Applied Statistics (MA30091) — Mock Project — 2022: NHANES: physical activity and diabetes

Set: 18.02.2022; Due: 28.02.2022

In this project you will analyse data from the US National Health and Nutrition Examination Survey (NHANES) 2005-2006. The data is a subset of survey participants aged between 21 and 59 years. The data can be downloaded from Moodle. To read in use the command:

load(file="NHANES0506.rda")

You can find information about NHANES in 2005-2006 at:
NHANES Questionnaire 2005-06

Your **report** must be no more than 4 pages long in total and should be divided into two parts.

1. A section aimed at members of the US/UK population, presenting and explaining your results. This section of the report should be readable without specialist statistical knowledge. This section should be no longer than one A4 page.

2. A technical section aimed at statisticians and epidemiologists, explaining exactly what you did, why you did it and what the conclusions were, in a manner that would allow the analysis to be repeated. By this I mean that the statistical structure of what you have done should be clear, so that it could be repeated using any statistical software the reader wanted to use. This means that models and results should be presented mathematically and with graphs or tables as necessary and not using computer code and output. When writing this section, you can assume the reader has read the first section.

Your report should address the following **questions**:

1. Glycohemoglobin was measured in participants to assess Diabetes mellitus. It reflects plasma glucose for the previous 120 days and is used to monitor diabetes. A glycohemoglobin value of 6.5% or greater indicates diabetes and values between 5.7% and 6.4% indicates pre-diabetes.

   Clinicians want to find out which are the risk factors for high levels of glycohemoglobin which are still below the threshold for diagnosis of diabetes/ pre-diabetes. That is, which of the variables are associated with glycohemoglobin in individuals who don't have diabetes or pre- diabetes? Describe the nature of the association.

2. Using a variable describing moderate to vigorous levels of physical activity, determine whether and how physical activity is associated with body measurements (bmi, waist circumference and height).

Variable description

The variables have been selected from a number of different NHANES data sets. For documentation of the variables (including factor levels, missing values etc) see the webpage given for the relevant data set. E.g. DEMO_D is the demographics data set for NHANES 2005-2006 and under this name you find the variables documented on the NHANES web-site. Note that you might have to define variables as factors.

Note that the physical activity variables given were generated from the raw minute-to-minute accelerometer data PAXRAW_D using the R package "nhanesaccel" [1], using the default settings. Thus, non-wear time was defined as any interval 60 minutes or longer in which all count values were 0. Days of monitoring with 600-1200 minutes of weartime were considered valid for analysis, and daily averages for participants with a minimum of 3 valid days are considered. Time spent in various activity intensities was assessed using standard cut-points (int.cuts) (sedentary: < 100; light 100-759; lifestyle 760-2019; moderate 2020-5998; vigorous >= 5999). Sedentary time accumulated in bouts of at least 30 minutes was defined as any interval of weartime 30 minutes or longer in which all count values were in the sedentary range. Bouted moderate-to- vigorous physical activity (bouted MVPA) and bouted vigorous physical activity (bouted VPA) were defined as any interval 10 minutes or longer in which all count values were in the MVPA or VPA range, respectively. Physical activity guideline minutes (PA guidelines) were calculated as the sum of bouted MVPA and bouted VPA. Because all bouted VPA is also bouted MVPA, bouted VPA minutes were given twice the weight of bouted MVPA minutes in this calculation, which is in line with the 2008 Physical Activity Guidelines for Americans [2].

Demographics http://wwwn.cdc.gov/nchs/nhanes/2005-2006/DEMO_D.htm

SEQN - Respondent sequence number
RIAGENDR - Gender
RIDAGEYR - Age at Screening Adjudicated
RIDRETH1 - Race/Ethnicity
DMDEDUC2 - Education Level - Adults 20+
DMDMARTL - Marital Status
INDFMINC - Annual Family Income

Glycohemoglobin http://wwwn.cdc.gov/nchs/nhanes/2005-2006/GHB_D.htm

LBXGH - Glycohemoglobin (%)

Body measurements http://wwwn.cdc.gov/nchs/nhanes/2005-2006/BMX_D.htm

BMXHT - Standing Height (cm) BMXBMI
- Body Mass Index (kg/m**2)
BMXWAIST - Waist Circumference (cm)

Triglyceride, LDL-cholesterol and Apoliprotein (ApoB)
http://wwwn.cdc.gov/nchs/nhanes/2005-2006/TRIGLY_D.htm

LBXTR - Triglyceride (mg/dL)

LBDLDL - LDL-cholesterol (mg/dL) LBXAPB - Apolipoprotein (B) (mg/dL)

Physical Activity Monitor

http://wwwn.cdc.gov/nchs/nhanes/2005-2006/PAXRAW_D.htm

valid_days - Number of days with valid_day = 1.
valid_min - Minutes determined to be valid weartime according to non-wear algorithm.
cpm - counts/valid_min.
sed_min - Sedentary minutes (counts < int.cuts[1]).
light_min - Light intensity minutes (int.cuts[1] <= counts < int.cuts[2]).
life_min - Lifestyle intensity minutes (int.cuts[2] <= counts < int.cuts[3]).
mod_min - Moderate intensity minutes (int.cuts[3] <= counts < int.cuts[4]).
vig_min - Vigorous intensity minutes (counts >= int.cuts[4]).
lightlife_min - Light-to-lifestyle intensity minutes (int.cuts[1] <= counts < int.cuts[3]).
mvpa_min -Moderate-to-vigorous intensity minutes (counts >= int.cuts[3]).
active_min - Active (i.e. non-sedentary) minutes (counts >= int.cuts[1]).
sed_percent, ..., active_percent - the above in percent, e.g. sed_min/valid_min
sed_bouted_30min - Sedentary minutes accumulated in bouts of length >= 30 min.
num_mvpa_bouts - Number of MVPA bouts per day.
num_vig_bouts - Number of vigorous bouts per day.
mvpa_bouted - MVPA minutes accumulated in bouts of length >= 10-min.
vig_bouted - Vigorous intensity minutes accumulated in bouts of length >= 10-min.

References

1. Dane R. Van Domelen, W. Steve Pittard, and Tamara B. Harris (2013). nhanesaccel: Functions for processing NHANES 2003-6 accelerometer data. R package version 1.0.

2. U.S. Department of Health and Human Services. 2008 Physical Activity Guidelines for Americans. Available at: www.health.gov/paguidelines. Accessed August 1, 2013.

General Instructions

The report should include no computer commands and no raw output. The minimum acceptable font size is 11pt and you must use A4 paper with normal margins (i.e. at least 20mm all round). In addition to the four-page report, you should submit an appendix of commented R code which could be used to exactly reproduce your results. There is no page limit for the appendix but note that the 4-page main report must stand alone and must not rely on references to parts of the appendix. The report should be in pdf format, and the appendix may be an R script (*. R) or an R markdown (*. Rmd) file.

To obtain feedback, the report and appendix should be submitted online via Moodle by 18:00, 28th February 2022.

I may answer generic questions about statistics and computing relevant to the coursework, but not specific questions about this specific analysis. To keep things fair, questions relevant to the whole class will not be answered individually but will be answered on a Moodle forum. Do not ask other members of staff, post-graduates or students for help.