THE RELATIONSHIP BETWEEN PHYSICAL ACTIVITY AND SOCIOECONOMIC FACTORS IN UKRAINE

by

Daniil Minakov

A thesis submitted in partial fulfillment of
the requirements for the degree of

MA in Economic Analysis.

Kyiv School of Economics

2024

Thesis Supervisor: _____ Professor Maksym Obrizan

Approved by _____
Head of the KSE Defense Committee, Professor

_____

_____

_____

Date _____

Kyiv School of Economics

Abstract

THE RELATIONSHIP BETWEEN PHYSICAL ACTIVITY AND SOCIOECONOMIC FACTORS IN UKRAINE

by Daniil Minakov

Thesis Supervisor: Professor Maksym Obrizan

Promoting physical activity is one of the keys to enhancing individual health and well-being. This study explores the relationship between physical activity, socioeconomic factors, and sports facility availability. By employing logistic regression analysis, the data from a household survey conducted in Ukraine is analyzed. Age demonstrates a strong negative relationship. Higher education, residency in the city, and income exhibit positive associations with physical activity. However, social status and gender do not show an interrelation with physical activity engagement. Also, there was no statistical significance between the number of sports facilities in the region and individual engagement in physical activity.

TABLE OF CONTENTS

LIST OF FIGURES

Number *Page*

LIST OF TABLES

Number *Page*

ACKNOWLEDGMENTS

I am grateful to my thesis advisor, Professor Maksym Obrizan, whose guidance had a significant impact on the completion of this research. I am also grateful to all the professors from the research workshops who participated in the review and guidance of this thesis. I appreciate being a part of the KSE community and am thankful for the support and opportunities it provides.

Chapter 1

INTRODUCTION

Physical activity has a positive impact on well-being and health (Liu and Zhong 2023) with consequent implications for public health at the macro level (Hunter et al. 2015). Level of engagement has been consistently associated with various health benefits, such as reducing the risk of chronic diseases, enhancing quality of life, or improving mental health (Peluso et al. 2005). Despite all the possible benefits of doing physical activities many individuals still do not do even the minimal recommended levels of physical exercises.

Understanding the level of physical activity in society is necessary to make policy decisions in the field of sports. In addition, disaggregating this information by criteria will provide a deeper understanding of the current physical activity situation in the country. In addition, the availability of sports facilities can also play an important role in physical activity.

This research aimed to investigate the relationship between engagement in regular physical activity, socioeconomic factors (people's income, gender, education, settlement, social status, region of residence), and availability of sports facilities across the regions in Ukraine. I identified disparities in physical activity levels based on income level, region of residence, and education. The results provide insights for policymakers to develop targeted (on a specific group of individuals) interventions to promote engagement in physical activity in society. I used logistic regression analysis to examine the complex relationship between individual-level characteristics, outer factors, and participation in regular physical activity, also explored the marginal effect at representative values for a deeper understanding of the non-linear nature of the relationship between participation in regular physical activity and income.

This research contributes to the existing literature by providing empirical results of the relationship between engagement in regular physical activity and different socioeconomic factors together with the availability of sports facilities in Ukraine. I used an approach from the study conducted in South Korea (Kim and So 2014) by applying logistic regression analysis, but extended it with more socioeconomic factors like other similar studies did (Hyytinen and Lahtonen 2013; Testoni et al. 2018).

Chapter 2

LITERATURE REVIEW

Existing studies include relationship analysis between physical activity and different social variables, such as income, gender, race, etc. Such types of studies cannot answer the questions: "How does a person's physical activity impact a person's income?" or "Which social factor determines the income?". But it does not mean that they are of low importance. Of course, there are cause-and-effect analyses of physical activity on wealthiness of the person. Those studies are distinguished by the complexity of gathering needed data.

The study for Korea (Kim and So 2014) explores the relationship between household income and physical activity among people in Korea.

They took the Korean Survey of Citizen's Sport Participation conducted by the Korean Ministry of Culture, Sports, and Tourism. Its structure consists of the 9,000 Koreans aged from 10 to 89 years. The survey's sports question: "Recently, on how many days did you do over 30 minutes of physical activity (or exercise), except walking, in your leisure time?" is the main for determining if a person does physical activity or not. The authors decided to differentiate respondents into two groups: those who did not do exercises and those who did at least once per week, this variable was further used as a dependent one. As an independent variable authors took the answers for the level of income, which was grouped into 12 ranges from less than 1 million won income to over 6 million won, with 5 hundred thousand won steps. Also, they included the participant's age and gender as covariant variables in the model.

The results show that there is a strong relationship between physical activity and income for Koreans controlling for age and gender. But the relationship is stronger for females than for males. Females' physical activity is rising for every income

range, while the relationship for men's physical activity is getting stronger only from 3.5 thousand dollars income. Meaning that there is no evidence of a strong positive or negative correlation between income and physical activity for men with income less than 3.5 thousand dollars in Korea.

The study does not provide a cause-and-effect analysis but only shows the interrelation between household income and physical activity. It can be because of the lack of data which is associated with the high complexity of data gathering for cause-and-effect analysis. It is needed to have yearly data for income and whether people doing sports to control for seasonality or other factors that can impact income.

The cause-and-effect analysis was done for twin males in Finland (Hyytinen and Lahtonen 2013). The authors took data from Statistics Finland which covers a period from 1990 to 2004 year. The sample size is 5,042 respondents, all are twin males. Other socioeconomic factors are not used in the analysis, which points to the limitations of the study – it does not consider gender, social status, education level, etc.

The most impressive thing about this research is that they used twins to control for unobserved genetic differences and family effects. Also, they differentiate the level of physical activity in respondents' early ages and explore the effect of this on long-term income. They found that the long-term income of physically active males is 14-17% higher compared with less physically active males.

Another work with Finland data (Kari et al. 2015), conducted on fresher data than the previous paper, explored the relationship. The authors had data for daily physical activities such as aerobic and basic steps made for more than 10 minutes at a time. The dataset consisted of 753 adults with a mean age of 41.7 years based on a survey conducted in 2011. They did not find evidence of a strong relationship between physical activity and income for men, but there is evidence for women.

Sibley et al. (2018) use bivariate analysis of the level of physical activity among young people in the USA by income, race, and gender. The authors focus on adolescents and young adults only, because physical activity in youth is associated with health in adulthood, they note.

They took data from the National Health and Nutrition Examination Survey from 2007 through 2016. It comprises 9472 respondents and is differentiated by gender, age, and race. The authors analyzed the respondents' answers to the questions about physical activity duration and intensity to check whether they meet the recommended guidelines for physical activity.

The reported results show that greater physical activity efforts are associated with younger age, white race, and higher income. Young females report less physical activity than males, black males have the longest duration of physical activity. This study includes more social factors than previously mentioned papers and provides its interrelationship with physical activity, which can be helpful for decision-makers in public health and sports.

Trying to measure how physical activity impacts income or the relationship between these variables can be widened to the question: "Are people who do regular physical activity happier than those who do not?". Another form of this question is vital for sports policymakers: "Does physical activity increase people's well-being?". Income can be treated as one of the parameters for the well-being of the person with a positive correlation: more income – better subjective well-being.

One way to explore these questions was presented in the recent study on subjective well-being (Testoni et al. 2018). To answer such a question, the authors had to define well-being. They evaluated subjective well-being based on self-assessments about how people feel overall. The authors reference other studies about how sport affects other areas of life while discussing the impact of doing sports or physical activity on subjective well-being, but the dependence on income is not considered.

They do not claim the positive impact on well-being, but the positive relationship between SWB and physical activity. Liu and Zhong (2023) support these findings.

With this research, I am trying to explore the relationship between income, education, social status, age, gender, and physical activity controlling for different available socioeconomic factors. It is an improved version of the Korean paper's approach, where the authors take the income on one side and physical activity on the other side of the equation and state the relationship omitting the fact that income is affected by a bunch of other social factors, for example education. It is hard to do the cause-and-effect analysis because of the lack of data. But checking and exploring the relationship is one of the steps for exploring the more fundamental question that is vital for Ukrainian policymakers: "How to make people happier?".

Chapter 3

METHODOLOGY

In this thesis, I follow the methodology used in Korean research (Kim and So 2014) – a multiple logistic regression model, but I extend it with more independent variables. Table 1 shows the list of the used independent variables in the model, next section discusses the data source.

Table 1. Independent variables description

Variable short name	Motivation
Person age	The younger generation might be more likely to engage in physical activity, older generation might be otherwise due to health conditions, energy, and lifestyle preferences.
Male	Studies show the gender difference level of physical activity engagement.
Education	It is assumed that more educated people are more aware of the importance of doing physical activities.
Is the person self-employed?	Usually, self-employed people have more control over their schedules which allows them to pick convenient times for physical activities.
Income	Higher income creates more opportunities for individuals to participate in physical activities, such as attending fitness clubs, buying sports equipment, etc.
Region of residence	Control variable.
BMI	Endogenous. People who do physical activities have lower BMI.
Health status proxy	Endogenous. People who do physical activities have worse health.
Having land in use or livestock	Control variable. Relate to lifestyle which can indirectly impact a person's engagement in physical activities such as outdoor work, gardening, farming, etc.
Is the person from the city	Control variable. Cities have more access to parks, sports clubs, etc.
Sports facility density	Count of publicly available places where people can do sports or physical activities, such as sports grounds of different types divided by region's area.

This model is usually used for the investigation of the relationship between binary dependent variables and independent variables, thus it is appropriate for analyzing the dependence of whether people do physical activities (dependent binary variable) and other socioeconomic factors

BMI and health status are proxy variables for health conditions. To check for the endogeneity I conduct two models: with those variables and without.

Multiple logistic regression models are used for predicting the probability of a binary dependent variable (outcome) accounting for one or more independent variables. The generic model is defined as:

$$P(Y = 1) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}} \quad (1)$$

Where:

- $P(Y = 1)$ is the probability of the dependent variable to equal 1.
- β-s are coefficients of the independent variable.
- X-s are independent variables.

For the convenience of the interpretation of the coefficients the model usually is re-arranged to have linear equation such as $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ on the right-hand side. The re-arranged form is defined as in (2).

$$ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \quad (2)$$

The left-hand side of the equation is called log odds or "the log of the odds". Odds are the probability of the dependent variable to equal 1 divided by the probability of the dependent variable to equal 0.

The typical interpretation of coefficients in such models is "*A one-unit change in X is associated with a one-unit change in the log odds of Y*". Basically, this interpretation is not practical because the initial model was re-arranged, so coefficients are in log odds units. To have the meaningful, or better name it straightforward, interpretation of the coefficients we must transform coefficients in odds ratios (OR) by exponentiating the coefficients. Then, the interpretation of log odds is the following: "*A one-unit change in X is associated with* $\{(e^{\beta} - 1) * 100\}$ *% change in the Y*", which is much easier to understand.

Unlike the simple Ordinary Least Square model Logistic model assumes the outcome is a binary and non-linear relationship between dependent and independent variables. Other assumptions of the logistic model include error independence and the absence of multicollinearity.

Here is the formal notation of the constructed Logistic model (3):

$$ln\left(\frac{P(DoPhysicalActivity)}{1 - P(DoPhysicalActivity)}\right) = \beta_0 + \beta_1 * Male + \beta_2 * PersonAge + \beta_3 * PersonIncome + \beta_4 * HasHigherEducation + \beta_5 * BMI + \beta_6 * IsFromCity + \beta_7 * IsSelfEmployed + \beta_8 * HelpMedCount + \beta_9 * LandPoul + \beta_{10} * IsFromWestRegion + \beta_{11} * IsFromCenterRegion + \beta_{12} * IsFromSouthRegion + \beta_{13} * SportFacilitiesDensity \quad (3)$$

The independent variables consist of two groups: variables under interest, and control variables. **Gender**, **PersonAge**, **PersonIncome**, **HasHigherEducation**, **SportFacilities, IsFromCity**, and **IsSelfEmployed** are variables under interest. Others: **BMI**, **HelpMedCount**, **LandPoul**, **IsFromWestRegion**, **IsFromCenterRegion**, and **IsFromSouthRegion** are added as control variables to reduce endogeneity.

Multicollinearity is tested via the variance inflation factor (VIF) for each covariate of the model. VIF is evaluated by taking each independent variable and regressing it for every other independent variable. Then R-squares are collected and VIF derived via the next formula (4):

$$VIF = \frac{1}{1 - R_i^2} \tag{4}$$

Where i correspondents to the independent variable we are checking.

The more the VIF value the more the evidence of multicollinearity. If VIF is 1 it means the variables are not correlated. Overall, the lower the VIF values the better as it points out that correlation is weaker, so multicollinearity is not severe.

As the Logistic model assumes no linear relationship between a dependent variable and independent variables – it is not enough just to calculate the coefficient of the independent variable. A non-linear relationship implies the variable's impact varies based on concrete values. To address this concern – the marginal effects must be calculated. It is taking the derivative with respect to the corresponding independent variable. For the Logistic model case, it shows the fact that probability changes when the independent variable is increased by one unit.

To construct the "representative" observation I picked values based on the prevalence of value in the sample. For example, the variable **IsSelfEmployed**. Only 186 people are self-employed among 5,547 people in the sample, so the representative value for this variable will be 0. For variables like **PersonAge**, there are three possible values which represent different age groups: value 2 corresponds to 18-35 y.o.; value 3 corresponds to 36-55 y.o.; and value 3 corresponds to 56-60 y.o, I chose value 3 for representative values as this group is the most represented in the dataset. For **PersonIncome**, which represents the individual's yearly income, I have calculated the mean value and divided by 10,000 to have values in tens of thousands for convenience. The same approach I used for **BMI** (Body Mass Index) variable. However, for **SportFacilityDensity** I calculated the mean also, but excluded the Kyiv city from the calculation as it is outlier with value 2.5, while the average for all other regions is 0.1. Table 2 presents the chosen values for representative observation.

Table 2. Representative observation values

Variable name	Value	Description
Male	1	Male
PersonAge	3	36-55 y.o.
PersonIncome	7.7	Mean value (₴, ten thousands)
HasHigherEducation	0	Secondary education
HelpMedCount	1	Sought medical help within the last year
LandPoul	1	Have land in use or livestock
BMI	25.6	Mean value
IsFromCity	1	Live in city
IsSelfEmployed	0	Not self-employed
IsFromWestRegion	0	-
IsFromCenterRegion	1	From center region
IsFromSouthRegion	0	-
SportFacilitiesDensity	0.1	The mean sport facility density in a region

The “representative” individual is male. He is within 36-55 y.o. He has secondary education, has land in use or poulty. Also, he seeked for a medical help within the last 12 months. His BMI is 25.6. He is not self-employed and lives in the center region of Ukraine in a big city. The sport facility density in his region is about 0.1 which corresponds to one sport facility per 10 squared kilometers.

Python and R are used as a primary tool for building and validating the model.

Chapter 4

DATA

The State Statistics Service of Ukraine has a dataset with microdata on the main indicators of income, expenses, living conditions, and others of households and their members for 2021 released at the end of 2022 (2022). It includes 7,614 households consisting of 15,824 people separated into two different tables named "households" and "members" (for convenience it is further called "persons", "individuals" or "members") tables. The statistics do not include conscripts, homeless people, prisoners, people permanently living in boarding schools or homes for the elderly. Also, people living in the temporarily occupied territories of Crimea and parts of the temporarily occupied territories of Luhansk and Donetsk regions did not participate in the survey. Also, the State Statistics Service of Ukraine reported amount of sports facilities present in Ukraine, which I divided by region area and included in the model as the density of sports facilities represents the proxy for the individuals' access to opportunities for physical activity (2018). The higher availability the more chances people will be engaged in physical activities (Cohen et al. 2013).

To prevent the establishment of a person's identity from the data, the State Statistics Service of Ukraine depersonalized data using global recoding, aggregation, and masking. For example, age is grouped into ranges: before 18 years, 18-35 years, 36-55 years, 56-59 years, and 60 and more years; the outlier values for expenses and income were changed with the average values. After all, the State Statistics Service of Ukraine claims that after the deanonymization the averages for the data do not differ more than 2%, so it can be safely used for analysis as the risk of revealing the person's identity is minimized. They provided weights for each observation in the dataset, it was accounted for in the model.

This is a very extensive survey that contains a lot of different variables, for example for households: region of residence, socio-economic status, total income and expenditure, income, and expenditure by different types of products or services, etc.; for individuals: age, level of education, height, weight, whether a person does sport at least once a week, income, etc. Overall, there are 139 variables in the "households" table and 94 in the "persons" table. So, I left only those that were used in the analysis and presented in figures variables, which are renamed for convenience, they are described in APPENDIX A with other derived variables.

If, for example, we add up all counts for age subgroups – the sum is 15,753, but the total number of persons is 15,824. It means that there are other values for the **age** variable compared to the reported by the State Statistics Service of Ukraine. It was decided to filter out such inconsistencies in the "households" and "persons" tables. The full information about how many such rows were detected and for which variables are described in APPENDIX B.

Also, the dataset includes information about whether the person sought medical help within the last 12 months and the primary reason for the required medical support such as trauma, sickness, medical prevention, renewal of prescriptions, and others. Such information will be used to create a proxy variable for health status. Health status is highly likely associated with the person's engagement in physical activity (Galán et al. 2013), if the person sought medical help, it could indicate the presence of underlying health conditions.

Height and weight information is also present. It is not directly related to the person's engagement in physical activity, but they can be used as control variables. It allows us to calculate BMI (Body Mass Index). The higher the index the more chances for a person to have barriers to physical activity. (Hemmingsson and Ekelund 2007)

The presence of livestock or land reveals the lifestyle of the household. This is also not directly related to physical activity engagement but could indirectly reflect the living conditions that can impact on person's engagement in physical activities. For example, the presence of land increases duties that involve physical activities such as outdoor work, farming, or gardening.

That is, the dataset is compiled separately from two tables: a table for households and a table for individuals. Each household has its unique identifier, for each person, the identifier of the household to which he/she belongs is indicated. So, for linking the data from two tables I rely on this identifier. But before merging the data it is filtered.

The income variable is represented in Ukrainian hryvnias. I reported it in dollars using the exchange rate of 27.29 (NBU). And also turned it into thousands to simplify the interpretation of the further regression results.

People under 18 years old and over 60 years old are excluded, because younger people may not have their income, and older people may not be able to do physical activity or not have a job at all. Retired people are also excluded as there is for sure no impact from doing physical exercises on the amount provided by the government pension. It is done by applying constraints for **PersonAge** and **IsRetired** variables: excluded rows if **PersonAge** equals 1 or 5, excluded rows where **IsRetired** equals 1.

Also, if the household has kids – leave only those with one or two kids as the reported share of households with one or two children among households with children is 97.5% (State Statistics Service of Ukraine. 2020. "Children, females and family in Ukraine"). If the household does not have kids – leave it as it is. For this restriction the next constraints are applied: if the variable **HhWithChildren** is 0 (household without children) – do not apply any restrictions, if the variable

HhWithChildren is 1 (household with children) – filter out all households with **HhSize** (number of members in the household) more than 4.

To find the corresponding person from the "persons" table who is head of the specific household – the **FamilyId** variable is used which is the linkage key variable between both. This variable is unique for each household, it means that it is unieuq for household table, but can be repeated in the individuals table.

Table 3 presents the summary of the dataset after applying the filtering discussed above.Figure 1, Figure 2, and Figure 3 show the relationship between physical activity and socioeconomic factors.

Table 3. Dataset summary

Variable/Characteristic	Count of persons
Count	5,547
Do physical activity	36.4% (2,018)
Average income	77,020₴ (2,822$)
Age:	-
18-35 years	29.0% (1,608)
36-55 years	63.1% (3,501)
56-59 years	7.9% (438)
Gender:	-
Male	50.2% (2,784)
Female	49.8% (2,763)
Settlement:	-
Countryside	40.9% (2,269)
Small City	25.2% (1,396)
Big City	33.9% (1,882)
Education:	-
Higher	48.6% (2,696)
Secondary	51.4% (2,851)
Socioeconomic status:	-
Employed	71.5% (3,966)
Self-employed	3.4% (186)
Other	25.1% (1,395)

There are 5,547 persons in the final table, while in the initial "persons" table there are 15,824 persons, which is about one-third. Two-thirds of them are 36-55 years old. Males and females are approximately equally distributed. Most of the people, or 40.9%, live in the countryside, then, 33.9%, live in the big cities, and others are from the small cities. Almost half of the respondents have higher education, others have secondary education. Only 3.4% are self-employed, others are either employed or have other type of employment.

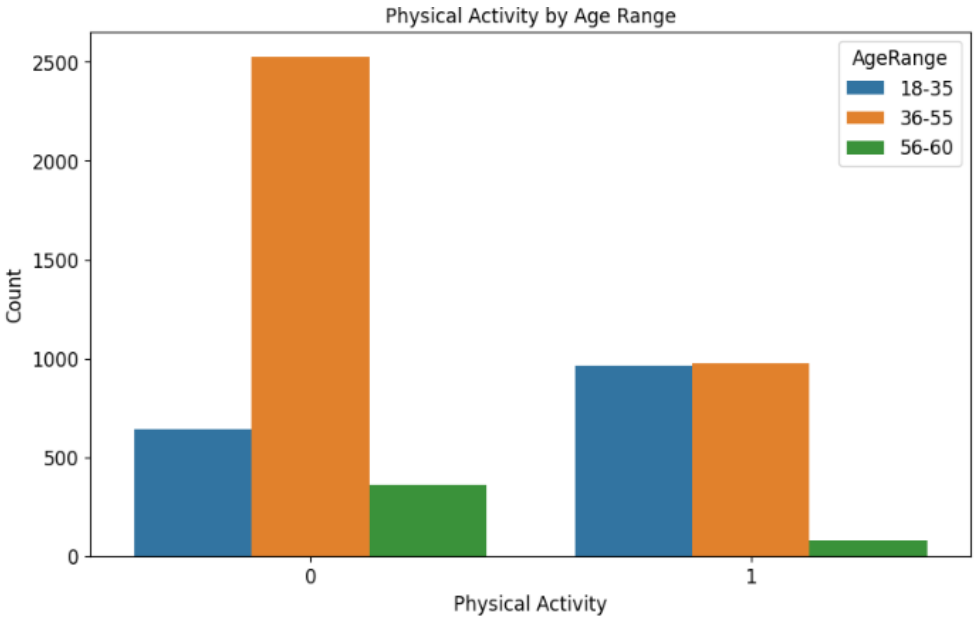

Figure 1 Physical Activity by Age Ranges

Figure 1 is a count plot that examines the distribution of physical activity between different age groups. The youngest group ranging – from 18-35 years old is the only group where the majority of individuals reported that they do physical activity at least once per week. At the same time, two older groups (36-55- and 56-60-year-old groups) have lower proportions of people who do physical activity. The oldest

group (56-60 years old) has the lowest number of respondents, which is expected as it is the smallest range. In absolute values, the 18-35-year-old age group has slightly fewer people engaged in physical activity than the 36-55-year-old group, but the portions within the groups are different, suggesting that younger people have higher engagement in physical activity.

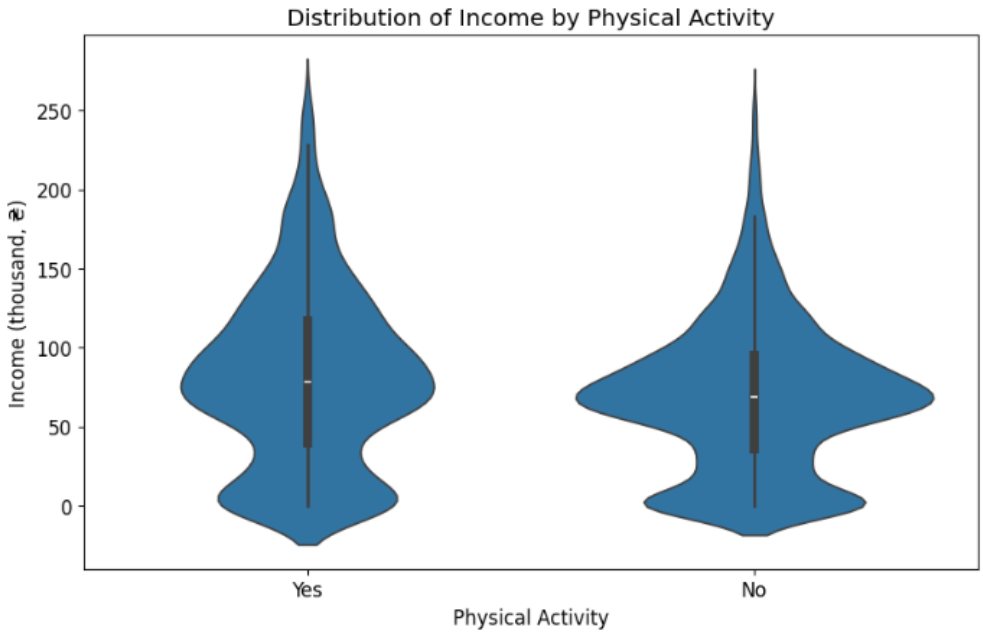

Figure 2 Distribution of Income by Physical Activity

The distribution of income by physical activity is depicted on violinplot (Figure 2). I have deleted outliers higher than three standard deviations in income to focus on the central tendency and make the plot clearer for understanding. It shows that individuals who engage in physical activity exhibit slightly higher median income compared to those who do not. Also, there are more individuals around the mean who are not engaged in physical activity, while the distribution of income of

individuals who engage in physical activity is thicker for income further from the mean.

Figure 3 shows the relationship between physical activity and city residence (a), gender (b), self-employment(c), and education(d). A higher proportion of individuals in the city engage in physical activity compared to those from the countryside. There is no difference if an individual is male or female, they both have approximately the same shares of individuals engaged in physical activity. It is hard to derive from the self-employment plot if there is any interrelation as few people are self-employed. Individuals with higher education exhibit higher engagement in physical activity.

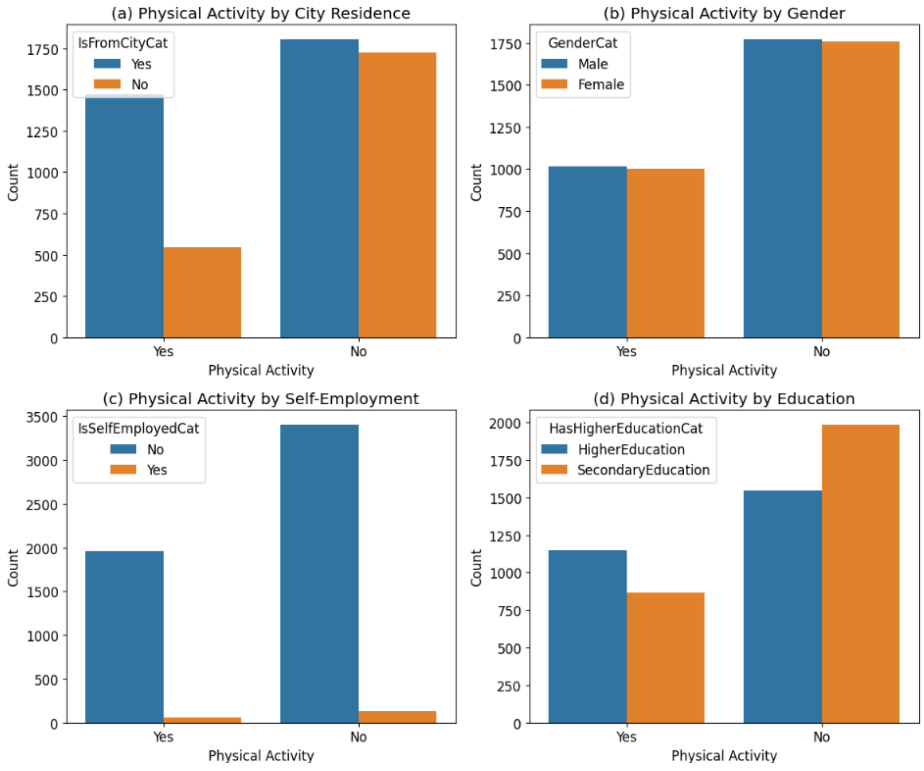

Figure 3 Physical Activity by City Residence (a), Gender (b), Self-Employment (c), and Education (d)

Figure 4 and Figure 5 show shares of individuals engaged in physical activity and available sports facilities correspondingly by region. The highest shares have large regions in Ukraine by population: Lviv, Dnipropetrovsk, Volyn, and Kyiv city. At the same time, Lviv, Dnipropetrovsk, and Kharkiv regions have a higher amount of sports facilities, but Kyiv city is not at the top. Other regions have lower physical activity and physical facilities.

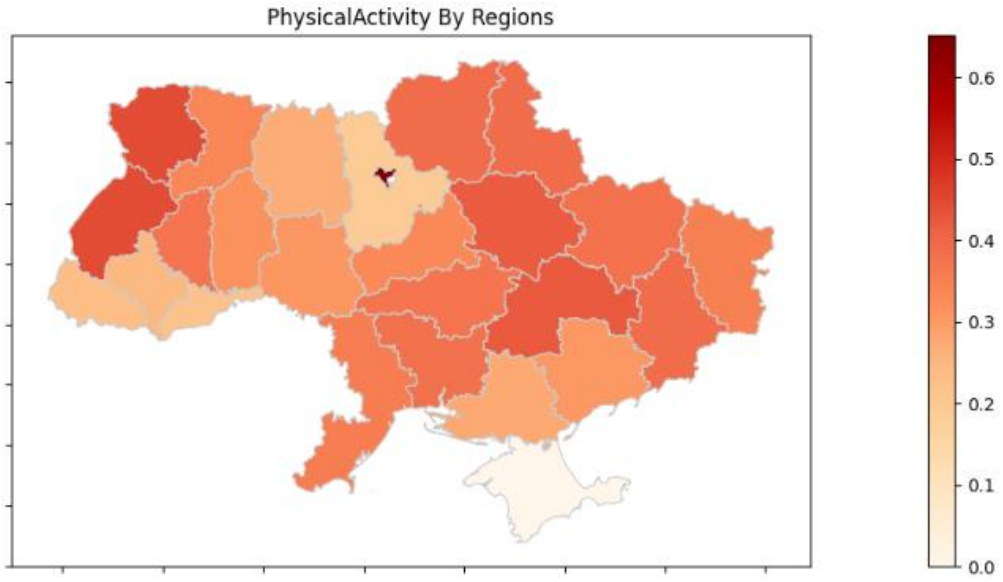

Figure 4 Shares of individuals engaged in physical activities by region.

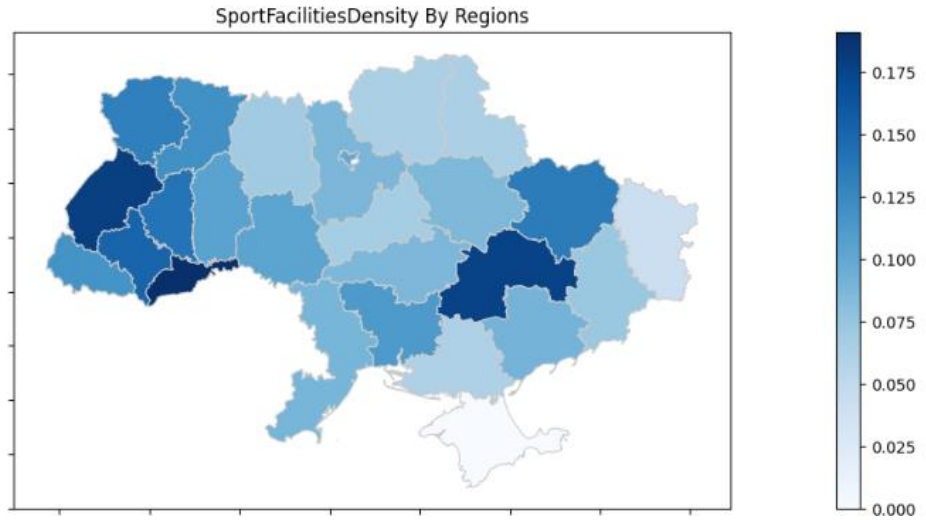

Figure 5 Sports facility density by regions (Kyiv's density is set to mean for visualization, actual value for Kyiv is more than 2.5).

Chapter 5

ESTIMATION RESULTS

In this chapter, I present the estimation results of the logistic model. Table 4 presents the models' summaries. It includes coefficient estimates, standard errors, and p-values. Stars show the significance of the coefficients based on p-values, they are standard for models evaluated in R, the breakdown is: "***" – the significance level is 0.01; "**" – 0.05; "*" – 0.1.

Table 4. Models Summaries

	Dependent variable:	
	DoPhysicalActivity	
Independent variables	(1)	(2)
(Intercept)	3.562***	1.805***
	(0.617)	(0.440)
Male	-0.073	-0.068
	(0.143)	(0.142)
PersonAge	-0.976***	-1.145***
	(0.139)	(0.134)
PersonIncome	0.034***	0.029**
	(0.012)	(0.012)
HasHigherEducation	0.280*	0.283*
	(0.151)	(0.149)
IsFromCity	0.505***	0.541***
	(0.180)	(0.177)
SportFacilitiesDensity	0.267***	0.294***
	(0.104)	(0.103)
IsSelfEmployed	-0.326	-0.350
	(0.470)	(0.469)

TABLE 4 — Continued

	Dependent variable:	
	DoPhysicalActivity	
Independent variables	(1)	(2)
BMI	-0.086***	
	(0.021)	
HelpMedCount	-0.077	
	(0.152)	
LandPoul	0.082	0.125
	(0.161)	(0.159)
IsFromWestRegion	-0.213	-0.214
	(0.205)	(0.203)
IsFromCenterRegion	-0.101	-0.141
	(0.191)	(0.189)
IsFromSouthRegion	-0.012	-0.025
	(0.242)	(0.238)
Observations	5,547	5,547
Log Likelihood	-300.309	-310.200
Note:	*p<0.1; **p<0.05; ***p<0.01	

The significant variables at 95% level are: income, age, a person from the city, and BMI. The higher education variable is marginally significant at a 90% level. The coefficient for the intercept is also significant. Age, income, city residence, BMI, and sport facility density are all strongly significant even for 99% confidence interval which indicates strong relationship with the dependent variable DoPhysicalActivity. Region variables are all statistically insignificant, this suggests that regionality is not related with physical activity. However, including the specific regions could show significance but such analysis is beyond the scope of this research. The gender variable also is not statistically significant which tells us that

there is no relation between the gender of an individual and his or her engagement in physical activity.

The second model is evaluated without **BMI** and **LandPoul** variables, which were chosen to try to control for the endogeneity problem in the model. The coeffcients and significance between two models are relatively the same, which is an evidence for the low impact of the endogeneity problem.

Table 5 contains calculated odds ratios and corresponding confidence intervals at a 95% confidence level.

Table 5. Odds Ratios

	OR	2.5%	97.5%
(Intercept)	35.221	10.651	119.714
Male	0.930	0.702	1.230
PersonAge	0.377	0.286	0.493
PersonIncome	1.035	1.010	1.061
HasHigherEducation	1.324	0.985	1.779
IsFromCity	1.657	1.167	2.364
SportFacilitiesDensity	1.306	1.066	1.607
IsSelfEmployed	0.722	0.272	1.754
BMI	0.918	0.881	0.955
HelpMedCount	0.925	0.687	1.248
LandPoul	1.086	0.793	1.490
IsFromWestRegion	0.808	0.540	1.206
IsFromCenterRegion	0.904	0.621	1.316
IsFromSouthRegion	0.988	0.613	1.585

A person's age has a negative relation with a person's choice to do physical activity. It is hard to tell the exact number as the person's age is represented as ranges of unequal length.

There is strong evidence that income and city residence are positively related to a person's choice to do physical activities. An additional 10 thousand hryvnia in yearly income is associated with a 3.5% higher chance for a person to do physical activities. The city residence is associated with 65.7% increase in odds of engagement in physical activities.

There is weak evidence that people with higher education are more likely to do physical activities by 32.4%.

All other control variables except the BMI index are not significant. The odd ratio for the BMI index variable is negative.

Figure 6 shows VIF values for each independent variable.

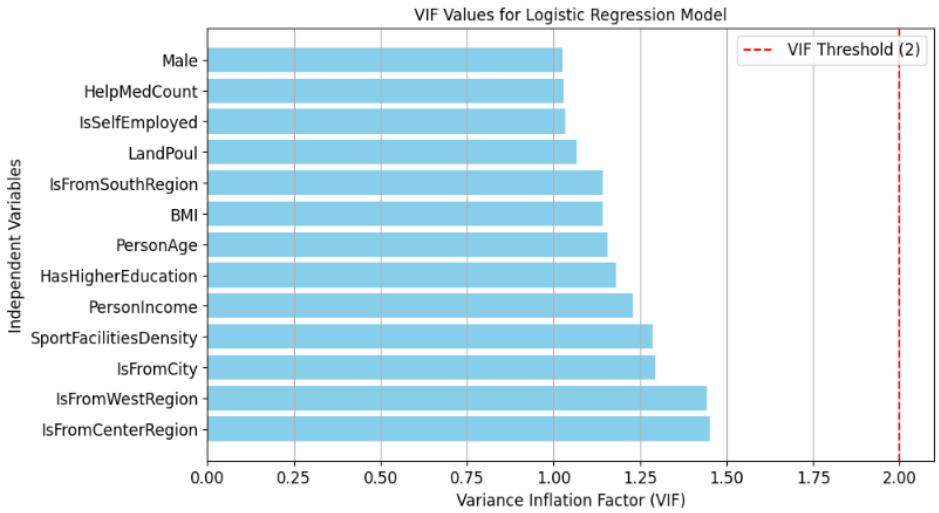

Figure 6 VIF Values

VIF values are below 1.5 for each variable which is evidence of not severe multicollinearity.

For margin calculation, the MER (marginal effect at representative values) method is chosen. The model includes many dummy and categorial variables, which is why the mean method for margin calculations is not appropriate in this case.

Let's focus on statistically significant marginal effects of variables: income, age, higher education, and if a person lives in a city.

Figure 7 represents margin values for a person's income holding other variables at representative values.

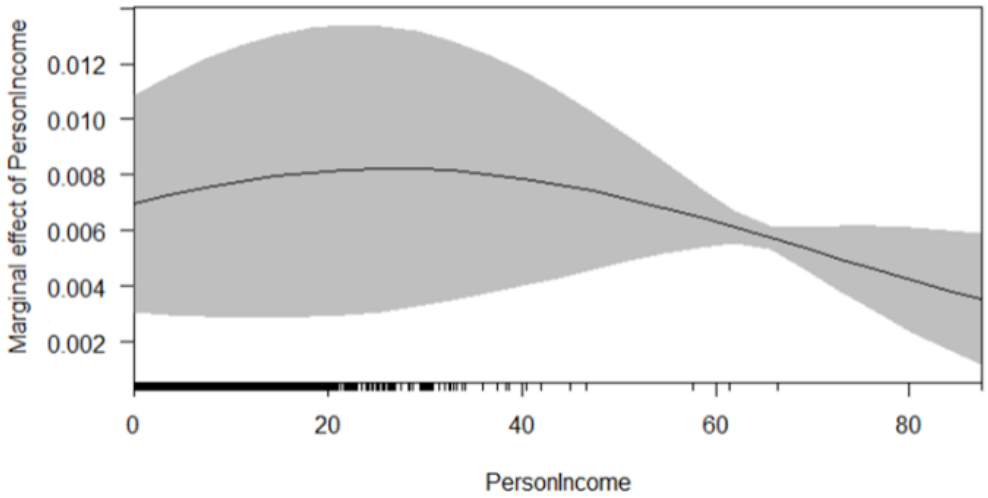

Figure 7. Marginal effect of PersonIncome

The marginal effect of a person's income shows a positive non-linear relationship with a 95% confidence level. It means that holding all other variables at their representative values – one unit change in a person's income is associated with a different increase in the probability of a person doing physical activity. The marginal effect starts at ~0.007 and increases till the 200 thousand hryvnias income,

then declines till 0.003 at the 850 thousand hryvnias income. However, the marginal effect is positive for the entire range of a person's income, which means that one unit change in a person's income is associated with a higher probability, but at a decreasing rate, up to 250 thousand hryvnas in income the one point increasing busts the probability of doing physical activity by approximately 0.9% (the confidence interval expands the range to 0.3-1.2%), after 250 thousand hryvnas income each increase of a person's income still associated with a positive chance increase, but at decreasing rate up to 0.3% at more than 850 thousand hryvnas income.

For dummy variables under interest, I calculate the predicted probability of the Y for two scenarios: holding all X-s at representative values (see Table 2) when the dummy variable is 0 and when it is 1. Table 6 shows the marginal effects of dummy variables. The table includes only variables with statistically significant coefficients: city residence variable and the variable for controlling the education level of an individual.

Table 6. MER for dummy variables

Dummy variable name (only significant)	Marginal effect on representative values
HasHigherEducation	0.0559
IsFromCity	0.0820

A coefficient on higher education is significant at 90%, it is almost significant at 95% and has a positive sign. It can be considered as, for sure, not strong, but evidence that if a person has higher education, she or he will more likely do physical activity by 5.6%. It is less than odd ratio values, but the marginal effect is calculated by holding other variables at representative values.

The average marginal effect on representative values of sport facility density is 5.1% (it is averaged as it does not deviate much from this value). Even though it is statistically significant and positive the value is only 5.1%, where at the same time the average value of density across regions is 0.1. It means that by doubling sports facilities in the region the effect will be expected only a tenth of 5.1%. So, the amount of sports facilities does not have a high interrelation with engagement in physical activities.

If a person is from a city (basically it means the person is not from the countryside) – the prediction of doing physical activity is increased by 8.2%. Its coefficient is strongly significant, so the evidence of a positive relationship is strong.

The age variable is statistically significant with a negative effect. However, the age variable has three possible values – 2, 3, and 4 thus it is more appropriate to calculate the marginal effect in the same way as for dummy variables –as a change in the predicted probability between two groups. Table 7 represents the marginal effects based on predicated values at representative values.

Table 7. MER via predicted values for PersonAge

PersonAge value and meaning	Predicted values at representative values	The difference
2 (18-35 y.o.)	0.467	-
3 (36-55 y.o.)	0.248	-0.219
4 (56-60 y.o.)	0.110	-0.138

People in the 36-55 age group have a significant change in predicted value compared with the 18-35 age group. The difference is -21.9% in predicted chance for a person to do physical activity. For the difference between the 36-55 age group and the 56-60 age group, the difference is lower and equals -13.8% If the age data

were represented not by age ranges, but exact years – the marginal effect for sure would be lower. Nevertheless, the impact is negative and strongly statistically significant.

Chapter 6

CONCLUSIONS AND POLICY RECOMENDATIONS

The findings of this study show insights into the interrelation between physical activity engagement and different socioeconomic factors, such as age, income, gender, region of residence, social status, and density of sports facilities in Ukraine. By applying the logistic regression, I identified significant predictors of physical activity among mentioned factors. Age is a negative predictor; the marginal effect reaches -21.9% between the 18-35 and 36-55 age groups. This highlights the importance of interventions to promote physical activity among older populations. Education has a marginally positive association, suggesting that educational programs could have the potential to promote physically active behavior. People with higher incomes also are associated with higher engagement in physical activities. Each 10 thousand UAH is associated with a 3.5% increase in the odds of an individual's engagement in physical activity. However, the relation is not linear, the association increases up to 0.8% with an income value of 230 thousand UAH, then gradually diminishes to 0.4% with an income value of 850 thousand UAH.

Moreover, city residence has the highest association in the model, as per the odd ratio of city residence variable people from cities are almost 65.7% more likely to be physically active than rural residents. It shows that we have a significant gap between these groups of population.

Additionally, the density of sports facilities showed a significant, but quite low effect. The increase of sports facility density by 1 is associated with an increase in odds of physical activity by 5.1%. At the same time, the mean density across regions is only 0.1. So, the doubling of the number of sports facilities will increase the density to 0.2. As per the developed model increase of 0.1 in the density of sports facilities would be associated with an increase of physical activity by about 0.5%.

Which is too low compared with the required resources to double the sports facilities.

This research did not show any significant differences in physical engagement in different regions in Ukraine. Also, I did not find any evidence of differences between males and females as well as differences between employed and self-employed people.

The dataset used in this study is from a survey conducted in 2021. However, after the full-scale invasion into Ukraine in February 2022 by Russia, a significant number of individuals have migrated. A large portion of these individuals are educated, urban residents, and have relatively higher incomes. This exacerbates the gaps highlighted in my analysis underscoring. Of course, concrete numbers could be calculated when we have a fresher survey.

The policies aimed to increase engagement among the population should be focused on people from 35+ age group and smaller cities. Also, straightforward decisions such as building more sports facilities might not work. It means that to handle the issue of low engagement the policymakers should focus on more complex solutions, not just building more facilities. The more complex solutions may include promoting, motivating, and encouraging people through marketing companies, or creating communities, etc.

WORKS CITED

Cohen, D. A., B. Han, K. P. Derose, S. Williamson, T. Marsh, and T. L. McKenzie. 2013. "Physical activity in parks: a randomized controlled trial using community engagement" *Am. J. Prev. Med.* https://doi.org/10.1016/j.amepre.2013.06.015

Galán, I., R. Boix, M.J. Medrano, P. Ramos, F. Rivera, R. Pastor-Barriuso, and C. Moreno. 2013. "Physical activity and self-reported health status among adolescents: a cross-sectional population-based study" *BMJ Open* 2013;3:e002644. https://doi.org/10.1136/bmjopen-2013-002644

Hemmingsson, and E. U. Ekelund. 2007. "Is the association between physical activity and body mass index obesity dependent?". *Int J Obes* 31, 663–668 (2007). https://doi.org/10.1038/sj.ijo.0803458.

Hunter, R. F., M. Boeri, M. A. Tully, P. Donnelly, and F. Kee. 2015. "Addressing inequalities in physical activity participation: Implications for public health policy and practice". *Preventive medicine*, 72 , pp. 64-69. https://doi.org/10.1016/j.ypmed.2014.12.040

Hyytinen, A., and J. Lahtonen. 2013. "The effect of physical activity on long-term income" *Social Science & Medicine* Volume 96, Pages 129-137. https://www.sciencedirect.com/science/article/abs/pii/S0277953613004188.

Kari, J. T., Jaakko Pehkonen, Mirja Hirvensalo, Xiaolin Yang, Nina Hutri-Kähönen, Olli T. Raitakari, Tuija H., and Tammelin. 2015. "Income and Physical Activity among Adults: Evidence from Self-Reported and Pedometer-Based Physical Activity Measurements" *PLoS ONE* 10(8). https://doi.org/10.1371/journal.pone.0135651.

Kim, Ill-Gwang, and Wi-Young So. 2014. "The Relationship between Household Income and Physical Activity in Korea" *Journal of Physical Therapy Science* Volume 26 Issue 12 Pages 1887-1889. https://www.jstage.jst.go.jp/article/jpts/26/12/26_jpts-2014-254/_article/-char/en.

Liu, N., and Q. Zhong. 2023. "The impact of sports participation on individuals' subjective well-being: the mediating role of class identity and health".

Humanit Soc Sci Commun 10, 544 (2023). https://doi.org/10.1057/s41599-023-02064-4

NBU, 2024. "The official exchange rate of the hryvnia against foreign currencies (average for the period)" https://bank.gov.ua/files/Exchange_r.xls.

Peluso, M.A.M., and L. H. S. Guerra de Andrade. 2005 "Physical activity and mental health: the association between exercise and moode". *Clinics*. https://doi.org/10.1590/S1807-59322005000100012

Sibley, L., S. Armstrong, C. A. Wong, E. Perrin, S. Page, and A. Skinner. 2018. "Association of Physical Activity With Income, Race/Ethnicity, and Sex Among Adolescents and Young Adults in the United States: Findings From the National Health and Nutrition Examination Survey, 2007-2016." *JAMA Pediatr.* 732–740. https://jamanetwork.com/journals/jamapediatrics/article-abstract/2684233.

State Statistics Service. 2018. "Institutions of culture, physical culture and sports of Ukraine in 2017".

State Statistics Service. 2022. "Anonymous microdata on basic indicators of income, expenditure and living conditions of households".

Testoni, S., L. Mansfield, and P. Dolan. 2018. "Defining and measuring subjective well-being for sport policy." *International Journal of Sport Policy and Politics* Volume 10, - Issue 4: Theory and Methods, Pages 815-827. https://www.tandfonline.com/doi/full/10.1080/19406940.2018.1518253.

APPENDIX A

FamilyId (original variable is ***code_fam***, exists in "households" and "persons" tables) – unique identifier for the family, it is unique in the "households" table, but repeated in the "persons" table. It is used as a linkage between households and persons.

HhSize (original variable is ***hzise*** from the "households" table) – the count of people in the household.

HeadGenderAndAge (original variable is ***gnd*** from the "households" table) – a person who is the head of the household. Possible values: 1 (female, 18-29 years), 2 (female, 30-59 years), 3 (female, more than 60 years), 4 (male, 18-29 years), 5 (male, 30-59 years), 6 (male, more than 60 years).

HeadGender (derived from the original variable ***gnd*** from the "households" table) – gender of head of the household. It is a dummy variable where if its value equals 1 – the head of the household is male, if 0 – the head of the household is female. Derivation rules: if ***gnd = 1 OR gnd = 2 OR gnd = 3*** then **HeadGender** = 0; if ***gnd = 4 OR gnd = 5 OR gnd = 6*** then **HeadGender** = 1.

DoPhysicalActivity (derived from the original variable ***SPORT*** from the "persons" table) – a dummy variable, indicates if the person does physical activity at least once per week. If the value equals 1 – the person replied he or she does physical activities at least once per week, 0 otherwise. Derivation rules: if ***SPORT =*** 1 then **DoPhysicalActivity** = 1; if ***SPORT = 2 OR SPORT = 9*** then **DoPhysicalActivity** = 0.

PersonIncome (derived from the original variables ***PPINC1*** and ***PPINC2*** from the "persons" table) – total person's income. Derivation rule: **PersonIncome = *PPINC1 + PPINC2***. ***PPINC1*** – main salary, compensations, dividends, etc.

PPINC2 – additional sources of income, such as scholarships, pensions, government unemployment benefits, etc.

Derived settlement dummy variables: **IsFromBigCity**, **IsFromSmallCity**, and **IsFromCountryside** are derived from the original variable ***tp_ns_p*** from the "persons" table. Dummy variables derivation rules: **IsFromBigCity** = 1 if ***tp_ns_p = 1***, 0 otherwise; **IsFromSmallCity** = 1 if ***tp_ns_p = 2***, 0 otherwise; **IsFromCountryside** = 1 if ***tp_ns_p = 3***, 0 otherwise.

HhWithChildren (derived from the original variable ***type_dom*** from the "households" table) – a dummy variable: 1 if the household has children, and 0 otherwise. Derivation: original value 1 is compiled into 1; original value 2 is compiled into 0.

PersonAge (original variable is ***AGE*** from the "persons" table) – person's age. Possible values: 1 – less than 18 years, 2 – 18-35 years, 3 – 36-55 years, 4 – 56-59 years, 5 – 60 years or more.

Gender (derived from the original variable ***SEX*** from the "persons" table) – person's gender. Derivation rules: original value 1 (male) is compiled into 1; original value 2 (female) is compiled into 0.

Derived education dummy variables: **HasHigherEducation**, **HasSecondaryEducation**, and **HasNoEducation** are derived from the original variable ***L_EDUC_M*** from the "persons" table. Dummy variables derivation rules: **HasHigherEducation** = 1 if ***L_EDUC_M = 1***, 0 otherwise; **HasSecondaryEducation** = 1 if ***L_EDUC_M = 2***, 0 otherwise; **HasNoEducation** = 1 if ***L_EDUC_M = 3***, 0 otherwise.

Derived socioeconomic status dummy variables: **IsEmployed**, **IsSelfEmployed**, **IsRetired,** and **HasOtherEmploymentForm** are derived from the original variable ***ses_mem*** from the "persons" table. Dummy variables derivation rules: **IsEmployed** = 1 if ***ses_mem = 1***, 0 otherwise; **IsSelfEmployed** = 1 if ***ses_mem***

= 2, 0 otherwise; **IsRetired** = 1 if ***ses_mem = 3***, 0 otherwise; **HasOtherEmploymentForm** = 1 if ***ses_mem = 4***, 0 otherwise.

Derived variable which indicated whether person sought for medical help because of one of the following reasons: trauma, sickness, medical prevention, renewal of prescriptions – **SoughtForMedicalHelp** = 1 if ((help_med == 1) AND (trauma == 1 OR sickness == 1 OR prf_insp == 1 OR prescript == 1), and 0 otherwise.

BMI (body mass index, derived from ***HEIGHT*** and ***WEIGHT*** source variables). The formula is the next: $BMI = \frac{WEIGHT}{HEIGHT^2}$.

Derived variable **HhHasLivestockOrLandInUse** which indicates whether household has livestock or land in use compiled from ***poultry*** and ***landplot*** source variables. It equals 1 if ***poultry*** or ***landplot*** equal 1, and 0 otherwise.

IsFromVinnytsiaOblast (derived from the original variable ***cod_obl*** from the "persons" table) – a dummy variable: 1 if the person from the Vinnytsia oblast, 0 otherwise. Derivation: the original value 5 is compiled into 1, and any other original value is compiled into 0.

The above pattern also applied to other regions, so to shorten the descriptions – below are: the names of the derived variables with the corresponding values of ***cod_obl***.

IsFromVolynOblast – for ***cod_obl*** = *7*.

IsFromDnipropetrovskOblast – for ***cod_obl*** = *12*.

IsFromDonetskOblast – for ***cod_obl*** = *14*.

IsFromZhytomyrOblast – for ***cod_obl*** = *18*.

IsFromZakarpattiaOblast – for ***cod_obl*** = *21*.

IsFromZaporizhzhiaOblast – for ***cod_obl*** = *23*.

IsFromIvanoFrankivskOblast – for ***cod_obl*** = *26*.

IsFromKyivOblast – for ***cod_obl*** = *32*.

IsFromKirovohradOblast – for ***cod_obl*** = *35*.

IsFromLuhanskOblast – for ***cod_obl*** = *44*.

IsFromLvivOblast – for ***cod_obl*** = *46*.

IsFromMykolaivOblast – for ***cod_obl*** = *48*.

IsFromOdesaOblast – for ***cod_obl*** = *51*.

IsFromPoltavaOblast – for ***cod_obl*** = *53*.

IsFromRivneOblast – for ***cod_obl*** = *56*.

IsFromSumyOblast – for ***cod_obl*** = *59*.

IsFromTernopilOblast – for ***cod_obl*** = *61*.

IsFromKharkivOblast – for ***cod_obl*** = *63*.

IsFromKhersonOblast – for ***cod_obl*** = *65*.

IsFromKhmelnytskyiOblast – for ***cod_obl*** = *68*.

IsFromCherkasyOblast – for ***cod_obl*** = *71*.

IsFromChernivtsiOblast – for ***cod_obl*** = *73*.

IsFromChernihivOblast – for ***cod_obl*** = *74*.

IsFromKyivCity – for ***cod_obl*** = *80*.

IsFromWestRegion – IsFromLvivOblast or IsFromVolynOblast or IsFromZakarpattiaOblast or IsFromIvanoFrankivskOblast or IsFromChernivtsiOblast or IsFromKhmelnytskyiOblast or IsFromTernopilOblast or IsFromRivneOblast.

IsFromCenterRegion – IsFromKyivOblast or IsFromZhytomyrOblast or IsFromVinnytsiaOblast or IsFromKyivCity or IsFromChernihivOblast or IsFromSumyOblast or IsFromCherkasyOblast or IsFromKirovohradOblast or IsFromPoltavaOblast.

IsFromSouthRegion – IsFromMykolaivOblastor or IsFromOdesaOblastor IsFromKhersonOblast.

APPENDIX B

The State Statistics Service of Ukraine provides a helper overview file. It describes the possible values for each variable for both tables: "households" and "members". But, while working with raw data, there were found inconsistencies between declared in helper file possible values and actual values in tables. It was decided to remove all the entries, where such inconsistency was found, for raw variables that were used in analysis or derivation of other variables.

For example, the ***age*** variable should have values from 1 to 5 inclusively as per the helper file, but there are 71 rows in the raw "members" table for which the ***age*** variable has 0 value. Other examples: in the "members" table: ***tp_ns_p*** (person's settlement, 65 entries), ***L_EDUC_M*** (person's education, 967 entries), ***L_EDUC_M*** (person's socioeconomics status, 721 entries); in the "households" table: ***cod_obl*** (household's region, 10 entries), ***tp_ns_p*** (household's settlement, 21 entries).

Also, there were observations with ***HEIGHT*** and ***WEIGHT*** variables with zero values. It is decided to delete them too.