DA3_3

Tamas Koncz

2017 december 8

1. Filter data: Keep respondents between 50 and 80 years of age.

The variables you will need are whether the person deceased within 6 years of the interview (“deceased”), gender (“female”), age (“age”), years of education (“eduyears_mod”), income group within country (“income10g), and the explanatory variables of your focus, physical activities (variable “sports”: 1: more than once a week, 2: once a week, 3: one to three times a month, 4: hardly ever, or never).

```
dt <- fread('mortality_oldage_eu.csv')
dt <- dt[age >= 50 & age <= 80, ]
dt <- dt[, c("deceased", "female", "age", "eduyears_mod", "income10g", "sports")]
```

2. Do exploratoty analysis: Create binary variables from the sports variable. Describe these variables in your dataset. Drop observations that have missing value for either.

The conversion to binary variables was done with the model.matrix R-function:

```
m <- model.matrix(~ -1 + deceased + female + age + eduyears_mod + income10g + spor
ts + factor(sports),
                  data = dt)
```

First, let’s look at the frequencies for our LHS variable, “deceased”. With the help of a crosstable, we will cross-reference it with another variable, “female”, representing genders in the sample.

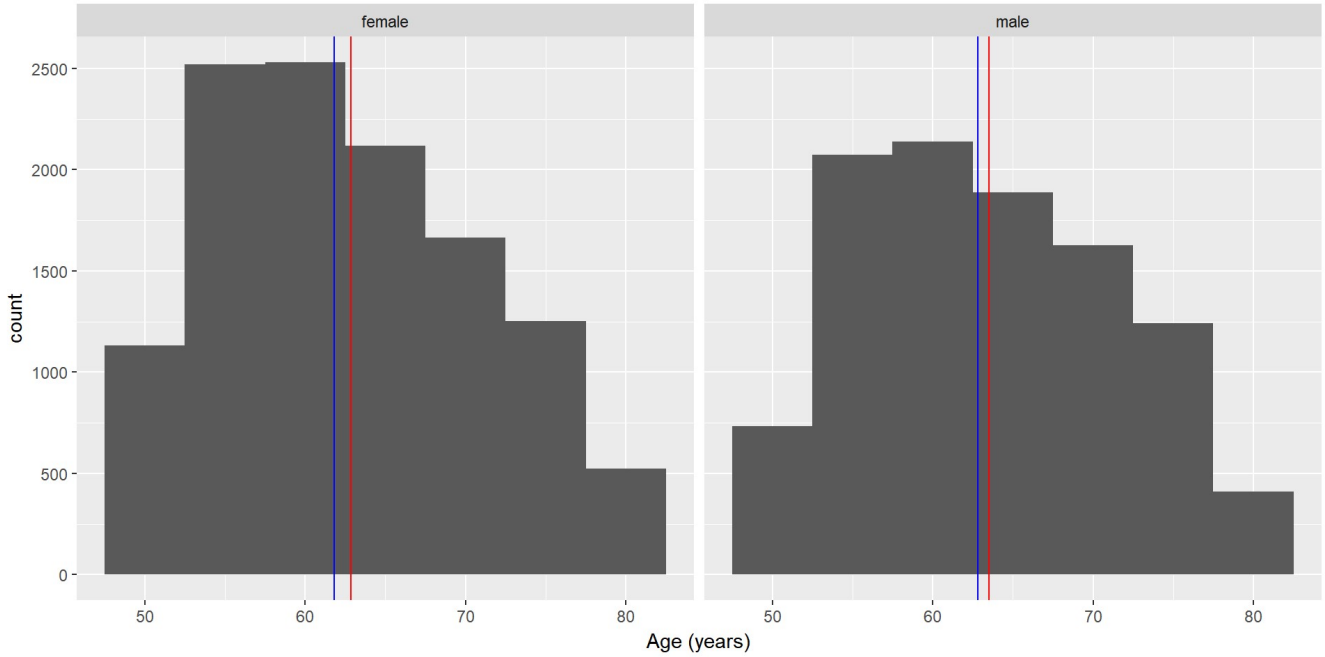
	Female		
Deceased	**0**	**1**	**Total**
0			
N	9470	11481	20951
Row(%)	45.2%	54.8%	94.2%
1			
N	786	498	1284
Row(%)	61.2%	38.8%	5.8%
Total	10256	11979	22235

This dataset is not balanced:
1. Deceased only consitute 6% of our sample
2. There are almost 10 percantage points more females than males
3. Males have died with a significantly larger frequency than females

The above should not be a problem for the exercises below (neither do I explore the reasons), but it is

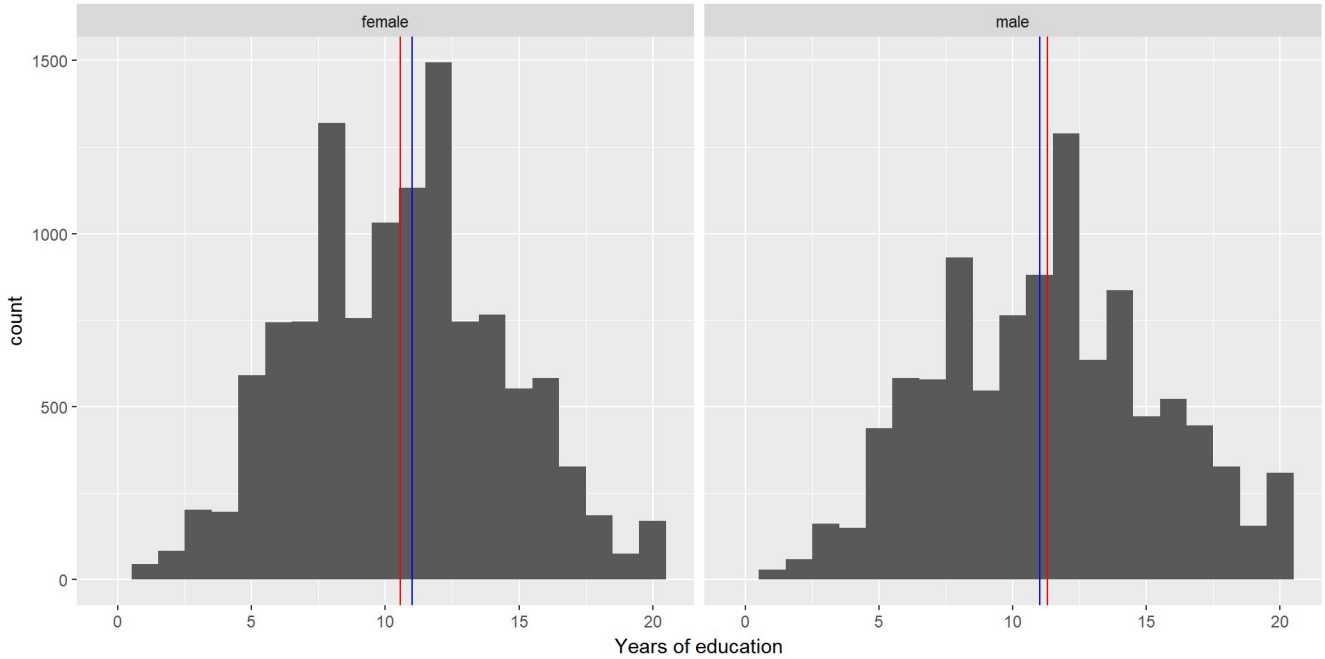
worth to keep in mind.

Now let's consider the below distribution plots (red line - mean, blue line - median):



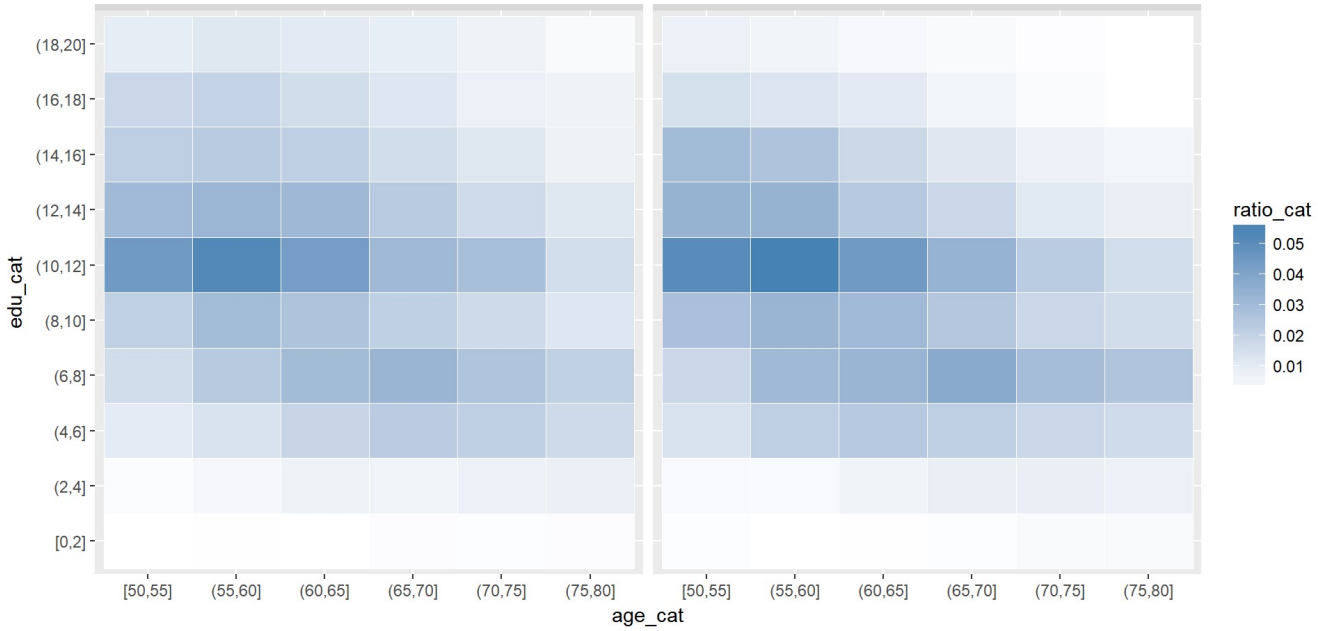
Age is very similarly (basically identically) distributed for both males and females. The distributions are not normal, but rather right-skewed. Visually they remind of the lognormal distribution, hence a log transformation can be considered for enhancing the regression's fit.

Next, the distribution of years spend in education, visuallized in the same way:



The distribution of education years is also similar between genders, however we can observe males having spent more time in higher education slightly more frequently. The distributions resemble that of the bell curve, however there are two apparent differences: most people finish their education after a certain amount of years (8-12-etc.), hence we can see these spikes on the bar chart as well, while there is a small uptake with people with 20 years - PhDs. (Note that the average person has just finished high school in this sample.)

It is also interesting to examine the joint-distribtion of age and education:



The above heatmap gives us a good indication of the share of different age - education 'groups', showing share %-s for genders separately. For both genders, most people are concentrated around ~12years of education for all age groups - older people are generally somewhat less educated than the younger groups. Also, a larger portion of males have spent 16+ years in education than females, both their share is very small for both genders, specially among older people.

There are two people with 0 years of education, something that we could consider exterme values. However, I have decided to keep them in the sample for two reasons: first, there is not enough evidance that this would be due to data quality issues. Secondly, they were not significantly influential on the regression results.

deceased	gender	eduyears_mod	income10g	sports
0	female	0	10	2
0	male	0	5	2

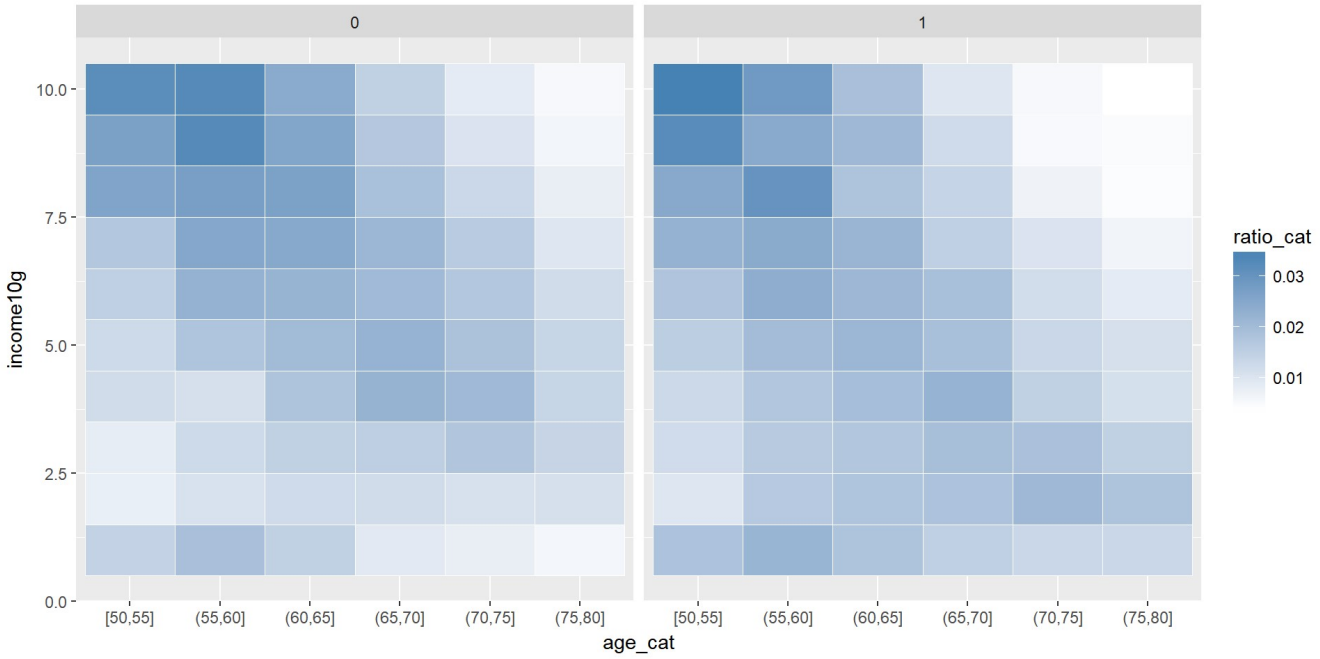
Lastly, let's take a glimpse at the correlation matrix for our potential RHS variables (excl. gender):

	age	eduyears_mod	income10g	sports
age	1	-0.24	-0.23	0.21
eduyears_mod	-0.24	1	0.28	-0.1
income10g	-0.23	0.28	1	-0.1
sports	0.21	-0.1	-0.1	1

(Note1: the correlation coefficient for sports should not be interpreted at face value, as the variable is categorical. However, given it's ordered, we can take some very cautious conclusions from it. Note2: the same is not a worry for income, as the 'income10' is treated variable as continous for the sake of this exercsie)

The first observation we can make is that older people tend to do less sports in general (not suprisingly). Income and education are positively correlated, while both are negatively correlated with age, also in line with expectations. Most correlation coefficients are significant, but none are high in absolute terms - part of this can be attributed to non-linear relationships in our data.

Relationship between income and age visualized:



Note the decrease with age, as well as the more even distribution among women.

2.1 Estimate a linear probability model (LPM) of mortality on sports. Report and interpret the results.

```
lpm <- lm(deceased ~ `factor(sports)1` + `factor(sports)2` + `factor(sports)3`,dat
a = dt)
dt$deceased_pred <- predict.lm(lpm)
```

Before interpreting model results, let's remind ourselves that the unconditional probability of someone dieing in the six years, based on the dataset is 5.7%:

```
sum(dt$deceased) / dt[,.N]
```

```
## [1] 0.05735079
```

Now, onto the model coefficients:

	Dependent variable:
	deceased
`factor(sports)1`	-0.059***
	(0.004)
`factor(sports)2`	-0.059***
	(0.005)
`factor(sports)3`	-0.042***
	(0.006)
Constant	0.092***
	(0.003)
Observations	21,848
R^2	0.014
Adjusted R^2	0.014
Residual Std. Error	0.231 (df = 21844)
F Statistic	106.030*** (df = 3; 21844)

Note: *p<0.1;* ***p<0.05;*** p<0.01

People who “hardly ever, or never” do sports die within 6 years with 9.2% chance People who do sports “one to three times a month” die with 4.2% less chance compared to people in the 4th (previous) group. They die with a 5% likeliness Peope who exercise even more die with 5.9% less chance compared to people in the first category (there were no significant difference between group 1 and 2). They die with a 3.3% likelihood Based on the p values, all coefficients are significant on the usual levels

People who do more sports are a lot less likely to die based on the dataset. Just based on this we cannot say however that doing sports make them “healthier” (clarification: but I do consider not dying healthier). We *only* know that there is a correlation in the dataset. However, it can be that even if the correlation is present in the general population as well, there is no causal relationship. Example: it could be that healthier people (who die less likely) are fit to more sports in general. Or that older people do less sport on average.

We are not interpreting the R2 for the model.

3. If you are interested in the causal effect of doing sports on mortality, would you want to control for some of the other variables in the dataset to get closer to the causal effect you are after?

Yes - including control variables could help us separate the relationship between just mortality and sports, by “filtering out the noise” of potential confounders.

Focusing on the variable groups with most impact, for better understanding I would enhance the model with: 1. Variables about lifestyle choices: drinking, smoking, and if they could be made available, then variables describing work conditions, stress levels, etc. 2. Variables about initial state-of-health: obesity, self-rated health, chronic conditions, etc.

I expect the above to play a very significant role in mortality rates. There are some other variables that could be influential as well, maybe to a lesser extent: like ones describing family status, as psychological state “dummies”, or countries to account for differences in health systems.

Would that controlling get you the causal effect you are after?

Partly, yes. By controlling for other variables, we can narrow down the interference between age and sports, separated from other impacts. In this case, the direction of causality is obvious, given the type and meaning of the “deceased” variable. However, it is very hard to measure the true impact of sports on mortality, as there can be omitted variable bias that are basically impossible to know in a large dataset. Human health is a very complicated question, with a practically unlimited set of influencing factors - factors that can influence human behavior (in this case, doing sports) just as much as the likeliness of dying. So, keeping it short, we might gain a better understanding by the introduction of control variables, however it could be very challenging to include all the relevant ones, something to be conscious about.

4. Control for those variables in another LPM, interpret its results on sports, and compare those to the previous regression estimates. Discuss the differences and similarities.

First let’s enhance the model with the “age” variable (transforming age to years above 50 for easier interpretation of results), which could account as a proxy for general health-state:

	Dependent variable:
	deceased

	(1)	(2)
factor(sports)1	-0.059***	-0.040***
	(0.004)	(0.004)
factor(sports)2	-0.059***	-0.046***
	(0.005)	(0.005)
factor(sports)3	-0.042***	-0.035***
	(0.006)	(0.005)
age_diff		0.005***
		(0.0002)
Constant	0.092***	0.016***
	(0.003)	(0.004)
Observations	21,848	21,848
R^2	0.014	0.042
Adjusted R^2	0.014	0.042
Residual Std. Error	0.231 (df = 21844)	0.228 (df = 21843)
F Statistic	106.030*** (df = 3; 21844)	242.083*** (df = 4; 21843)
Note:		*p<0.1;* ***p<0.05;*** p<0.01

We can easily observe that controlling for age seems to be a good decision. The coefficients are showing similar effects (actually, there is much overlap among 95% CIs), but the impact of sports have decreased on the LHS variable, with age accounting for parts of it.

Being conscious about ink space, I am jumping to the final model I deemed best:

	Dependent variable:		
	deceased		
	(1)	(2)	(3)
factor(sports)1	-0.059***	-0.040***	-0.044***
	(0.004)	(0.004)	(0.004)
factor(sports)2	-0.059***	-0.046***	-0.045***
	(0.005)	(0.005)	(0.005)
factor(sports)3	-0.042***	-0.035***	-0.037***
	(0.006)	(0.005)	(0.005)
age_diff		0.005***	0.005***
		(0.0002)	(0.0002)
female			-0.037***
			(0.003)
eduyears_mod			-0.002***
			(0.0004)
Constant	0.092***	0.016***	0.062***
	(0.003)	(0.004)	(0.007)
Observations	21,848	21,848	21,848
R^2	0.014	0.042	0.049
Adjusted R^2	0.014	0.042	0.049
Residual Std. Error	0.231 (df = 21844)	0.228 (df = 21843)	0.227 (df = 21841)
F Statistic	106.030*** (df = 3; 21844)	242.083*** (df = 4; 21843)	187.637*** (df = 6; 21841)
Note:			*p<0.1;* ***p<0.05;*** p<0.01

Interestingly, controlling for education years gives better results than controlling for income, while based on p-values including both might not be a good option (remember the correlation between these variables).

Looking at the intercept, we can say that the conditional mortality rate on average is 6.2% for 50-year old males with no education, who hardly ever do sports. (Referring back to the distribution of the education variable, one could argue that the intercept is meaningless and should not be interpreted at all.)

Accounting for 95% CIs constructed with robust SEs, we see that doing sports more than hardly ever reduced the chance of death by around 3%-5.5% on average (separate coefficients could be more precisely interpreted, however given the overlap among the CIs I'll not do that here), controlling for the gender, age and education years.

Every year being older increases the chance of death by 0.5% in the dataset on average, controlling for other variables. Females are 3.7% less likely to die by the end of the observation period, on average, controlled for other factors. An additional year spent in education also has a negative impact on mortality years, -0.2% for every year on average, controlled for other factors.

The results of this "best" model are similar to the simple model run on just the "sports" variable. Even though the benefical impact of doing more sports is somewhat smaller, the magnitudes are similar. This can be possibily attributed to the fact that people in better general state might also do more sports (because they are able to do so), hence the first model overestimated its impact by taking credit for other health factors as well.

5. Re-do exercises 3 & 5 using logit. Calculate and interpret the marginal differences of the sports variables. Discuss the differences and similarities to the LPM results.

First, the results of the simple model:

	dF/dx	Std. Err.	z	P>|z|
factor(sports)1	-0.051	0.003	-16.7	0
factor(sports)2	-0.043	0.003	-14.4	0
factor(sports)3	-0.028	0.004	-7.7	0

Interpreting the marginal effect for "factor(sports1)": on average, people who reported doing sports more than once a week, had an expected mortality rate 5.1% less than people who never did sports. (Other variables should be interpreted analogusly)

And then the model with variables from the "best" LPM:

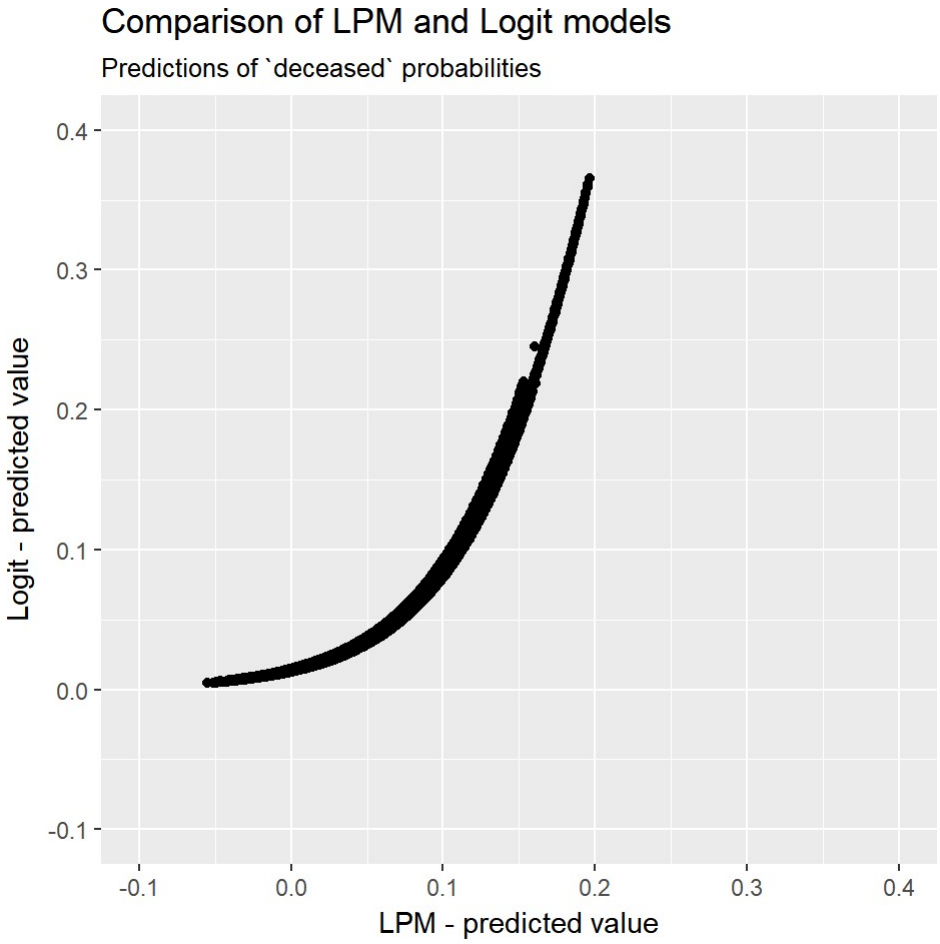
	dF/dx	Std. Err.	z	P>|z|
factor(sports)1	-0.038	0.003	-12.17	0
factor(sports)2	-0.034	0.003	-10.40	0
factor(sports)3	-0.025	0.004	-6.65	0
age_diff	0.005	0.000	17.96	0
female	-0.038	0.003	-11.85	0
eduyears_mod	-0.002	0.000	-4.82	0

Interpreting the marginal effect for "factor(sports1)": on average, people who reported doing sports more than once a week, had an expected mortality rate 5.1% less than people who never did sports, controlling for the impact of gender, age, and education years. (Other sports variables should be interpreted analogusly) For other variables, we can interpret marginal effects as we did coefficients in the LPM model: - Every year being older increases the chance of death by 0.5% in the dataset on average, controlling for other variables. - Females are 3.8% less likely to die by the end of the observation period, on average,

controlled for other factors. - An additional year spent in education also has a negative impact on mortality years, -0.2% for every year on average, controlled for other factors.

The impact of these control variables are almost virtually the same as it was in the LPM model. For the “sports” variable, we can observe slightly smaller absolute effects than we have seen in the LPM model, but otherwise the results are very similar.

We can see a comparison of predictions below:



As expected, we can generally observe that the logit model predicts larger probabilities than the LPM, as the “mitigation effect” of sports is smaller. Otherwise the predictions are similar, but the LPM model does predict some negative values (which are meaningless), as without a transformation in the formula, like the one logit has, the predicted probabilities are not bounded by 0 and 1.

So, in conclusion, both models could be appropriate for different reasons. Logit is better at actually making (meaningful) predictions, however if we are just interested in the impact of different variables, LPM could be a better choice, as its coefficients are directly interpretable.