

MATH 598 - PROJECT 4

*Please upload a pdf to myCourses by
10.00pm on Monday 30th April.*

This project is based on an analysis of a Medical Expenditure Panel Survey (MEPS) data set covering the year 2013 (data set labelled h163) that can be loaded and simplified using the following code:

```
library(foreign)
download.file("https://meps.ahrq.gov/mepsweb/data_files/pufs/h163ssp.zip",
             temp <- tempfile())
unzipped_file<- unzip(temp)
h163 <- read.xport(unzipped_file)
unlink(temp)
my.h163<-data.frame(h163$TOTEXP13,h163$AGE13X,h163$REGION13,h163$SEX,
h163$RACEV1X,h163$MARRY13X,h163$EDRECODE,h163$POVCAT13,h163$FAMINC13,
h163$HIBPDX,h163$STRKDX,h163$CHOLDX,h163$DIABDX,h163$ASTHDX,
h163$ARTHDX,h163$BMINDX53,h163$ADSMOK42,h163$EMPST31,h163$INSCOV13)
```

The codebook for the data set, explaining the different codings for the variables, can be found here:

https://meps.ahrq.gov/data_stats/download_data/pufs/h163/h163doc.pdf

For the purposes of this project, these data may be treated as a random sample from the US population. The data set comprises 36940 records.

In the data frame `my.h163`, 19 variables have been extracted from the general MEPS data. The research question focusses on the relationship between

- total medical expenditure (in US dollars) in 2013, `TOTEXP13`, the *outcome*, and
- having a diabetes diagnosis, `DIABDX`, the *exposure*,

on the basis of these data. The codebook entry for the `DIABDX` variable reads

DIABDX indicates whether each person (aged 18 or older) had ever been diagnosed with diabetes (excluding gestational diabetes). Persons aged 17 or younger were coded as "Inapplicable" (-1). The age of diagnosis for diabetes (DIABAGED) is included in this file. This variable is top-coded to 85 years of age.

Objective: Perform and report on an analysis using causal adjustment and other statistical procedures to quantify the average (causal, unconfounded) difference in expenditures between individuals who did (`DIABDX==1`) and did not (`DIABDX==2`) have a diabetes diagnosis.

Note that the data set contains missing values for some variables. Please note in your report how these missing values are treated in your analysis.

25 MARKS