Project: Learning from Data

Project Assignment 2nd Chance Exam

2022-2023

1 Background and Research Question

Smoking is known to have a negative effect on the overall health, particularly on the lungs, but also on e.g. the weight. On the other hand, physical exercise is known to have a beneficial effect on the BMI. In this study the focus is on the effect of smoking cigarettes and physical exercise on the body mass index (BMI) among children aged 3 to 19 years. The BMI is defined as weight (in kilogram), divided by the length (in meter) squared. The following table shows the BMI ranges that are considered as healthy (this is age-dependent).

age (year)	BMI (kg/m^2)
$[3-7]$	$[14-18]$
$[8-11]$	$[13-19]$
$[12-17]$	$[14-21]$
$[18-21]$	$[17-24]$

The primary research question is to assess the effect of smoking cigarettes and physical exercise on the BMI. Smoking information is available as a binary indicator (smoker / non-smoker). Information on physical activities, is given by two variables: Sport and SportDays (see the Data Dictionary for more information).

2 Study Design

A longitudinal observational study was set up to investigate the effect of smoking on the lung function among children. Families with young children in the East Boston area (USA) were asked to participate in this study. The data that will have to be analysed here, is a cross-sectional subset of data, containing the results of the latest available survey of each family. Only the data of one child per family is available (654 children).

3 Assignment

The assignment consists of the following items:

- Write a Statistical Analysis Plan (SAP).
- Write a final report on the data analysis.
  Note that the data analyses must be aligned with your SAP; if you deviate from your SAP, you must give a motivation. In this report you also need to reflect on ethical issues and on the stakeholders that may be involved.
  Submission deadline: August 17.
- Presentation of the report on August 21. Questions will be asked by the lecturers (oral exam). The time schedule for the presentations will be posted on August 18.

This is **individual work**.

4 Data

Disclaimer: the data used for this project, is inspired by the data of Tager et al. (1979), but it is not the original data set.

The data is available on BB.

4.1 Data Dictionary

Name	Type	Description	Values
Age	integer	Age of the child (in year)	3-19
FEV	real	FEV (litres)	3 decimals
Gender	integer	Gender of the child; indicator: 0=female, 1=mail	0,1
Smoke	integer	Smoker status of the child; indicator: 0=non-smoker, 1=smoker	0,1
height	real	height of the child (metres)	2 decimals
BMI	real	BMI of the child (kg/m^2)	1 decimal
SES	string	Socio-Economical Status of the family	"low", "middle" and "high"
ParentSmoke	integer	Smoker status of the parents; 0=noone is smoker; 1=at least one parent is smoker	0,1
Sport	real	Average number of hours per week of sport activities (averaged over previous year)	0 decimals
SchoolResults	string	School results of previous school year	"poor", "average" and "good"
T1D.	integer	Indicator for type I diabetes (0=no T1D, 1=T1D)	0,1
ColorBlind	string	Is the child color blind?	"yes", "no"
MotherEdu	string	Highest degree of the mother	"secondary school", "high school" and "university"
SportDays	integer	Average number of days of sport activities (averaged over previous year)	0 decimals
LungDisease	integer	Indicator of a lung disease; 0=no lung disease; 1=lung disease	0,1

Missing values are represented as NA in the datafile.

Reference

Tager, I., Weiss, S., Rosner, B., and Speizer, F. (1979). Effect of Parental Cigarette Smoking on the Pulmonary Function of Children, *American Journal of Epidemiology*, **110(1)**, 15-26.