Final Take Home Exam
Due: Friday April 28, 2023 at 4P

1 Instruction

- This exam is "open book," which means you are permitted to use any materials handed out in class, your own notes from the course, the text book, internet, and anything on the course website.
- The exam must be taken completely alone. Showing it or discussing it with anyone is forbidden.
- You may not consult with any other person regarding the exam. You may not check your exam answers with any person.
- In this final take home exam, you will conduct your analyses and then write a 4-page summary (including tables and graphs). I will grade this final exam based on the 4-page summary report. **The closer the report looks like a published paper, the higher the score.** According to the syllabus, the grade for this class will be calculated as follows: homework (50%), midterm-exam (25%) and this final take home exam (25%).
- Please hand in a print-out of your report and also email your R code to me (hoyen@stat.sc.edu).
- To make your report looks more like a paper, you can use the **echo=FALSE** option to prevent code from appearing in the finished file. In the beginning of your R Markdown file, set up the code chunk option as follows:

```
```{r setup, include=FALSE, echo=FALSE}
knitr::opts_chunk$set( warning = FALSE, message=FALSE, echo=FALSE,
                       fig.height = 4, fig.width = 8)
library(knitr)
```
```

2 Relative Size of Medical Expenditures among Persons With- vs Without Major Smoking-caused Disease

Medical Expenditure Panel Survey (MEPS) is a set of large-scale surveys of families and individuals, their medical providers (doctors, hospitals, pharmacies, etc.) and employers across the United States. MEPS collects data on the specific health services that Americans use, how frequently they use them, the cost of these services, and how they are paid for, as well as data on the cost, scope, and breadth of health insurance held by and available to U.S. workers. We will use the MEPS data collected during 2009 to determine the societal "costs" of smoking and estimate the smoking-attributable costs of treating major smoking-caused disease.

We will employ a two-part models: a key characteristic of medical costs is that a non-trivial fraction of persons have zero costs and the remaining have a distribution of costs that are highly skewed to the right. Data analyst of medical expenditures favors "two part models" that comprise:

(1) A model for the probability of any expenditure (p)

(2) A second model for the size of the expenditures ($) given some occurred (S)

Your analysis should produce an estimate of the expect medical expenditures for each smoker with MSCD (major smoking caused disease) if that person had never smoked. The attributable expenditure 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 MSCD and then how the presence of MSCD effect 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 MSCD
- Expected expenditures
- Attributable medical expenditures

The report should be written for a health-related journal. You can consider the following guide for the analysis.

3 Guide for Analysis

First focus on the second component, a model for the size of the expenditures given a non-zero value occurred. Then build the first component later and put the two parts together to estimate expenditure. Other potential confounding or effect-modifying variables include:

- Demographics variables: age (AGE0: age in years), gender (Gender: Male=1; Female=0)
- Socioeconomic status (SES): education(Educyr: years of education), poverty (Poverty: Poor=1; Not poor=0)
- Disease: major smoking-caused diseased (MSCD: Yes=1; No=0).

Analysis guide:

1. Under the two part model. Show that give an estimate of each component for a person, their expected expenditure are E($)=pS
2. Take the individuals with **positive expenditure**; log transform the medical expenditure and create the lexp variable such that $lexp = log_{10}(expenditure)$. Fit a **good** model based on lexp.
3. Assuming lexp following normal distribution

$$\log(expenditure) \sim N(\mu, \sigma^2)$$

   What is the mean expenditure?

4. Complete the table below. Δ is the difference of **mean** positive medical expenditure with vs. without MSCD. Std error and 95% CI are standard error and 95% confidence interval of the corresponding differences.

	Men			Women		
Age	Δ	**Std error**	**95% CI**	Δ	**Std error**	**95% CI**
40						
65						
80						

Table 1: Difference in **Mean** Positive Medical Expenditure with and without MSCD.

5. Build a **good** logistic regression for predicting the probability of any expenditure.
6. Based on your models in (2) and (5), produce estimates of the expect medical expenditures. Write down the calculation you used to obtain the estimates of the expect medical expenditures in detail. Summarize the attributable expenditures due to smoking over all people and by gender and age groups.