BIOSTATISTICS 650 GROUP PROJECT

**Analysis of the National Health and Nutrition Examination Survey Data**

**Background:**

The National Health and Nutrition Examination Survey (NHANES) program at the Centers for Disease Control and Prevention (CDC) is a cross-sectional survey study designed to assess the health and nutritional status of adults and children in the United States. It began in the 1960s and since 1999 the survey moved to continuous data collection in two-year cycles. Approximately 5,000 individuals of all ages are interviewed in their homes every year and complete the health examination component of the survey. The sample for each two-year cycle is representative of the NHANES target population, which is the non-institutionalized civilian resident population of the United States. NHANES is used to determine the prevalence of major diseases and risk factors for diseases, as well as to assess nutritional status and its association with health promotion and disease prevention. NHANES findings are also the basis for national standards for measurements such as height, weight, and blood pressure. You can read more about NHANES on the [CDC’s website](http://www.cdc.gov/nchs/nhanes/).

The CDC uses sampling strategies ([2007-2010](https://www.cdc.gov/nchs/data/series/sr_02/sr02_160.pdf), [2011-2014](https://www.cdc.gov/nchs/data/series/sr_02/sr02_162.pdf)) to purposefully oversample certain subpopulations like racial minorities. Naive analysis of the original NHANES data can lead to mistaken conclusions because the percentages of people from each racial group in the data, for example, are quite different from the way they are in the target population. In this project, you will use the “NHANES” data from the R package “NHANES”. Please refer to [this file](https://cran.r-project.org/web/packages/NHANES/NHANES.pdf) for variable definitions and usage of the package/data. This dataset contains 10,000 individuals resampled from the original 2009-2010 and 2011-2012 NHANES data and can be treated as if it were a simple random sample from the US population. Please investigate the association between X (predictor of interest) and Y (outcome of interest). You will determine what X and Y are, as well as your own scientific question(s) of interest. Note that the outcome of interest can be either continuous or discrete as long as you can justify why using linear regression is appropriate for answering your scientific question.

**Miscellaneous comments:**

1. As you may have noticed, the research question is left open. Albert Einstein said “The formulation of a problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill”. It is a common mistake for statisticians to rush into a complex analysis without paying attention to what the objectives are and whether the data are appropriate for the proposed analysis. Hence for this project, you will start with **defining your own scientific question**. Few or none of you is an expert in health and nutrition. I am not an expert. Extensive literature review is beyond the scope of this project. A limited literature review may be warranted, particularly to justify your scientific question and choice of adjustment variables.

2. One of the main things I look for is whether you justify your choices. For example, if you use a regression model and adjust for variable W, I would like to know why you adjusted for variable W (something you might not typically see in most published scientific papers). I don’t need to agree with you as long as your reasoning is sound.

3. Beware of “fishing expeditions” -- if you torture the data long enough, it will confess to anything (Ronald H. Coase). You’ll almost always find something but that something may just be a coincidence. Hence, it is strongly recommended to use a pre-defined analysis plan and stick to it. When you determine your analysis plan, note that it’s fine to have more than one model. Usually different questions require different analyses, and sometimes a single question even requires multiple models. Clearly lay out your rationale for each model/analysis.

4. The report should include an abstract, introduction, methods, results, and conclusions, like many scientific papers. In the methods section, please clearly describe and justify your choice of outcome(s), predictor(s) of interest, and covariates (something you might not typically see in most published scientific papers); please also clearly describe your statistical analyses. An example list of contents is on the last page. I will also provide a sample report. Note that the sample report is just an example of a good structure (what to include) for a group project report. Please don’t spend time on the statistical analyses in the sample report as some of the analyses in the report are beyond the scope of the class. The tables and figures in the sample report are not necessarily of publication quality – please search for high quality papers published in top medical journals and mimic the way tables and figures are presented in these papers.

5. The report should be a maximum of ten pages including figures and tables, not including bibliography. After the bibliography, please add another page that describes group member contributions. The author contribution page is required by certain scientific journals and doesn’t count towards the page limit. All text should be 12-pt font, double-spaced, and left-justified. Although recommended against, you can use a smaller font and line space for a table or figure. The first line of every paragraph should be indented. All page margins (top, bottom, left, and right) should be 1 inch. Along with the report (pdf file), an executable R Markdown file (or SAS code) should be submitted, with detailed comments to make sure that your codes are self-contained and readable. Lastly, I’d like to recommend a very helpful article written by Dr. Rod Little: "Some style and grammar tips for biostatistics and statistics students" ([link](https://docs.google.com/a/umich.edu/viewer?a=v&pid=sites&srcid=dW1pY2guZWR1fHJvZC1saXR0bGV8Z3g6MmQzZjVhM2VjOWVlN2Q0Mw)). Below I quote one paragraph:

*Less is more! Scientific writing should be clear and concise. When you write anything, go over it carefully and attempt to reduce the length by 20% without reducing content, simply by more concise wording, removing adjectives that do not add anything, and avoiding repetition.*

7. Finally, you will **present your results in class (combining two sessions) on December 6th and 8th**. Please turn in your presentation slides to Canvas **before 5pm on December 5th**. Once the slides are submitted, no more changes to the slides can be made (that is, the slides presented on the presentation days must be the same as the slides submitted to Canvas). Each group has 12 mins, and every group member should speak during the presentation. The presentation slides should be no more than five pages (not including title page and thank-you page if you have these, although these are not required). A series of slides adding one element at a time constitutes one page, e.g., if you would like to have an animated slide that draws a figure piece by piece. During the presentation, each group will share their own screen and the instructors will be a backup in case there is a technical issue for screen sharing. The order of presentation is randomly assigned (except for students currently not in the US time zone).

8. Please turn in your final report and R Markdown file (or SAS code) to Canvas **before 11:59pm on December 10th**. Late submission is not accepted. Please label your files with your group number and no names. Please also label the file that you upload with your group number, e.g. “Group1Report.pdf” for Group 1.

9. You can discuss the project only with your group members, the instructor, and the GSIs. All writing should be that of your group. Your analysis cannot be a copy of any analysis already published in the literature or posted on the internet.

10. This group project will be graded on a 60-point scale: 20 points for statistical and scientific content, 20 points for the quality of the report, and 20 points for the quality of the group presentation.

11. You may find that there are potential data errors. Please make your own decision on how to address these issues and justify your approach.

**Your title**

**Abstract**

Background

Methods

Results

Conclusions

**Introduction**

**Methods**

Study population

Variables

*Predictor(s) of Interest*

*Outcome(s)*

*Covariates*

Statistical Analyses

**Results**

**Conclusions and Discussion**