences health are described by control variables, but at this stage we do not estimate it. Secondly, we come across an omitted variable bias problem. Essentially, there are lots of other parameters that explain *SRH*, like engagement in vitamin intakes, regular check-ups, etc. We use SAGE database for this regression, but the dataset does not provide us with the comprehensive range of variables describing respondents way of life. The omitted variables are closely connected with education: the amount of sports done, check-ups made, vitamins taken, etc. correlate with persons education. Basically, this is an **indirect** way of education influence on health. The correlation leads to endogeneity of *Education* variable and inconsistency of $\hat{\gamma}$. To satisfy the needs of these two key points and solve the raised problems, we use IV regression. For the instruments we use a set of three variables: *Mother education*, *Father education*, *Freedom*. This method will allow us to obtain a consistent estimate of γ, getting information about the **direct** effect of education on health.

There are multiple reasons why we use *SRH* as a variable describing respondent's health instead of other quantitative indicators. Firstly, state of health is difficult to quantify. The best indicators of health are probably results of medical tests, however we don't have such information. Secondly, Self-Reported Health or the way how people rate their own health is the main indicator that determines a person's behavior in the economic world: labor supply, consumption and spending, etc.

To estimate the **indirect** effect, we use OLS regressions of health related activities variables on *Education*:

$$y = const + \beta * Education + \alpha * Control + \epsilon, \tag{2}$$

where y is a health related activity variable and *Control* is a vector of control variables.

The full list of activities is reported in section 5. We have taken a comprehensive list of health-related activities performed by a person available in the data. OLS regressions will allow us to obtain β estimates, gaining understanding of how education influences health-related behavior.

4. The direct influence: self-reported health

The first part of the estimation process is devoted to the direct influence of the education. We calculated $\hat{\gamma}_{IV}$ using 2-OLS procedure. The results are reported in equation 3:

$$SRH = \underset{(0.021)}{4.54} + \underset{(0.009)}{0.036} * Education + \hat{\alpha} * Control \tag{3}$$

First of all let us discuss the statistical results. The estimated $\hat{\gamma}_{IV}$ is statistically significant and has the expected sign: education has a direct effect on health. The doubts with exogeneity of the first two instrumental variables are cleared up by formal tests. As we have more instruments than equations, the Overidentifying Restrictions Test (the Sargan test) is available. For the set of three instruments we get $P_{value} = 0.48$, which leads us to a non-rejection of H_0 hypothesis that the set of instruments is exogenous. The results stated above lead to the confirmation that the set of three instruments is valid. To add up, $F_{stat} = 130 \gg 10$ that shows that instruments are far from being weak. Finally, $P_{value} = 0.003$ for the Hausman test, which leads to the rejection of H_0 hypothesis that OLS estimates are consistent on 1% level. (Although we have to admit that the P_{value} is somehow close to the 1% border). Nevertheless, the method is passing a statistical test successfully, allowing us to move on to the interpretation.

While the estimated value shows that the influence exists, we think that it is not that strong compared to other countries. One year of education increases the SRH by roughly 0.04 points. 10 years of school education add 0.4 points, which is a rather moderate rise. Additional 5 years of university education account for 0.2 points. Summing up the figures, we get that the average educational pattern "School + University" is worth 0.6 points. As the SRH is a discrete variable, we can conclude that the aforementioned pattern increases the reported health by roughly half of the step. This effect is milder compared to other countries, however the half a point increase may sound substantial, just because it was caused by means of education only. If we consider a switch between "Good" and "Very Good" response, definitely the mind has to undergo profound changes. Education qualitative changes the perception of health and helps individuals understand available public health

services. People understand their bodies better and respond to unknown health issues faster and more intelligently, utilizing a wide range of the services and using medicines properly.

5. The indirect influence: health-related behaviour

The second part of our estimation process is devoted to inferring the casual effect of education on people's behaviour related to health. We ran 17 OLS regressions with an available list of dependent variables and 3 groups of controls. Gretl software was used to conduct estimations.

The results are reported in table 1. The table shows estimated values for $\hat{\beta}$ from equation 2 divided into three columns corresponding to different sets of control variables. SAGE database allowed us to obtain a reasonable sample with a number of observations ranging from roughly 2250 (for female specific activities) to 3600. While RLMS provides with up to twice as much observations compared to SAGE, the database has a few reliable independent variables.

In this section we will take a closer look at health-related activities. The first 5 variables describe a respondent's attitude to smoking. We can clearly see that education is highly beneficial for this type of behaviour, reducing the probability of smoking. To be more precise, let us take a look at the estimated numbers. One year of education reduces the probability of a person to be a heavy smoker by roughly 1% and lowers the amount of cigarettes smoked by 0.2 cigarettes. It will contribute to probability reduction for current smoking by approximately 0.9 %. While these interpretations may sound unconvincing, the results have to be treated from a slightly different perspective. Education has to be taken into consideration in "steps". As the sample mainly consists of 50+ year-olds, these people got an education in the USSR. They spent 5 years at the university to receive the first degree. Consequently, these 5 years have to be considered as a whole, because they form an education "step". During this "step" the student will be exposed to the positive influence of education, changes in the outlook and attitude to health. What do the numbers say about the "steps"? The Speciality degree in the USSR (5 yrs) leads to a 5% reduction in probability of smoking. The results statistically confirm the hypothesis which show a positive influence of education on giing up smoking, although the effect is not that substantial, compared to USA with a year of education accounts for a 3% probability decrease (4).

Turning to sports and fitness, the same trend is clearly visible. People will be more likely to participate in vigorous and moderate sport activities with a 1.3% probability increase in total per one year of education. Consequently, the earned degree leads to a substantial 7.5% rise. Similarly, the probability of spending a sufficient amount of time walking or cycling increases as well.

The next point to discuss is fruit consumption. The amount of fruits[1] eaten on a typical day is treated as a profound change in a diet and attitude to it. Educated people switch to a healthier diet increasing the amount of fruit eaten and reducing consumption of cheap and unhealthy nutrition. We do not take vegetables into consideration because they mostly account for potatoes in the dataset, consumption of which is not beneficial for health. Returning to numbers, we can see a small but statistically significant positive effect. The figures are roughly similar to the USA estimates (4) [2]. It is also important to mention that education does not only contribute to a healthier diet, but also encourages people to take more vitamins, with a 5 year step of education accounting for a 4.5% increase in probability.

Education also increases awareness of such serious diseases as cancer. This includes regular check-ups and having gender specific examinations. The results show that better educated people are more likely to visit a doctor to check their health. The "step" of education discussed above accounts for a 5% rise in probability of having a mammogram and a 2.5% increase for regular check-ups. In other words, educated people are more serious and responsible for their health, devoting more effort to check their bodies properly and regularly.

There is also a hypothesis that educated people tend to be fitter. We test this statement using *BMI* as an independent variable. Although the hypothesis was not rejected statistically, the effect turned out to be moderate. The 5 years of education result in a 3% decrease in probability of being obese. *BMI* tends to decrease for higher educated people by rougly 0.3 points for 5 years of education.

The last dependent variable is a dummy indicating a person who drank vodka during the last 30 days. The dataset provided comprehensive information on drinking patterns for various types of alchocol such as wine, beer, vodka, etc. It was decided to take vodka as the only type of alcohol into consideration because of the reason that we call "quality switching". Consider two different people: one with a school education, another with a PhD degree. It is not reasonable to believe that people with more years of education will refrain from drinking alcohol. In fact, better educated people change the type of alcohol they consume to a more expensive and qualitative one. So a more realistic assumption will be that better educated people will switch from inexpensive drinks (e.g. vodka) to more costly ones (e.g. expensive wine). We consider that vodka can be a good representative of a cheap type of alcohol. If our hypothesis is true, then we should find that probability of drinking this type of alcohol will decrease with the years of education. Unfortunately, we can use only dummy variables for alcohol because the quantitative data measuring the amount drank in mL is not reliable: people refused to tell it or simply reported inaccurate and rounded amounts. Turning to the estimated figures, we can clearly see a beneficial influence. 5 years of education reduce probability of drinking vodka by 5%.

[1] Banana, mango, apple, orange, papaya, tangerine, grapefruit, peach, pear

[2] The research showed $\hat{\beta} = 0.067$ for fruits **and** vegetables, while our research has $\hat{\beta} = 0.023$ for fruits only

Variable	$\mathbf{N}_{obs}$	**Demographic and social controls**	**+ Income controls**	**+ Income and other economic controls**
SAGE				
Ever smoked tobacco	3593	−0.0070* (0.0019)	−0.0064* (0.0019)	−0.0071* (0.009)
Currently smoking	3590	−0.0097* (0.0017)	−0.0090* (0.0017)	−0.0085* (0.0017)
Currently smoking daily	3590	−0.0106* (0.0016)	−0.0099* (0.0016)	−0.0094* (0.0016)
Average amount of cigarettes smoked a day	3512	−0.2044* (0.0314)	−0.1891* (0.0314)	−0.1757* (0.0322)
Smoked formerly	3588	−0.0074* (0.0018)	−0.0067* (0.0018)	−0.0067* (0.0018)
Amount of fruits eaten on an average day	3075	0.0291* (0.0070)	0.0250* (0.0070)	0.0230* (0.0072)
Walking or riding a bicycle for at least 10 min	3596	0.0121* (0.0023)	0.0124* (0.0023)	0.0101* (0.0024)
Hours spent on walking or riding a bicycle	3580	0.0301* (0.0061)	0.0313* (0.0061)	0.0291* (0.0064)
Doing vigorous sports	3596	0.0033* (0.0001)	0.0032* (0.0001)	0.0028* (0.0001)
Doing moderate sports	3596	0.0106* (0.0015)	0.0102* (0.0016)	0.0089* (0.0015)
Had pelvic examination during last 2 years	2267	0.0077* (0.0003)	0.0073′ (0.0030)	0.0055 (0.0030)
Ever had mammogram	2322	0.0137* (0.0031)	0.0129* (0.0031)	0.0093* (0.0032)
RLMS				
Had check-ups during last 3 months	5918	0.0076* (0.0016)	0.0052* (0.0016)	0.0050* (0.0017)
BMI	5688	−0.0470′ (0.0225)	−0.0481′ (0.0230)	−0.0558′ (0.0231)
Obesity (BMI>30)	5688	−0.0054* (0.0021)	−0.0050′ (0.0021)	−0.0056* (0.0021)
Vitamin intakes during last 30 days	5918	0.0114* (0.0017)	0.0090* (0.0017)	0.0088* (0.0017)
Drank vodka during last 30 days	2496	−0.0101* (0.0034)	−0.0097* (0.0035)	−0.0109* (0.0035)

Table 1: The OLS estimates of the effect of education on various activities and indicators
* - significant at the 1% level, ′ - significant at 5% level, HC1 standard errors are reported in brackets

6. Nonlinear effect of education on health-related behaviour

In section 5 we give evidence that education has an indirect effect on health that channels through changes in people's behaviour. However, we do not actually know whether the influence is linear or not. To infer this from the data, we decided to use a Generalized Additive Model (GAM). To move beyond linearity we can use **polynomials**, **step functions** and, finally, **splines**. Splines are the extensions of the first two functions, as splines have high variance when predictors take extremely large or low values. That means that splines have higher variance at the boundaries. To solve the issue we chose a **natural spline**, because it has special boundary constraints, allowing it to produce more stable estimates at the boundaries. We could use both 10-fold Cross-Validation and AIC criterion to choose the number of knots for the splines. We used *GAM* package for R to conduct estimations. We also use a polynomial for *age* to address any possible issues with possible non-linearity.

We use several GAM models for some of the activities listed in table 1. Splines are statistically significant at a 1% level in every model used. Before interpreting the graphs we should familiarize the reader with the USSR system of education. There were three basic steps:

1. Compulsory school education (8 yrs)
2. Optional school education (2 yrs)
3. University (5 yrs)

We marked these steps in every graph. Additionally, every graph shows twice-standard-error lines for illustration of the 95% confidence intervals. As the majority of the respondents finished their compulsory school studies, the most interesting part of the graphs is 8-20 completed years of education. Although we have some respondents with 20+ yrs of education, they account for only a small proportion in the data, causing higher variance and standard errors at the right boundary. The same is true about the people who did not finish their school. The left boundary is also subject to higher variance. It is also noticeable that variables of the same type of activity (e.g smoking or cycling) have roughly the same curvature.

Firstly, the graphs show that most of the time education has a positive effect. Although, some splines show a negative effect, it occurs only on the boundaries where the confidence interval is much broader, so it can not be interpreted clearly. Practically every spline indicates that university education has a greater effect, which is exhibited by steeper curves with a higher absolute value of the derivative. This is mostly noticeable in figures 1-4. The effect starts from the compulsory school, then it is accompanied by a moderate increase during the next 2 years of education. The next 5 years of the university accounts for a bigger influence.

However, a different picture is seen in Figures 5 - 7. Let us start with the Figure 5. The spline shows that school doesn't create any incentives to do moderate sports. Conversely, further education exhibits great influence. We have to mention that the effect is

not immediate, it takes some years spent at the university to start. This is an important point that we will discuss later. How can we explain such a great significance of a higher education? This may be true because of the abundance of opportunities. Most universities have a developed sports system with various kinds of sports available. This allows people to get a better understanding of the types of existent physical exercises and find the most suitable one. Positive and responsible attitude to sport is also developed, that results in a higher probability for a person to do sports after the graduation.

Now let us turn to Figure 6. ompulsory education starts to influence an individual. The following two years of the school education support the effect. However, the effect stops for the next 2 - 3 years of education and then suddenly starts to rise again and continues after 15 completed years. We came up with an explanation for the paradoxical spline. We believe that the clue is in steps again. First of all, let us recall that spline illustrates the influence of education on fruit consumption. We have already mentioned, that fruit consumption is a good proxy for the profound change in the attitude to a diet. The attitude is a delicate thing and does not change suddenly. The healthy attitude to the diet and life starts to form during the years of compulsory education which can be considered as the first mark in the mind of the individual. We move on to university education. We strongly believe that university education yields fully only when it is finished. The university is not only knowledge, it is culture. It forms a new mindset but the change is finished **only** when a person is fully exposed to the variety of activities and subjects, resulting in endless amounts of impressions, feelings and some cultural shocks. Only when the person goes through the aforementioned, his/her mind will change attitude to everything to the world and to their health in particular. This explains why the unfinished education doesn't affect a person in terms of profound changes in a diet. Education starts to yield Only to the close of the educational step, which is clearly seen on the graph. As it was already noted, the same is true about Graph 5 and Graph 7. All in all, the graphs clearly show that the unfinished education does not yield as much, as the finished one.

To conclude, the splines confirm the positive effect of education on health-related behaviour. They also show that the relationship is not linear: university education appears to have a stronger effect than school education. Finally, some graphs provide a fascinating insight into the difference between the influence of finished and unfinished university education.

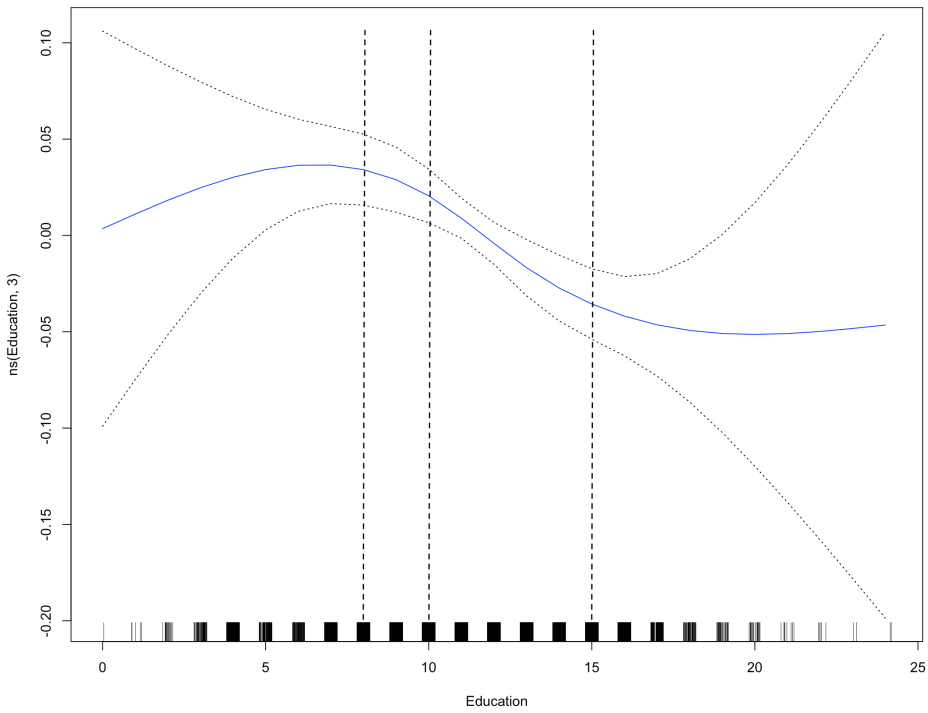

Figure 1: Ever smoked tobacco, df=3

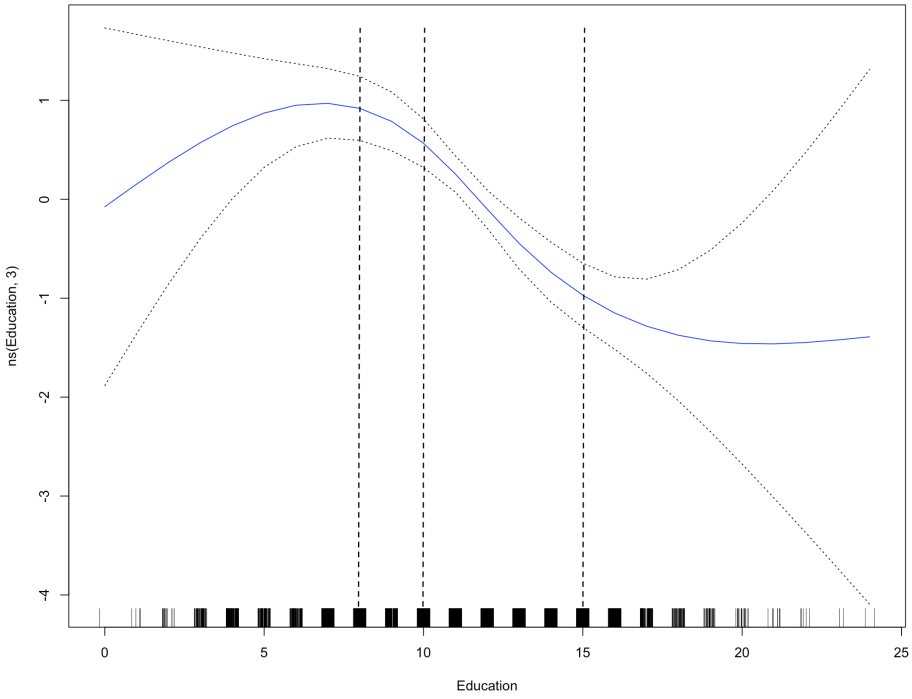

Figure 2: Amount of cigarettes smoked, df=3

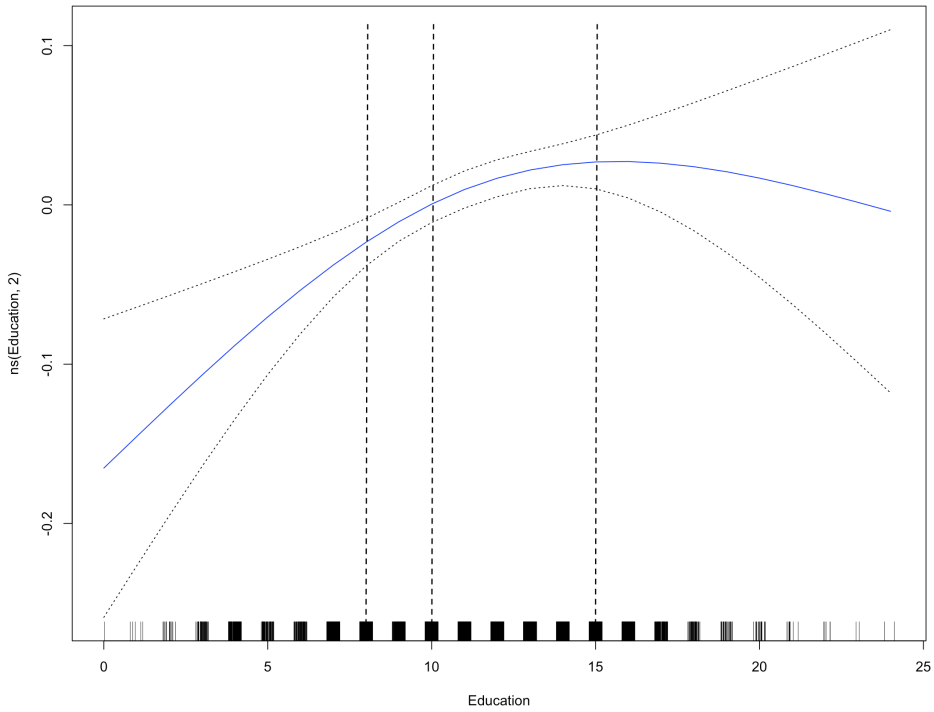

Figure 3: Walking or riding a bicycle for at least 10 minutes, df=2

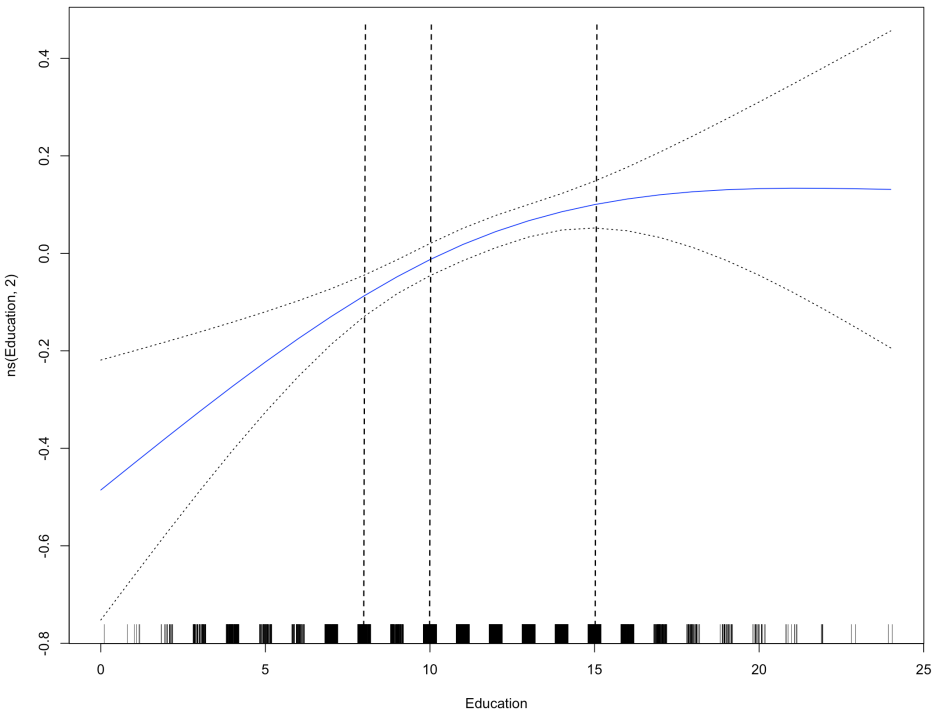

Figure 4: Time spent on walking or riding a bicycle, df=2

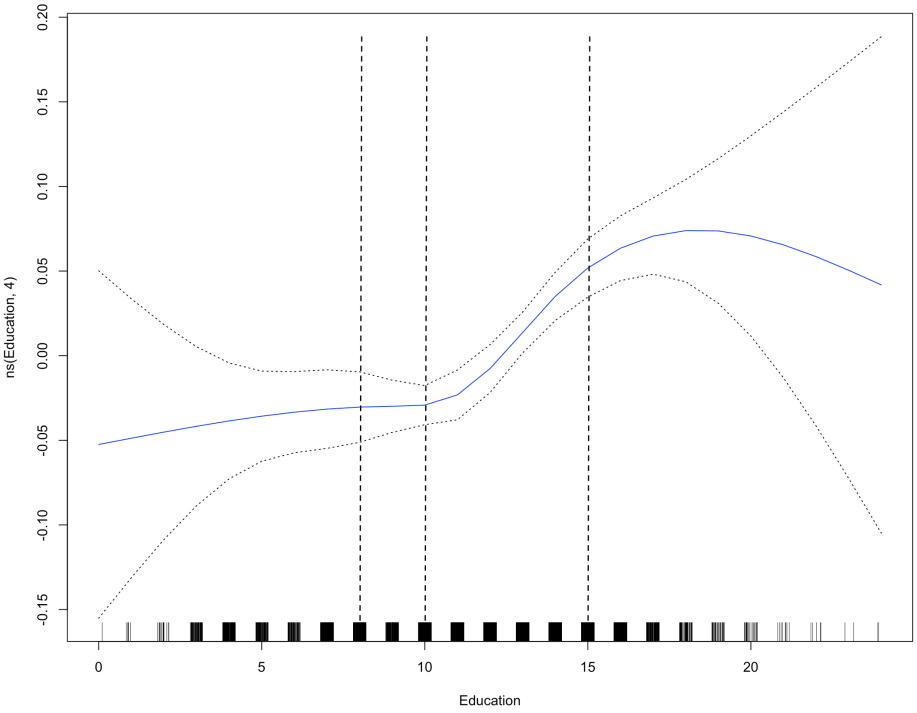

Figure 5: Doing moderate sports, df=4

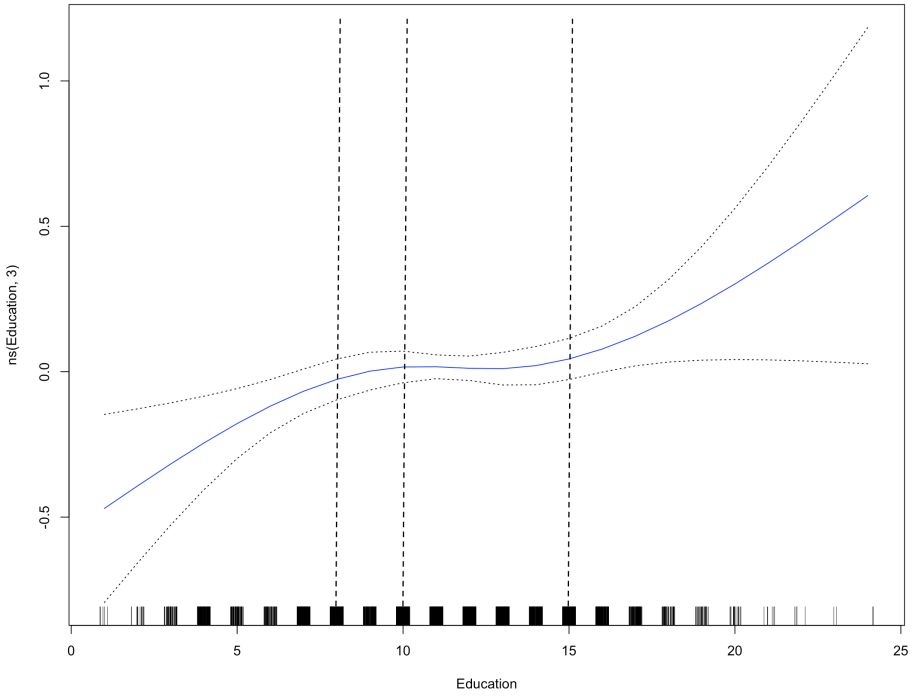

Figure 6: The amount of fruit consumed, df=3

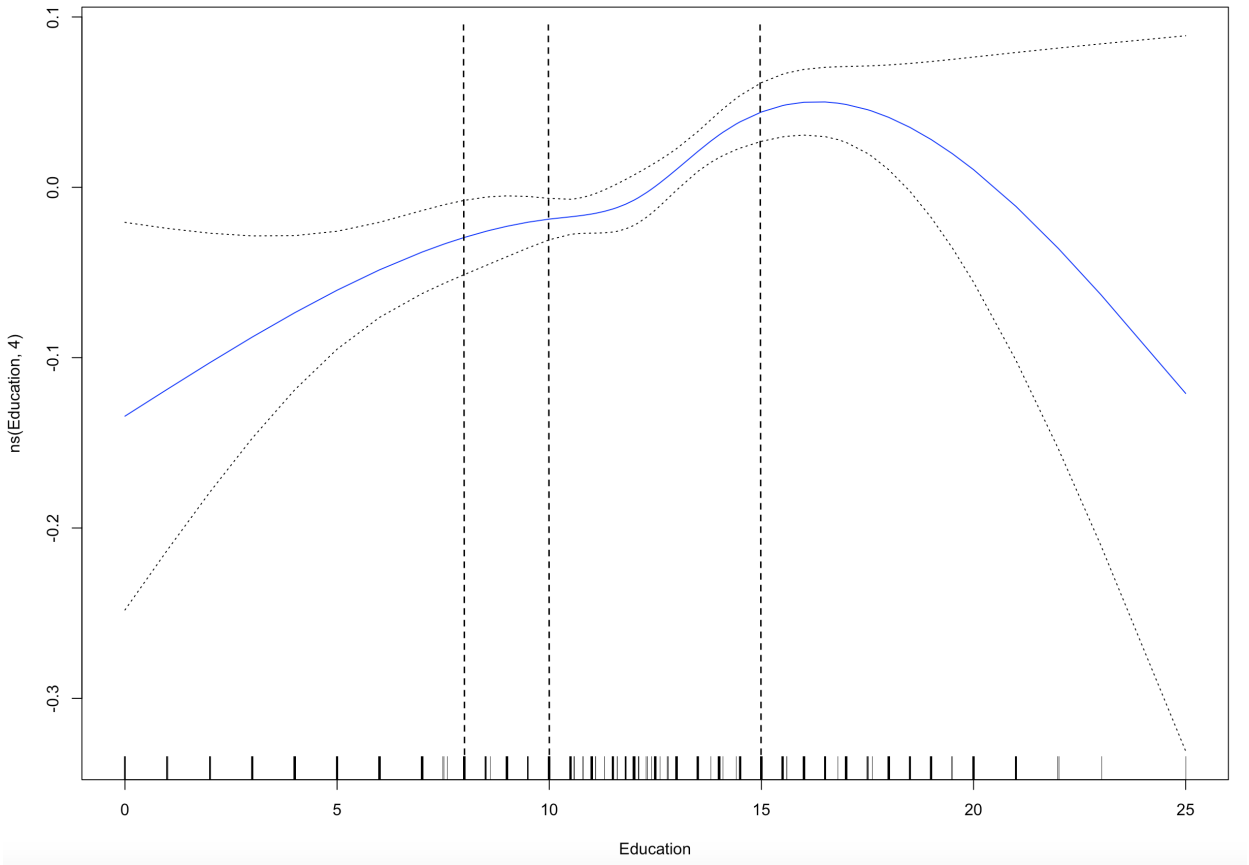

Figure 7: Vitamin intakes during the last 30 days, df=4

7. Robustness checks

In this section we do some robustness checks for the OLS estimates. We also infer to what extent education actually influences health. This means that we rank variables by their importance and then check the *Education* variable. To achieve the aforesaid, we applied one of the modern alternatives to the OLS regression. We use the shrinkage method model called *Lasso*. The method allows to fit the model using all available predictors, however it estimates unnecessary ones exactly to zero. The model is more interpretable than OLS, it also significantly reduces the variance (although it is also connected with a small bias) and performs a variable selection.

First of all, we estimated the Lasso models for all 17 activities that we mentioned above. To do this, we used an R package *glmnet*. We conducted the 10-fold cross-validation to find the most suitable shrinkage parameter λ for every model. The results of the estimation process are reported in Table 2.

Variable	$\mathbf{N}_{obs}$	$\hat{\beta}_{OLS}$	$\hat{\beta}_{Lasso}$	**Rank**
SAGE				
Ever smoked tobacco	3593	-0.0071^* (0.009)	-0.0070 (λ=0.00015)	8
Currently smoking	3590	-0.0085^* (0.0017)	-0.0079 (λ=0.00096)	5
Currently smoking daily	3590	-0.0094^* (0.0016)	-0.0089 (λ=0.00108)	5
Average amount of cigarettes smoked a day	3512	-0.1757^* (0.0322)	-0.1707 (λ=0.01401)	5
Smoked formerly	3588	-0.0067^* (0.0018)	-0.0063 (λ=0.00052)	8
Amount of fruits eaten on an average day	3075	0.0230^* (0.0072)	0.0230 (λ=0.00187)	2
Walking or riding a bicycle for at least 10 min	3596	0.0101^* (0.0024)	0.0098 (λ=0.00213)	2
Hours spent on walking or riding a bicycle	3580	0.0291^* (0.0064)	0.0276 (λ=0.00463)	2
Doing vigorous sports	3596	0.0028^* (0.0001)	0.0027 (λ=0.00074)	2
Doing moderate sports	3596	0.0089^* (0.0015)	0.0089 (λ=0.00370)	1
Had pelvic examination during the last 2 years	2267	0.0055 (0.0030)	0.0047 (λ=0.00455)	2
Ever had mammogram	2322	0.0093^* (0.0032)	0.0097 (λ=0.00572)	1
RLMS				
Had check-ups during the last 3 months	5918	0.0050^* (0.0017)	0.0049 (λ=0.00191)	2
BMI	5688	$-0.0558'$ (0.0231)	-0.0355 (λ=0.04007)	5
Obesity (BMI>30)	5688	-0.0056^* (0.0021)	-0.0028 (λ=0.00175)	7
Vitamin intakes during the last 30 days	5918	0.0088^* (0.0017)	0.0085 (λ=0.00135)	4
Drank vodka during the last 30 days	2496	-0.0109^* (0.0035)	-0.0075 (λ=0.00764)	2

Table 2: Lasso estimates for the health-related activities and indicators
* - significant at the 1% level, HC1 standard errors are reported in brackets for OLS and shrinkage parameter for Lasso, rank 1 indicates the most important variable according to Lasso

As we can see the Lasso model did not estimate the coefficient β for the *Education* variable to be zero in any model. This confirms that education is important for the explanation of the health-related behaviour. Additionally, the estimated $\hat{\beta}_{Lasso}$ figures are practically equal to the $\hat{\beta}_{OLS}$. The results reassure us that the effects mentioned in Section 5 are true.

The next step is to rank the variables. We do this with the help of the profile plots and by examining the order in which the dependent variables turn zero with the increase of the shrinkage parameter λ. The ranks are reported in Table 2. The *Education* variable is highly-ranked for most of the activities and indicators, underlining the importance of education. However, the variable had a 7-8 rank for some of them. Not less important is the fact that these activities indicate the probability of the past events (e.g. former smoking) where education has a less serious effect, although it is still significant.

The same is noticeable on the profile plots for coefficient paths against λ sequence. Figures 8-10 show how the coefficients will change, when the shrinkage parameter λ is tuned. The variables, that turn zero, are to be considered the least important. Consequently, the ones that are zero only on the right side of the graph are the most useful for the explanation. The graphs show that *Education* variable is of prime important, it is ranked 1-4 by the Lasso model.

The Lasso model confirms that education has a statistically significant effect on the health-related behaviour. It also shows that for most of the cases it has a considerable and dominant influence. The model supports and extends the conclusions obtained from the OLS model.

References

[1] E. M. Kitagawa, P. M. Hauser, Differential mortality in the United States, Harvard University Press, 1973.

[2] I. T. Elo, S. H. Preston, Educational differentials in mortality: United States, 1979-1985, Social Science & Medicine 42 (1) (1996) 47–57. doi:10.1016/0277-9536(95)00062-3.

[3] A. E. Kunst, J. P. Mackenbach, The size of mortality differences associated with educational level in nine industrialized countries., Am J Public Health 84 (6) (1994) 932–937. doi:10.2105/ajph.84.6.932.

[4] D. M. Cutler, A. Lleras-Muney, Understanding differences in health behaviors by education, Journal of Health Economics 29 (1) (2010) 1–28. doi:10.1016/j.jhealeco.2009.10.003.

[5] D. Snowdon, Aging with grace, Bantam Books, 2001.

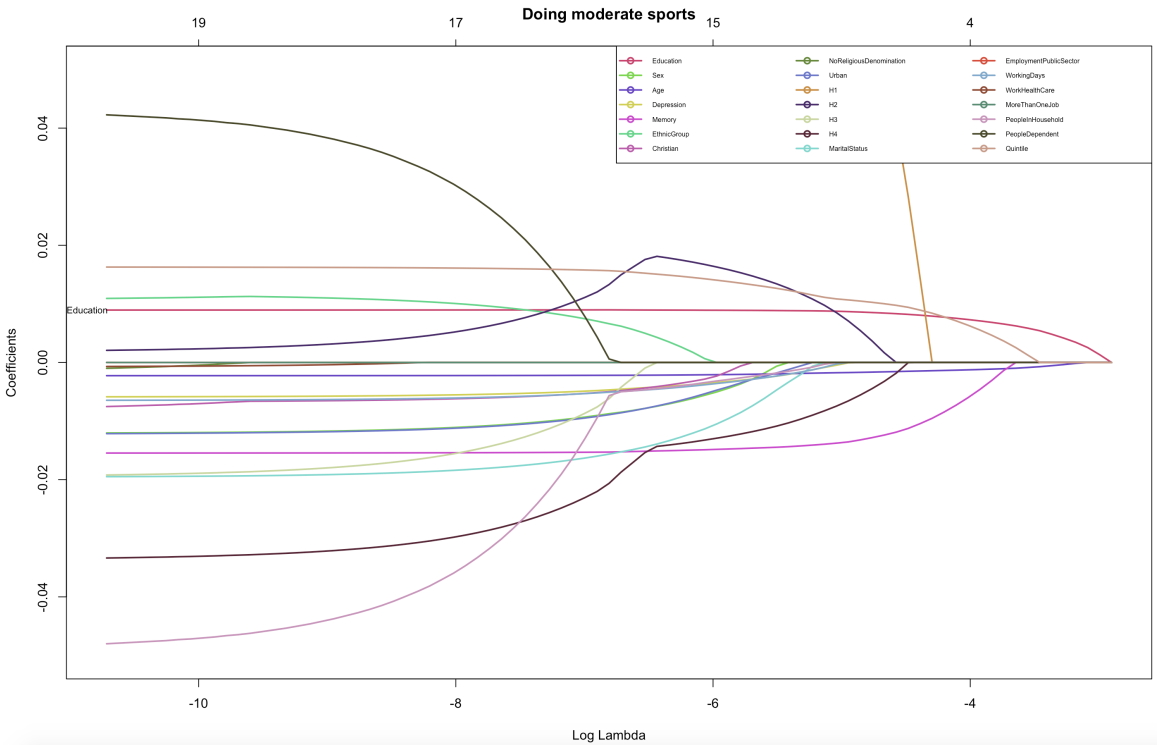

Figure 8: Profile plot for coefficient paths against the log-lambda sequence: doing moderate sports

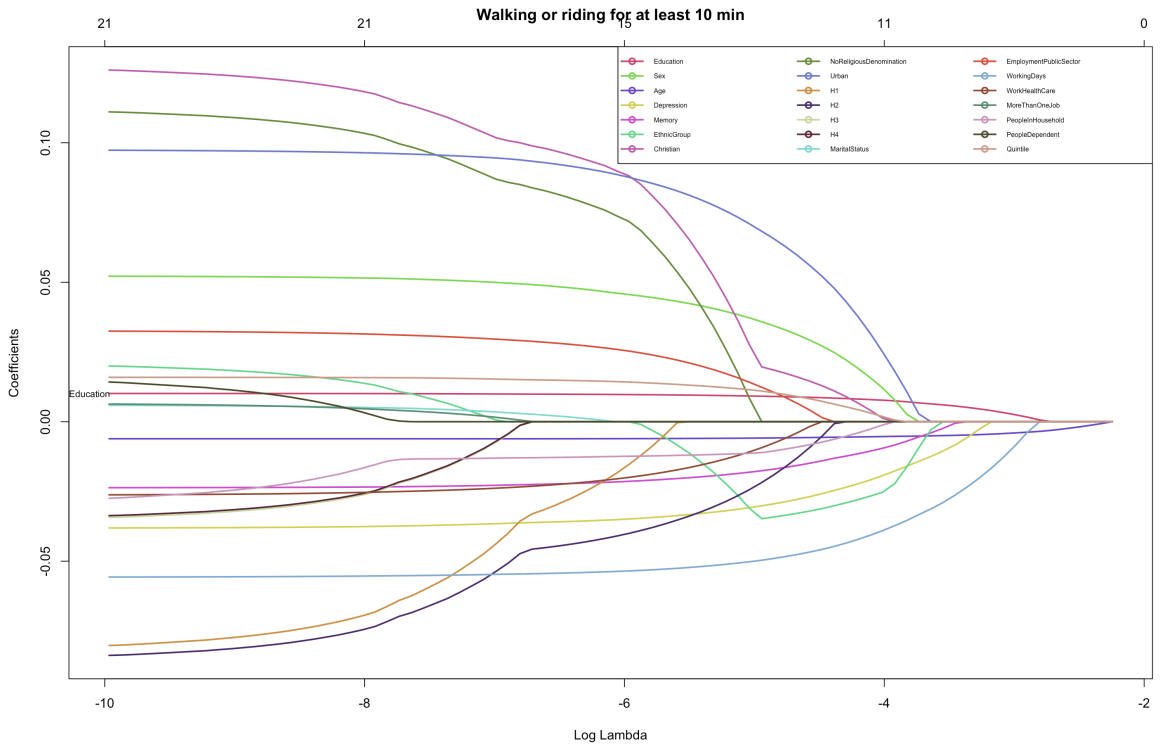

Figure 9: Profile plot for coefficient paths against the log-lambda sequence: walking or cycling (10+ min)

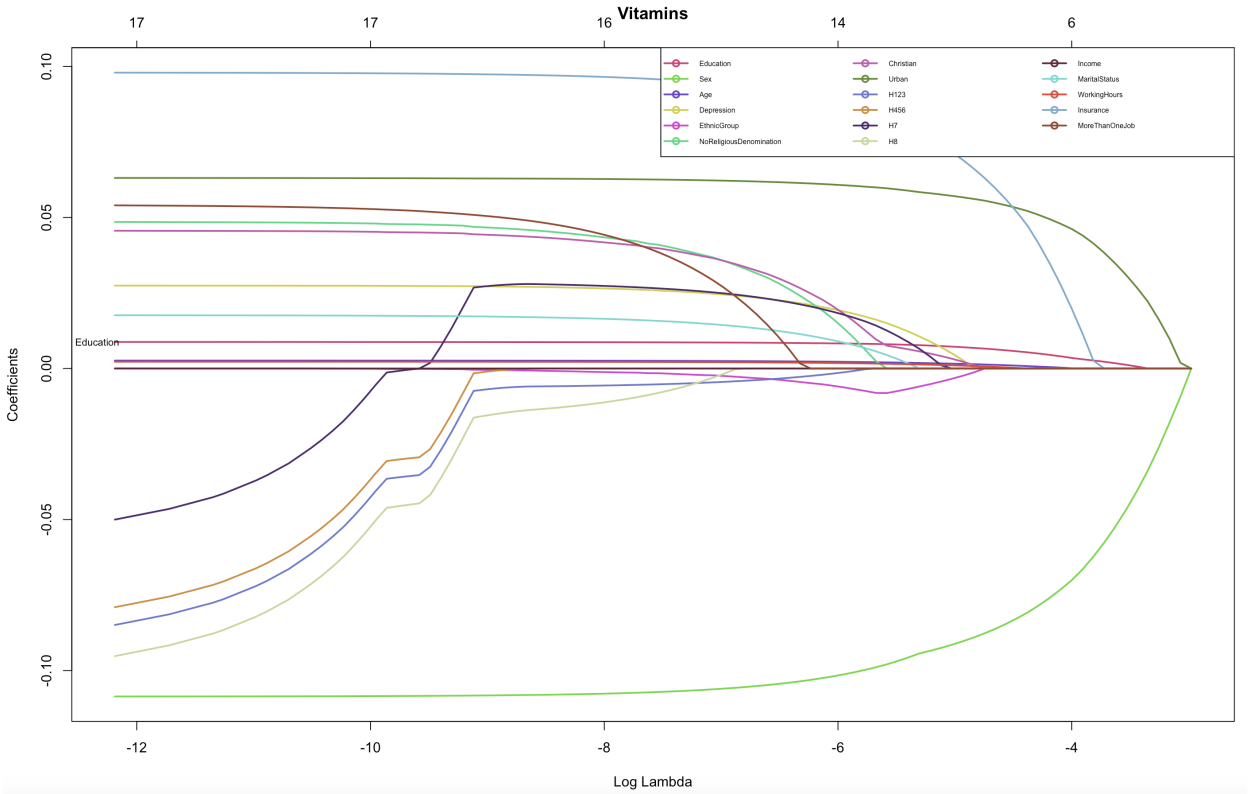

Figure 10: Profile plot for coefficient paths against the log-lambda sequence: taking vitamins during the last 30 days