Homework Assignment 4

Due Date: Wednesday, March 18, 2019 at 5P

Total Points: 100

1 Case study in logistic regression: predicting positive expenditures

(60 points) 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/Stat704-2018/Data/h129.RData

(a) Use the Medical Expenditure Panel Survey (MEPS) data set for persons 40 and older to build a logistic regression model that estimates the risk of a positive expenditure. Your model should include **mscd** and **other demographics and socio-economic variables** you find are useful predictors of having a positive expenditure. Note that nearly every person with a mscd has a positive expenditure. The coefficient for mscd in a logistic model will therefore be large and it does not make much sense to dwell on potential interactions of mscd with other variables since all of the mscd probabilities will be close to 1.0 by any sensible model. The real goal of the analysis is to estimate the rate (risk) of positive expenditures among non-mscd controls as a function of covariates.

(b) For each model you consider, estimate its coefficients and check the model for consistency with the observations by comparing the observed rates within several bins of predicted rates. Check for extremely influential observations in your final model.

(c) Use your final model to calculate the sensitivity and specificity for classifying a person as having a positive expenditure (or not) as a function of classification threshold. Estimate the area under the ROC curve. Compare your final model with one that only has mscd, age, and gender (main effect only) using area under the ROC curve. Compare your area under the curve with and without cross-validation. Does it make a difference in this case; why or why not?

(c) Summarize your final model and its ability to predict positive expenditures in a paragraph as if for a public health journal.

(d) Estimating the Attributable Risk of Positive Expenditure: Use your model to estimate the average risk of positive expenditure for age × gender subgroups of mscd subjects. Compare this value to the average risk for a group of people with the same covariate profile except without a mscd. Fill out the table below using the following steps:

1. Obtain the estimated risk of a positive expenditure of each person with a mscd.
2. New obtain their predicted probabilities of expenditure with all variables remaining the same except mscd status that is set to 0. [Hint: get the linear precditor and subtract the mscd effect, then transform back to the probability scale].
3. Average the predicted probabilities by age × gender stratum
4. Define the attributable risk of a positive expenditure as:

$$\text{AR} = \frac{P(E > 0|mscd) - P(E > 0|\text{no mscd})}{P(E > 0|mscd)}$$

	Men			Women		
Age	P(E> 0| mscd)	P(E> 0| no mscd)	AR	P(E> 0| mscd)	P(E> 0| no mscd)	AR
< 50						
$51 - 60$						
$61 - 70$						
$70+$						