STAT704 Final Project

Due: Friday Dec 2, 2022 at 5PM

Please hand in a print-out of your answer and R code, and also email your R code to me (hoyen@stat.sc.edu).

Instructions: feel free to discuss the project with other students. However, each student must conduct their own analyses and write-up their own solutions. Write as if for a scientific journal. Be brief and accurate.

Case study: The Health Care Cost for Smoking

I. Analysis Goals

The analysis goal of this project is to estimate the fraction of total medical expenditures among smokers with coronary hear disease, stroke, and lung cancer that can be attributed to their having smoked.

The medical model that underlines this question can be expressed as:

Smoking -> major diseases -> medical expenditure.

The analyses consist of two parts:

(1) Using MEPS, to estimate the difference in average expenditures between the diseased and non-diseased subgroups.

(2) Using MEPS, to estimate the fraction of total medical expenditures among smokers with CHD and other smoking-related diseases that can be attributed to their having smoked.

Reference: Johnson E., Dominici F., Griswold M., Zeger SL. (2003) "Disease cases and their medical cost attributable to smoking: an analysis of the national medical expenditure survey." Journal of Econometrics, 112: 135-151.

II. Dataset

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 for the following analysis.

The MEPS data is available at

http://people.stat.sc.edu/hoyen/Stat704/Data/h129.RData

Use the following code to get needed variables for the following analysis,

```
dat<-data.frame(h129$TOTEXP09, h129$SEX, h129$RACEX,
h129$ASTHDX, h129$DOBMM, h129$DOBYY,h129$ADSMOK42, h129$CALUNG,
```

h129$CHDDX, h129$EDUCYR, h129$POVCAT09, h129$MARRY09X, h129$SEATBE53, h129$PERWT09F, h129$AGE09X, h129$CHBRON31, h129$TTLP09X, h129$STRKDX) and the codebook at https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_codebook.jsp?PUFId=H129&sortBy=Start to look up the meaning of each variable.

The key predictor variables available to us to predict expenditures are:

(1) Demographic variable: age, gender
(2) SES: education (EDUCYR), income (TTLP09X)
(3) Disease: the disease for which smoking is the predominant causes: presence of a diagnosis for lung cancer (CALUNG), presence of a diagnosis of coronary heart disease (CHDDX), stroke (STRKDX).

Perform Analyses Described in III, IV, V

III. Building a Model for the Size of Expenditures

In this final project, we seek to estimate the difference in median expenditure size for persons who as similar as possible except for the presence of a major smoking caused diseases (MSCD). Toward this end, we will start by studying the dependence of expenditure size and presence of an MSCD on the demographic and SES variables. Once we have a reasonable model for predicting expenditure size using these variables, we will introduce the MSCD variables to the model.

1. Determine the dependence of expenditures on age and gender. Use F-statistics to test the null hypothesis that only age and gender are needed and that the spline and interaction terms all have coefficients equal to 0.
2. Study the dependence between expenditures and the two SES variables: education and poverty use scatter plot matrix and AVPLOT.
3. Building a new model with demographic and SES variables. Instead of assuming linear relationships, use dummy variables and interaction terms if needed.
4. Use F-test to determine whether the newly added variables are significant.
5. For the model obtained in (4), perform regression diagnostics and comments on model assumptions. Implement necessary fix-ups and update your model.

IV. Modeling Major Smoking-Caused Diseases

6. Display a side-by-side boxplots of the residuals from the model obtained in (5) stratified by MSCD status: no disease, CHD/stroke/lung cancer (CHD or

stroke or lung cancer). Do you observe any difference in the residuals from the diseased and non-diseased groups?

7. Add a variable for disease status (CHD or stroke or lung cancer) in the model and interpret the regression estimates.

V. Confounding

8. An interesting question is whether the demographic and SES variables "confound" the disease effects. To check, we refit the model with only the disease variables. Compare the disease regression coefficients in the two models.
9. Considering interactions between disease status and other variables in the model. Add the interaction terms in the model if needed.
10. Using the model in (9) and complete the table below.

Age	Estimated **difference** of Median expend(**$**): Diseased -Non-diseased	Std Error (Delta Method)	Std Error (bootstrap)	95% CI
20				
40				
65				
80				

11. Summarize your regression findings in a brief paragraph as if for a public health journal. Use coefficient estimates and confidence intervals in the texts.