As part of the first year qualifying exam, Biostatistics PhD/MHS/ScM students conduct a detailed statistical analysis of a data set to address a scientific question. The purpose of this project is to give you experience doing such an analysis and feedback that you can use to enhance your skills and performance of the first year evaluation. This project is optional, but you are encouraged to do it if you think it will benefit your development as a data analyst.

The goal of this analysis is to estimate the fraction of total medical expenditures among smokers with coronary heart disease and other similar diseases (similar in that their smoking relative risk is roughly the same) that can be attributed to their having smoked.

The medical model that underlies this question can be expressed

Smoking -> major diseases -> Medical expenditures

Note we are only interested in quantifying the effect of smoking on medical expenditures that results from the additional CHD that it causes.

Attached find a copy of data from the National Medical Expenditure Survey that is a representative sample of non-institutionalized U.S. adults and includes data on smoking, disease and medical expenditures.

Smoking is represented by 2 variables: total pack years smoked and time since quit for prior smokers.

CHD is a binary indicator of whether the person has a confirmed diagnosis of one of the CHD cluster of diseases during the current year.

Other potential confounding or effect-modifying variables include: age, gender, education and SES.

Your analysis should produce an estimate of the expected medical expenditures for each past or current smoker with a CHD if that person had never smoked. The attributable expenditures are the difference of their actual expenditures and these expectations. Summarize the attributable expenditures over all people and by gender and age groups.

To achieve this final goal, you must conduct separate analyses of how smoking effects the rate of CHD and then how the presence of CHD effects expenditures.
Conduct your analyses, and then write a 4-page summary (including tables and graphs) organized into the following sections:

- Executive summary (less than 1 page)
- Problem statement
- Risk of CHD
- Expected expenditures
- Attributable medical expenditures

The report should be written for a public health journal.

Good luck!