

Data project - R course MSc Epidemiology - WS25/26

Introduction

The purpose of this data project is to allow you to apply the principles and methods introduced in the QM lecture, and also some of the concepts from the Epi Research Design lecture, to real-world data. The data we will use come from the National Health and Nutrition Examination Survey (NHANES) study, which is described below. During the project, your task will involve answering the questions listed below by familiarizing yourself with the dataset (summarized below), performing statistical analyses appropriate to the question at hand, and interpreting the results of these analyses. To complete the project, you need to submit a report describing your analyses and interpretations by **December 22th of 2025**.

About the NHANES study

The NHANES study is an ongoing program of the Centers for Disease Control and Prevention (CDC) designed to assess the health and nutritional status of adults and children in the United States. Every year, a new sample of the resident non-institutionalized US population is interviewed and examined. Data are collected on a wide range of topics, such as demographics, occupation, medical diagnoses, self-rated health, nutrition, substance use, environmental exposures, and blood parameters. The data are used, among other things, to determine the prevalence of major diseases and risk factors, to design health promotion campaigns, and to establish national reference values (e.g. for height, weight, blood pressure). You can find more information on the NHANES website:

<http://www.cdc.gov/nchs/nhanes.htm>

http://www.cdc.gov/nchs/nhanes/about_nhanes.htm

Comment:

The sampling procedures for the NHANES surveys are quite complex. They utilize a four-stage stratified probability cluster sampling method. Instead of conducting a simple random sample of the entire U.S. population—which would require a complete registry of all citizens and their addresses—researchers first randomly select a number of districts. Within those districts, they then sample neighborhoods, followed by households, and finally individual study participants. Additionally, researchers intentionally sampled a larger number of participants from specific subgroups, such as ethnic minorities, which might otherwise be too small to allow for meaningful analyses. These survey design features affect representativity and precision of the statistical estimates derived from the data. To fully account for this complexity, advanced statistical procedures would be necessary. However, our data project focuses on fundamental statistical techniques rather than advanced methods, so we will treat the data as if we have a simple random sample that is representative of the U.S. population.

NHANES data are released by the CDC to the general public in biennial cycles as datasets containing data from the previous two survey years. Our dataset is based on the 2011-12 cycle. The NHANES general public releases can be downloaded by anyone without prior registration from their website. (The data are stored in SAS XPORT format.) If you are interested, you can, for example, take a look at the original 2011-12 data, along with variable descriptions and questionnaires, by following this link:

<https://www.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2011>

Of course, all data are to be used for the purpose of statistical analysis only, not for identifying specific individuals. Any such effort would constitute a breach of confidentiality and is prohibited (see http://www.cdc.gov/nchs/data_access/restrictions.htm). This also applies to the dataset we will be using in the data project.

The data set

The dataset used for the data project is not identical to the NHANES 2011-12 dataset. In fact, there are several sub-datasets in NHANES, each containing different variables on various topics, such as nutrition and drug use. The data project's dataset has been assembled as a selection of variables from these subsets. Additionally, it includes some new variables that have been derived from the original NHANES variables, for example, by combining values from several variables.

The table below provides an overview of the variables contained in our dataset. The variable name is used to refer to the variable in R. The variable label describes what is measured by the variable. For categorical variables, value labels are provided which explain the meaning of the variable values. "Data type" tells you what kind of values the variable can take ("boolean" means that the variable is a binary "TRUE-FALSE" variable; note that for these variables you can infer the meaning of the values from the wording of the variable label, which is always a question). For those of you who want to look into the original data: If the variable corresponds directly to a NHANES variable, the corresponding NHANES name is given in the last column. You can find the original dataset by entering the variable name in this search interface and choosing the appropriate data release cycle: <http://wwwn.cdc.gov/nchs/nhanes/search/default.aspx>. "Derived" means that the variable is calculated from NHANES variables in a more complex way.

Variable list

Variable name	Variable label	Data type	Value labels		NHANES
seqn	Subject identifier	integer			SEQN
cd	Blood cadmium level (nmol/l)	real			LBDDBCDSI
pb	Blood lead level (umol/l)	real			LBD8PBBSI
hg	Blood mercury level (umol/l)	real			LBDTHGSI
hdl	Serum HDL cholesterol level (mmol/l)	real			LBDHDDSI
hivpos	HIV test result positive?	boolean			LBDHI
weight	Weight (kg)	real			BMXWT
height	Standing Height (cm)	real			BMXHT
bmi	Body Mass Index (kg/m ²)	real			BMXBMI
rr_sys	Systolic blood pressure (mm Hg)	real			Derived
rr_dia	Diastolic blood pressure (mm Hg)	real			Derived
srhgnrl	Self-rated health in general	integer	1	'Excellent'	HSD010
			2	'Very good'	
			3	'Good'	
			4	'Fair'	
			5	'Poor'	
srphbad_prv30d	Bad physical health in prev. 30 days (self-rated)	integer	1	'On no day'	Derived
			2	'On 1-14 days'	
			3	'On more than 14 days'	
srmhbad_prv30d	Bad mental health in prev. 30 days (self-rated)	integer	Same as previous variable		Derived
adlimp_prv30d	Activities of daily life impaired due to bad health in prev. 30 days (self-rated)	integer	Same as previous variable		Derived
age	Age at the time of interview (yrs; ages >80 coded as 80 to maintain anonymity)	integer			RIDAGEYR
educ	Educational level	integer	1	'Less than 9th grade'	DMDDEDUC2
			2	'>= 9th grade but no high school diploma'	
			3	'High school diploma'	
			4	'Undergraduate degree'	
			5	'Graduate degree'	
martlst	Marital status	integer	1	'Married'	DMDMARTL
			2	'Widowed'	
			3	'Divorced'	
			4	'Separated'	
			5	'Never married'	
			6	'Living with partner'	
			77	Refused to answer	
			99	Don't know	
male	Male gender?	boolean			RIAGENDR
ethnic	Race/Ethnicity	integer	1	'Hispanic'	Derived
			2	'White'	
			3	'Black'	
			4	'Other/Mixed'	
incred	Family income relative to poverty line (PL)	integer	1	'< PL'	Derived
			2	'PL – 2*PL'	
			3	'2*PL – 4*PL'	
			4	'> 4*PL'	

asthma_ever	Ever diagnosed with asthma?	boolean	MCQ010
asthma_now	Have asthma now?	boolean	MCQ035
ovrwhgt_ever	Ever diagnosed with being overweight?	boolean	MCQ080
arthrit_ever	Ever diagnosed with any kind of arthritis?	boolean	MCQ160A
stroke_ever	Ever diagnosed with stroke?	boolean	MCQ160F

Data project questions

1. Familiarize yourself with the NHANES study design and the variables in the dataset by reading the sections “About the NHANES study” and “The data set” of this document. Use the information provided to answer the following questions:
 - a) What are important characteristics of the NHANES study (study design, target population, study objectives, study period, ...)?
 - b) What general categories (such as “demographics”, “lab results”) can the variables in the dataset be sorted into?
2. Using the subsample you have chosen, describe the US population with regard to
 - a) Demographic characteristics. How can the strange age distribution be explained? To deal with this problem, recode the age variable into the following categories: 18-34, 35-49, 50-64, 65-79, 80 or higher.
 - b) When asked about marital status, some of the participants refused to answer, while some didn't know which category they belonged to. Hence, they were coded differently. Take that into account and recode the variable `marlst`.
 - c) Self-rated health. Looking at the results of your descriptive analysis, what do you have to consider when interpreting the results?
3. Diabetes and ethnicity
 - a) How is the diabetes status distributed in your data set? Summarize the two diabetes groups into a single group, and recode the variable for diabetes status into two groups: non-diabetes and diabetes. Give an interval estimate for diabetes in your data set.
 - b) Work with the recoded variable. Use an appropriate statistical test to test the relation between diabetes status and ethnicity. Interpret the results of the test.
4. Weekly working hours and self-evaluated health status
 - a) The variable `hrsworked_prvwk` codes the number of weekly working hours. 77777 or 99999 are missing values. Recode these missing values with NA.
 - b) Recode the variable `hrsworked_prvwk` into categories $[,40)$, $[40,)$ weekly working hours. Number or name the categories properly. (Hint: Set the new variable as a factor variable, if necessary)
 - c) Plot weekly working hours against self-evaluated health status using mosaic plot. Looking at the plot, what would you say about the relation between these two variables?
 - d) Use an appropriate statistical test to test for the statistical significance of this relation. Interpret the results of the test.
5. Blood mercury level (umol/L) and sex
 - a) Describe and plot the distribution of blood mercury level in men and women.
 - b) Use both parametric and non-parametric statistical tests to test for the statistical significance the relation between blood mercury level and sex. Interpret the results of the tests.
6. BMI and HDL
 - a) How strong is the relation between BMI and HDL (the “good” cholesterol)? Is it statistically significant? How much of the variation in HDL can be explained (in a statistical sense) by variation in BMI?

- b) Does the relation between HDL and BMI change when you adjust for age (categorized)? Interpret the coefficients of the resulting model. (When you mean-center BMI before fitting the model, you can also interpret the intercept). Does BMI have a clinically relevant impact on HDL according to your model?
- c) Try to find a better model to predict HDL by including more covariates. Select a number of candidate covariates which, in your opinion, may be related to HDL, and then choose a model selection strategy and a criterion/test for comparing models. Describe the model with the best fit according to your search, and interpret the model coefficients.

7. Cancer

- a) Estimate the lifetime prevalence of cancer. Can you also give an interval estimate?
- b) What are the prevalence estimates in those who were exposed to pollutants at work for a longer time period, and in those who were not? Is there a statistically significant difference in prevalence between these two subgroups?
- c) Adjust for age in the relation between lifetime diagnosis of cancer and exposure to pollutants, using the categorized age variable. (Note: No information on pollutant exposure was collected from participants aged 80+, so these cannot be included in the analysis.) Does the adjustment for age change the picture? Interpret the model coefficients, including the intercept.
- d) Try to find a good model of cancer diagnosis, describe and interpret it, as you did for HDL.