207.4441 Applied Statistics I, first semester 2020-21

Assignment 1: U.S. population nutrition survey

Report due at the beginning of class on November 15

The United States Department of Agriculture conducted a survey in 1995 and 1996, with a representative sample of U.S. households, to examine the relationships among dietary habits, knowledge and attitudes, and health. The data for the project, a subset of the full sample comprising 4036 observations, appear in four sheets of the Excel file called nutrition\_data.xlsx. Descriptions of the variables appear in the file nutrition\_vars.rtf, and full information about the survey questions whose names start with “kq” is given in the file scale\_vars.pdf.

Goal: To examine the variables affecting the BMI of the respondents.

All of the following tasks are very important…but the last two will require the most thought.

1. Combine the four sheets (main and scale, males and females) into a single R data frame. (Hints: You can save the sheets as four .csv files and read these in to R using read.csv(). Use merge() to combine the two sets of variables, and rbind() to combine males and females.)
2. Use descriptive methods (frequency tables, histograms, the pairs() function, etc.) to check the integrity and completeness of the following variables: region, urbanization, income, age, sex, race, education (highest grade completed), the five “diet” variables, exercise frequency, and self-reported weight status. Don’t forget to convert categorical variables to “factors”. (Note: In “real life,” questions about the data should generally be referred to the investigator to determine if there are errors in the data and, if so, whether these can be corrected.)
3. For each variable, values ​​that specify missing data ​​(e.g., “indeterminate”, “not ascertained”) should be changed to NA.
4. Check the distribution of the response variable and whether there is a need for a transformation.
5. Use plot() and boxplot() to examine whether BMI appears to be related to the other variables listed above, and as an additional check for anomalies in the data.
6. Regress BMI on each of the other variables individually.
7. Build a multiple regression model for BMI. Explain the choice of variables in the final model and assess the model fit. There is no one “correct” approach here, but you will need to justify your approach. Some strategies you may consider include:
   1. “Screen” variables based on the results of the separate regressions for each predictor—i.e., only predictors found to be significant in separate models are candidates for inclusion in the multiple regression.
   2. Forward selection, backward elimination, or both, e.g., using stepAIC().
   3. Consider interaction effects.
8. Summarize the results in a report that includes a brief introduction to the problem; a description of the data set; an explanation of the data analysis methodology; appropriate text, tables and figures (not raw R output!) to present the results, and a brief conclusion. Attach your R code as an appendix.