MA30091 : Applied Statistics 2023/2024

Coursework 1: Health Survey England (2019) Data Analysis

Set: 14:00, 1st March 2024

Due: 14:00, 15th March 2024

Estimated time required: The project should take not more than 10 hours work by each group member.

Conditions: This is a group coursework.

Marks: 50 (contributes 50% to final grade for the unit)

Length: Your report must be **no more than 5 pages long** including graphs and tables. You may have an extra page (5+1) for the title of your report. The plots and tables can be interspersed within the report. Your report should be divided into two parts:

a) A section aimed at members of the general public, presenting and explaining your results. This section of the report should be readable without specialist statistical knowledge.

b) A technical section aimed at statisticians, 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, it is meant 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. When writing this section, you can assume the reader has read the first section.

The report should include no computer commands and no raw output. The minimum acceptable font size is 11pt.

In addition to the five-page report, you should submit an appendix of **commented** R code which could be used to exactly reproduce your results. This will not be marked for credit, but it might be checked to see if it is unclear what you have done in your report, so ensure that it is intelligible and does not contain redundant material. There is no page limit for the appendix, but note that the five-page main report must stand alone, and must not rely on references to parts of the appendix.

Submission: The report and appendix should be submitted online via Moodle by **14:00, 15th March 2024**. The report should be in pdf format, and the appendix may be either an R script (*.R) or an R markdown(*.Rmd) file. Please don't mention the names of the group members while submitting, only the candidate numbers should be included in the report.

Support and advice: Your lecturer(s) may answer generic questions about statistics and computing relevant to the coursework, but not specific questions about specific analyses. 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.

Feedback: You will receive feedback within a maximum of three semester weeks following the submission deadline. The feedback will consist of your marked work and an overall feedback document commenting on the assessment across the cohort.

Late submission of coursework: If there are valid circumstances preventing you from meeting the deadline, your Director of Studies may grant you an extension to the specified submission date, if it is requested before the deadline. Forms to request an extension are available on SAMIS.

- If you submit a piece of work after the submission date, and no extension has been granted, the maximum mark possible will be the pass mark.
- If you submit work more than five working days after the submission date, you will normally receive a mark of 0 (zero), unless you have been granted an extension.

Academic integrity statement: Academic misconduct is defined by the University as "the use of unfair means in any examination or assessment procedure". This includes (but is not limited to) cheating, collusion, plagiarism, fabrication, or falsification. The University's Quality Assurance Code of Practice, QA53 Examination and Assessment Offences, sets out the consequences of committing an offence and the penalties that might be applied.

Recommended group work policy

- Agree on how you will work together. There are many ways to work together efficiently - the University recommends Microsoft Teams. You can regard this as an opportunity to practice an important business skill.
- Make an early start to allow time for revising and proofreading your work.
- Please be considerate of the other group members - be responsive and do your fair share. If some members of the group do not contribute substantively to the work, participating members may submit the coursework independently. Notify me and the non-participator if you intend to do this. Keep a record of communications so the claim can be verified. Participating group members will not get more credit but verified non-participants will receive none.

Generative AI

GenAI tools may be used while doing the analysis or writing. Under the University's Academic Integrity Statement, you 'must not present content created by generative AI tools as though it were your own'. Any text or code produced by genAI must be checked for correctness and cited. In addition, you must include a short statement (max 250 words) at the beginning of your submission indicating: what tools you used and how you used them OR that you have not used genAI. You should be prepared to explain anything in your submission to an examiner if asked to do so.

The GenAI statement can be part of the title page.

Background

The Health Survey for England (HSE) started in 1991, and is a series of cross-sectional annual surveys sponsored by the Department of Health. It collects information about the health of people living in England and combines questionnaire-based answers from a face-to-face interview with physical measurements taken by a trained nurse and the analysis of blood samples.

The data for this project come from the 2019 HSE. Participants were selected using a multi-stage stratified random sample. All persons at the selected address were eligible for interview. All adults were selected for interview.

A nurse visit was arranged for all participants who consented; this included measurements and the collection of blood and saliva samples, as well as other questions. A total of 8205 adults aged 16 and over and 2095 children aged 0-15 were interviewed, including 4947 adults and 1169 children who had a nurse visit.

Task

In this coursework you will analyse data related to health, age, socio-economic factors and lifestyle habits in adults (from the age of 16) from the population in England.

Your report should address the following issues:

1. Provide estimates and confidence intervals of the prevalence of smoking, e-cig smoking and alcohol consumption in the population.
2. Using an appropriate model describe to what extent is smoking associated with socioeconomic status and age.
3. Using an appropriate model comment on the lifestyle habits that are associated with systolic blood pressure and how are they associated? (Hint: Note that the appropriate model may not only involve lifestyle habits as predictors)

All arguments should be quantitative and using the techniques from the course.

The file hsesub.Rdata (available on moodle) contains a subset of variables of the 2019 HSE data. The variables are listed below. Supporting files are provided on the moodle unit page. These are the full dataset, a script readdata.R which shows you how I created hsesub.Rdata and the codebook for the variables 'hse_2019_eul_20211006_ukda_data_dictionary.rtf'.

You are not expected to impute missing values. For modelling, you may assume that each observation has equal weight. You may need some exploratory analysis before choosing the appropriate model. Not all the variables will be useful or even sensible to include, and you may recode variables, or even derive new ones, where appropriate.

Variable description

Missing values are coded as NA and include categories Refused, Don't know, Refused/not obtained, and Not applicable.

SerialA Archive respondent serial number

wt_int The interview weights are a combination of the household weight and a component which adjusts the sample to reduce bias from individual non-response within household

Sex Gender, male (coded 1), female (coded 2)

Age35g Respondent age - grouped, approx 3 year bands for 0-15, 5 year bands 16+

ag16g10 Age 16-75+ in ten year age bands

topqual2 Highest Educational Qualification - students separately coded

marstatD Marital status including cohabitees coded

qimd19 Quintile of index of multiple deprivation (IMD) score 2015 - least deprived to most deprived. This is an indicator of socio-economic status.

urban14b Rurality of dwelling unit (urban/rural) - binary

origin2 grouped ethnic categories

cigdyal_19 Number of cigarettes smoked a day - inc. non-smokers

cigsta3_19 Cigarette Smoking Status: Current/Ex-Reg/Never-Reg

NDPNow_19 Current use of E-cigarettes and/or nicotine delivery product (NDPs), (16+yrs)

drinkYN_19 Drink alcohol in last 12 months - binary

dnoft_19 Frequency drank any alcoholic drink last 12 mths

d7many3_19 Number of days drank in last week, including none

GOR1 Government Office Region - numeric

BMIval Valid BMI measurements using estimated weight if >130kg

omsysval Valid mean Systolic blood pressure

You may consider the variables **Sex**, **topqual2**, **marstatD**, **qimd19**, **urban14b**, **origin2**, **GOR1** as representing socio-economic status and the variables **cigdyal_19**, **cigsta3_19**, **NDPNow_19**, **drinkYN_19**, **d7many3_19**, **BMIval** as representing lifestyle habits.

Teaching access agreement

In order to submit the coursework you have to sign the access agreement. The access agreement is available as an activity in the moodle which you need to agree with by selecting the appropriate option and that will allow you access to the coursework dataset. By signing the form you agree that you will use the data only for the purpose of this coursework and delete the data after the project is finished.

References and resources

Full documentation of the complete HSE 2019 data: https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=8860

Full citation for the dataset: NatCen Social Research, University College London, Department of Epidemiology and Public Health. (2023). Health Survey for England, 2019. [data collection]. UK Data Service. SN: 8860, DOI: http://doi.org/10.5255/UKDA-SN-8860-1

Marking Scheme

The project will be marked out of 50. There is no single correct analysis for this type of project, so you will not be marked on the basis of how close you get to some particular model answer. The marks are not

subdivided, but will be allocated on a combination of statistical approach and justification, interpretation of results in context and presentation.

35 - 50 (First)

A report that could be presented with little or no revision. Analysis is sound so that conclusions are well-supported statistically. Interpretation is reasonably mature. The project should demonstrate a clear overview of the work, without getting lost in details, and be free of all but minor statistical errors.

30 - 34 (2.1)

A report that could be presented after a round of revision, but without having to re-do much of the actual analysis. Some flaws in the analysis or presentation (or minor flaws in both), but basically sound. A good grasp of the statistics and context, so that interpretation is reasonable.

25 - 29 (2.2)

Major re-working required before the report could be presented, but containing some sound statistics demonstrating understanding of statistical modelling and its application. Reasonable presentation and organisation.

20 - 24 (Third)

Major flaws in analysis and presentation, but demonstrating some understanding of statistics, and a reasonable attempt to present the results.

0 - 19 (Fail)

Flawed analysis demonstrating little or no understanding of statistics, and/or incomprehensible or overly bad organisation/presentation.

Contact details:

Dr. Sandipan Roy
Room: *6 West 1.06*
E-mail: *sr2081@bath.ac.uk*