R package for observational studies

Objective

Observational studies are often used for assessing the prevalence of a certain exposure or outcome and evaluating association between the exposure and the outcome. Many packages are created to solve some problems in the analysis while there is no specific package to do an integrated analysis for observational studies. To make it more convenient for researchers and scholars to generate complete and well-formatted results, this package is built to perform the whole EDA process and provide related summary tables and graphs.

Dataset

For the example dataset, the NHANES dataset which is contained in the package NHANES in R is used. This data was originally assembled by Michelle Dalrymple of Cashmere High School and Chris Wild of the University of Auckland, New Zealand for use in teaching statistics. The outcome variables can be either binary variables or continuous variables. For binary outcome variable, the indicator of whether patients have diabetes or not is used. For continuous outcome variable, the patients' total HDL cholesterol in mmol/L is used as an example. The covariates will be determined by the users, in the example, BMI, exercise levels, alcohol intake and smoking are used for both univariate regression and multivariate regression with potential confounding variables.

Brief Description

4 functions are developed in the package to deal with the most common data analysis in the observational analysis. For the first function, the incidence, incidence rate, hazard rate, incidence proportion, and period prevalence during 2009 to 2010 and 2011 to 2012 are calculated and output in a table. The second function is designed to generate the descriptive statistics table with numerical summaries of the total number of cases, arithmetic mean and percentiles. For the third function, it generates results for assumption checking and univariate regression results, i.e. tables for outputs from univariate regression with measurement such as risk ratio, odds ratio, hazard ratio, rate difference, risk difference and incidence difference, graphs and potential analysis results. For the fourth function, it outputs the assumption checking results and summary tables and graphs for multivariate regression with potential model selection and mediation regression results for checking confounding variables.