UKRAINIAN CATHOLIC UNIVERSITY
APPLIED SCIENCE FACULTY
INTRODUCTION TO APPLIED ECONOMETRICS

KOSTYANTYN HRYTSYUK

ANALYSIS OF THE IMPACT OF DIFFERENT HEALTH-RELATED HABITS ON THE ACADEMIC RESULTS BASED ON COMPLEX SURVEY DESIGN

LVIV
2020

ABSTRACT

HRYTSYK, Kostyantyn. Analysis of the impact of different health-related habits on the academic results based on complex survey design. 2020. 7 p. Course Project – Introduction to Applied Econometrics, Applied Science Faculty, Ukrainian Catholic University. Lviv, 2020.

This work presents a study with the use of complex survey design focused on the health-related behaviors in high school grades (9-12) for analyzing the impact of different health-related habits on the academic success of pupils. The need for its realization arose from the personal observation of unsatisfactory student performance in relation to the small amount of sleep. In order to do so, we chose to use the Young Risk Behavior Survey that is conducted by the Centers for Disease Control and Prevention of the U.S. Department of Health and Human Services since 1991 every 2 years using complex survey design. These data are representative for all U.S. states and are eligible for visualizing distributions and relations between different variables, providing hypothesis testing and creating regression models for defining the significance of impact amount of sleep and other factors on academic results.

Keywords: Complex Survey Design. Health-related habits impact.

CONTENTS

1 INTRODUCTION

Nowadays, learning becomes a vital part of everyday life. Since global development temp is quite fast, people have to gain new knowledge and skills during life.

As school is the first stage of learning in each person's life, in this project, we decided to investigate which and how health-related habits affect academic results in high-school students (9-12 grades) in the United States of America.

For this project, R language was chosen as a programming tool. For handling complex survey design the R package `survey` was chosen. For data manipulation the R package `dplyr` was chosen. For data visualization the R package `ggplot2` was chosen.

As a data source Youth Risk Behavior Survey(YRBS) was chosen, particularly, the survey that was taken in 2017. One more aim of this paper is to use a complex survey design data (which YRBS is) in conducting analysis. The details of such an approach of creating surveys are described in chapter two, along with the description of data loading, cleaning, and manipulation is provided in chapter 2.

In chapter 3 conduct different stages of analysis are described such as:

- time-series analysis;
- visualizing distributions of different categories among respondents with histograms;
- visualizing distributions of academic results by different parameters such as doing sport, having breakfast and enough amount of sleep with stacked bar charts;
- hypothesis-testing (t-test, χ^2 test) for defining the significance of graphs above;
- apply linear regressions for modeling relations in visualizations and others.

In the concluding chapter, the short summary of the conducted analysis and obtained results are provided.

2 DATA LOADING AND CLEANING

2.1 Data Loading

As a data source Youth Risk Behavior Survey(YRBS) was chosen. Due to the inner complex survey design, such a data set is appropriate to the aim of this assignment. Here are some reasons why YRBS was chosen:

- It contains recent data, last available survey was conducted in 2017
- It covers majority of U.S. states;

Data for the research was loaded from the Centers for Disease Control and Prevention (CDC) official site. There is combined data set with all years of survey. There is a combined data set with all years of survey available. On the site, it is presented in several formats such as **.dat** and MS Access database **.mdb**.

In R there is a function `read_table` for handling **.dat** files, but **.dat** files provided by CDC return an error during loading.

So, .mdb files which contain data about each specific state in alphabetical order were chosen as a primary data source. To convert them in **.csv** format we used MS Excel to read the table in **.mdb** and save it as a comma-separated file.

The result of this operation you can find as files `A-M.csv` and `N-Z.csv`. All other descriptions of manipulation for the sake of data cleaning you can find in attached R file.

2.2 Data Cleaning

Originally, the YRBS survey contains answer for more than 130 questions. The next set of questions was chosen:

- **Q32:** During the past 30 days, on how many days did you smoke cigarettes?
- **Q42:** During the past 30 days, on how many days did you have at least one drink of alcohol?
- **Q46:** During your life, how many times have you used marijuana?
- **Q49:** During your life, how many times have you used any form of cocaine, including powder, crack, or freebase?
- **Q50:** During your life, how many times have you sniffed glue, breathed the contents of aerosol spray cans, or inhaled any paints or sprays to get high?
- **Q51:** During your life, how many times have you used heroin?
- **Q52:** During your life, how many times have you used meth-amphetamines?
- **Q53:** During your life, how many times have you used ecstasy?
- **Q54:** During your life, how many times have you used synthetic marijuana?
- **Q57:** During your life, how many times have you used a needle to inject any illegal drug into your body?

- **Q71:** During the past 7 days, how many times did you eat fruit?
- **Q72:** During the past 7 days, how many times did you eat green salad?
- **Q73:** During the past 7 days, how many times did you eat potatoes?
- **Q74:** During the past 7 days, how many times did you eat carrots?
- **Q75:** During the past 7 days, how many times did you eat other vegetables?
- **Q78:** During the past 7 days, on how many days did you eat breakfast?
- **Q79:** During the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day?
- **Q88:** On an average school night, how many hours of sleep do you get?
- **Q89:** During the past 12 months, how would you describe your grades in school?

The most important question in this set is the question 89 about grades in school. We will use it as a measurement for academic results. So, records from data set were chosen only when question 89 was answered and only since the first year when this question was answered which is 2001.

Since columns with answers on the question have names by the pattern ”q{question number}”, column names were changed for informative ones.

Also, were introduced dummy variables for regression modeling. They were created for different categories (race, grade, sex) and for indicating patterns in respondents' habits (sleep enough, used drugs).

In the last stage of data cleaning, we created a new data frame based on answers since 2017. Later, we compared distributions of the base categories to be sure that data distribution in 2017 is the same as in previous years.

2.3 Survey Design

Since YRBS is a complex survey design, an important moment in data loading is specifying how this design was constructed. The R package `survey` provides a huge variety of tools for creating and analyzing complex survey designs such as YRBS.

Firstly, we will explain how YRBS 2017 design was created. Among all U.S. states, 43 were selected. Also, 3 tribal surveys were included from Cherokee, Navajo, and Winnebago nations. And 21 big cities such as New York City, San Francisco, or Boston. Together 67 places where the survey was conducted. These places are called **stratum** in terms of complex survey design.

Each strata contains **Primary Sampling Units (PSU)**. They are respondents answer inside each stratum. Each pair of stratum and PSU create unique **id** for each response. And each PSU has a **weight** - number of individuals that this record presents and it is an alternative to sampling probabilities.

In research, design was created by function `svydesign`. In loaded data **id** is presented as column **PSU**, **strata** is presented as column **stratum** and **weights** is presented as column **weight**.

3 DATA ANALYSIS

3.1 Visualization

3.1.1 Time Series

Time series were constructed for the average of 4 values displayed in the table below. Change of value in each case isn't quite big, but still, we can see some tendencies:

Average value	Tendency description
Mark	The average mark is growing over the years. Since 2001, it has grown for 6%
Number of days having breakfast	We can see volatility over the years, but in 2017 it becomes close to the value of the base year.
Number of days doing sport	We can see volatility over the years, but in 2017 it becomes close to the value of the base year.
Amount of sleep	Against the average mark, the average amount of sleep is decreasing over the year. Since 2005, it decreased by 4%.

3.1.2 Histograms

Histograms were used to visualize distribution of respondents among different categories to get better understanding of data.

3.1.3 Stacked bar charts

The stacked bar charts were used to visualize percentage distribution of one category relatively to another.

3.1.4 Scatter plots

During research were used different approaches to visualization to get the answer to the question ”Does health-related habits affect academic results?”. Such graph type as a scatter plot of the relation of the mean amount of such values as amount of sleep, average number of days doing sport and the mean value of marks at school can show us is there any correlation between these values. Since all data is taken from US states, on scatter plots data is grouped by state name. Also, these relations will be checked later with regression modeling.

3.2 Hypothesis Testing

3.2.1 t-test

To conduct t-test new dummy variables were created, which was described in the previous chapter. α was chosen with value 0.1%. That means if p-value is less then 0.001/2=0.0005, we can reject H_0.

The results of t-testing are in the table below.

H_0 hypothesis	t-value	p-value	Reject H_0
There is no difference in marks by the enough amount of sleep	4.8334	1.557e-06	Yes
There is no difference in marks by having breakfast	9.9977	2.2e-16	Yes
There is no difference in marks by doing sport	9.0002	2.2e-16	Yes
There is no difference in marks by eating fruits	11.391	2.2e-16	Yes
There is no difference in marks by eating vegetables	6.7629	2.321e-11	Yes
There is no difference in marks by smoking	-16.33	2.2e-16	Yes
There is no difference in marks by drinking alcohol	-6.785	2.006e-11	Yes
There is no difference in marks by using drugs	-17.463	2.2e-16	Yes

What we can see is that all factors have a significant effect on marks and other variables. Important to point out that all bad habits affect badly on academic results. The interesting moment here is that the effect of smoking is quite close to the drag using.

All positive habits such as doing sport or sleeping enough have a significantly positive results on grades at school. Such conclusions will be proved in next sections.

3.2.2 χ^2-test

We used χ^2-test to investigate whether belonging to some category (race, grade, sex) affects the fact does a person sleeps enough or not. The main dummy variables for this test was `sleep.more.8.hours`, which has value 1 if a person sleeps 8 hours or more, and 0 otherwise. α is the same as for t-test.

The results of χ^2-testing are in the table below.

H_0 hypothesis	χ^2-value	p-value	Reject H_0
Enough amount of the sleep is not depend on age	1340.1	2.2e-16	Yes
Enough amount of the sleep is not depend on race	146.38	0.0001638	Yes
Enough amount of the sleep is not depend on grade	1410	2.2e-16	Yes
Enough amount of the sleep is not depend on sex	198.07	5.302e-10	Yes

In all tests we can reject H_0. That means that belonging to any of the categories used in the testing affect the fact does a person sleep enough. This fact confirms our stacked bar charts in section 3.1.3.

3.3 Regression Modeling

To test the effect of different factors on the academic results also was used linear regression modeling. Here, we will explain the results only of the one regression. The results of the rest you can find in the code file attached.

Formula: marks ~ sleep.more.8.hours + is_eating_vegetables + is_eating_fruits +

+ is_doing_sport + is_having_breakfast + is_smoking +is_drinking_alcohol + used_drugs

Coefficient	Estimate	t-value	p-value	significance
sleep.more.8.hours	0.01259	0.612	0.54	
is_eating_vegetables	0.12521	3.703	0.000225	***
is_eating_fruits	0.16833	7.082	2.72e-12	***
is_doing_sport	0.13682	5.735	1.30e-08	***
is_having_breakfast	0.17804	6.823	1.56e-11	***
is_smoking	-0.34264	-8.807	2e-16	***
is_drinking_alcohol	0.05783	2.402	0.016503	*
used_drugs	-0.36739	-14.075	2e-16	***

From the regression output, we can see the confirmation for almost all t-tests. What is different is the coefficient for drinking alcohol. Here it is positive, but the significance of it is quite low. So, we can't accept it as an evidence of positive effect of drinking alcohol on the academic results.

4 CONCLUSIONS

After conducted analysis we can state that health-related habits has an effect on the academic results for US pupils of 9-12th grade surveyed in terms of YRBS.

The most effective habits are having breakfast and eating fruits every day. Also, doing sport regularly and eating vegetables provide positive benefits for studying.

However, such an important activity as sleep doesn't make a significant difference in the marks at school.

On the contrary, such habits as smoking, drinking, and using drugs negatively affect grades at school.

Also, this research showed us that complex based survey is an effective tool for conducting surveys and carrying out analysis based on data from it.