Medical Expenditures Analysis

Cynthia Kineza

Objective

The aim of this analysis is to estimate from the National Medical Expenditure Survey (NMES) data the fraction of total medical expenditures that are attributable to having a major smoking-caused disease (MSCD) — including lung cancer, laryngeal cancer, COPD, CHD, stroke, and other cancers — for different age, sex and demographic strata.

Methods and Results

To estimate MSCD-attributable expenditures, we fit two models: (1) a logistic regression model for the probability of a positive expenditure and (2) a linear regression model for the size of the expenditure given a positive value. For each model, we begin by exploring the association of the outcome with MSCD, age, sex and demographic variables to select the model form.

Model 1. Probability of positive expenditures

First we plot each variable of interest against the indicator of having zero expenditures. Let Y_i be the total expenditures for subject i. Figure 1 shows that nearly all persons with an MSCD (male and female) have positive expenditures, regardless of age or gender. Furthermore, we see that the probability of positive expenditures increases with age for those persons without an MSCD. We also see that females tend to have a higher probability of positive expenditures than males, but this gap closes with increasing age. Therefore, we include in our model interaction terms between age and MSCD, age and gender, and MSCD and gender.

We perform similar exploratory analysis for the association with race, geographical region and poverty status. We find that subjects with race = 'other' tend to have higher probability of positive expenditures across all ages. We therefore include an indicator of race='other' in our model. Furthermore, subjects in the high income category appear to have a slightly higher probability of positive expenditures, possibly due to an increase in elective expenditures. Therefore, we include an indicator of high income in our model. However, there does not appear to be a strong association between positive expenditures and geographical region. There are missing values for income; however, there are only 56 missing values in a dataset of over 13,000 observations, which will be dropped from the analysis.

We now fit the following logistic regression model for zero expenditures, with age modeled as a cubic spline with a knot at 60:

$$\begin{aligned}\text{logit}Pr(Y_i = 0) = {} & \beta_0 + \beta_1 MSCD_i + \beta_2 Sex_i + \beta_3 MSCD_i Sex_i + \beta_4 RaceOther_i + \beta_5 HighIncome_i \\ & + \beta_6 Age_i + \beta_7 Age_i^3 + \beta_8 (Age_i - 60)_+^3 + \\ & + \beta_9 Age_i Sex_i + \beta_{10} Age_i^3 Sex_i + \beta_{11} (Age_i - 60)_+^3 Sex_i \\ & + \beta_{12} Age_i MSCD_i + \beta_{13} Age_i^3 MSCD_i + \beta_{14} (Age_i - 60)_+^3 MSCD_i\end{aligned}$$

To test whether the age effects are truly nonlinear, we perform a likelihood ratio test comparing with a model that only includes linear age effects, and we find that the age spline terms significantly improve model fit ($p = 0.00529$). We now test whether each of the interaction effects in the model contribute significantly to model fit. First, we perform a t-test on the interaction between MSCD and gender, and do not find a statistically significant effect ($p = 0.808$). Dropping the MSCD-gender interaction effect, we now perform a likelihood ratio test for the age-MSCD and age-gender interactions. We do not find a statistically significant

improvement in model fit by allowing for age-gender or age-MSCD interactions (p = 0.129), so we drop these terms from the final model. Therefore, our final model for zero expenditures is

$$\text{logit}Pr(Y_i = 0) = \beta_0 + \beta_1 MSCD_i + \beta_2 Sex_i + \beta_3 RaceOther_i + \beta_4 HighIncome_i + \beta_5 Age_i + \beta_6 Age_i^3 + \beta_7 (Age_i - 60)_+^3.$$

Figure 2 shows the observed and fitted probabilities of zero expenditures by age, sex and presence of an MSCD for the group with RaceOther=FALSE and HighIncome=FALSE. We group age into deciles and compute the observed probabilities of zero expenditures within each decile. To generate predicted probabilities, we use the median age within each decile.

Figure 1

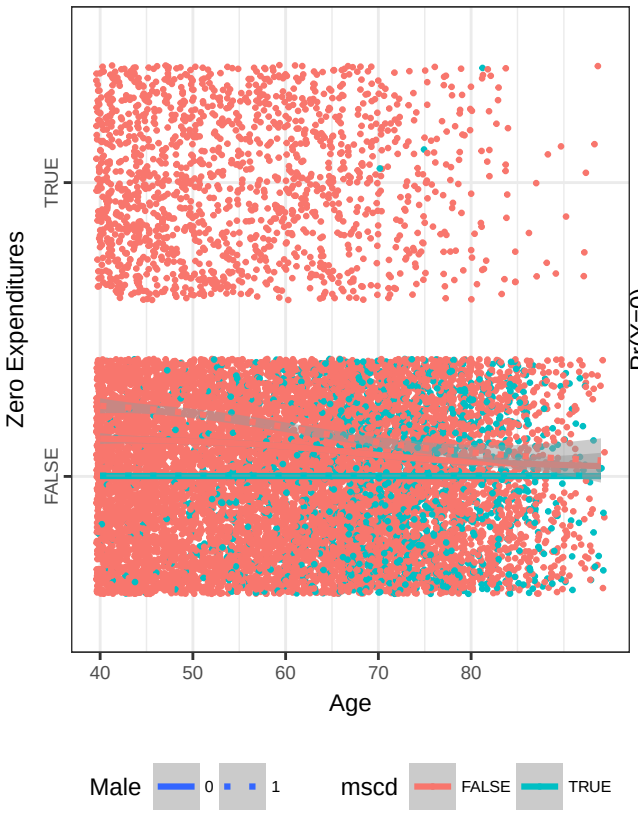
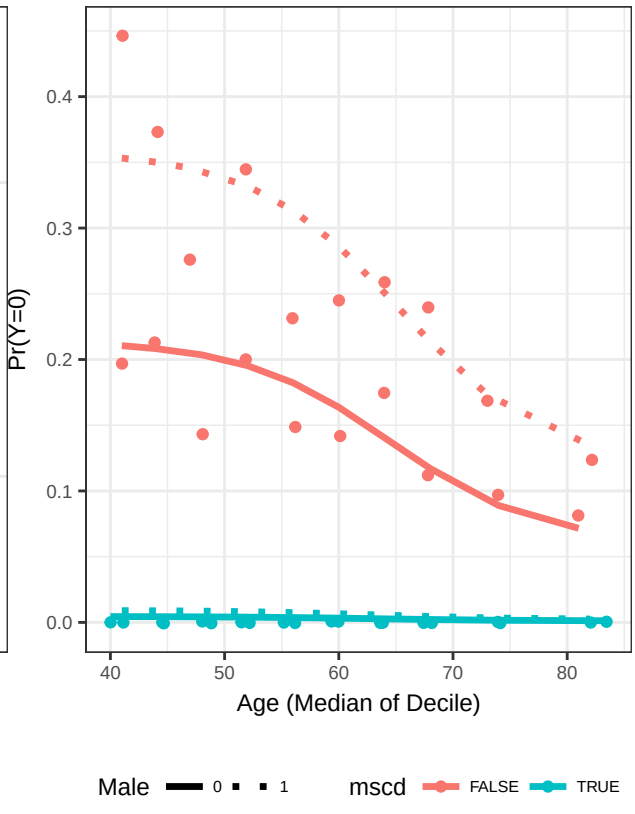


Figure 2



We now summarise the probability of non-zero expenditures for each age-gender strata (fixing RaceOther=FALSE and HighIncome=FALSE). To generate predicted probabilities, we use a fixed age that is representative of each group (e.g. 45, 55, 65, 75). We compute attributable risk as risk difference (the difference in rate of a condition between an exposed population and an unexposed population), namely

$$AR(Y > 0|X) = Pr(Y > 0|X, \text{MSCD}) - Pr(Y > 0|X, \text{no MSCD})$$

	Males			Females		
Age	Pr(Y>0,MSCD)	Pr(Y>0,no MSCD)	AR	Pr(Y>0,MSCD)	Pr(Y>0,no MSCD)	AR
45	0.99	0.65	0.34	1.00	0.79	0.20
55	0.99	0.68	0.31	1.00	0.81	0.18
65	0.99	0.76	0.24	1.00	0.87	0.13
75	1.00	0.84	0.16	1.00	0.92	0.08

Model 2. Size of medical expenditures given positive expenditures

We again begin by exploring the association of expenditures with MSCD, age, sex and demographic variables. Exploratory plots show that for persons with an MSCD, medical expenditures are fairly constant across age and gender. For subjects without an MSCD, expenditures increase with age, and females have slightly higher expenditures than males. Exploratory analysis for the potential role of smoking variables (e.g. total pack-years, age started smoking, age stopped smoking) shows weak associations with size of medical expenditures. However, as the sample contains a large proportion of missing data (up to 50% for some variables), we have not included these variables in the prediction model. Exploratory plots do not show evidence that race, geographical region, or income are associated with size of medical expenditures. Therefore, we stratify by MSCD and fit the following models for expenditures:

$$E(Y_i|MSCD_i = 1) = \beta_0 + \beta_1 Age_i$$
$$E(Y_i|MSCD_i = 0) = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \beta_3 Age_i Sex_i$$

The data are highly skewed, which implies that a Normality assumption for the residuals is inappropriate. However, since we are interested in modeling the *mean* expenditures, we can use ordinary least squares to obtain unbiased and efficient coefficient estimates and use bootstrapping to obtain confidence intervals. (Note: an alternative approach would be to fit a GLM with a log link and a Gaussian or Gamma assumption for the residuals. In this case, we would check the parametric assumption through QQ plots of the model residuals, and use bootstrapping or the Delta method to obtain confidence intervals on the original data scale.)

Disease Attributable Fraction of Expenditures (DAFE) due to MSCD

We now compute the expected expenditures by age, gender and MSCD, fixing RaceOther=FALSE and HighIncome=FALSE. For each covariate profile X, we then compute expected expenses as $E[Y|X] = E[Y|Y > 0, X] \times Pr(Y > 0|X)$ and Disease Attributable Fraction of Expenditures (DAFE) as

$$DAFE(X) = \frac{E[Y|X, \text{MSCD}] - E[Y|X, \text{No MSCD}]}{E[Y|X, \text{MSCD}]}.$$

$E[Y|X]$ and $DAFE(X)$ are both nonlinear functions of quantities that are asymptotically Normal, so they do not follow known distributions. Therefore, we bootstrap subjects 1000 times to obtain 95% confidence intervals for $E[Y|X]$ and $DAFE(X)$. We also use the bootstrap samples to obtain confidence intervals for $Pr(Y > 0)$ and $E[Y|Y > 0]$. Bootstrap confidence intervals are more appropriate than a Normal approximation for $Pr(Y > 0)$, particularly for the MSCD group, which has $Pr(Y > 0)$ close to 1.

In the table below, we summarise the expected values of $Pr(Y > 0)$, $E[Y|Y > 0]$ and $E[Y]$ by age, gender and presence of an MSCD. For each age and gender group, we then compute DAFE. Let Z represent the presence of an MSCD. Expected expenditures are given in thousands.

males

Age	Pr(Y>0|Z=1)	E[Y|Y>0,Z=1]	E[Y|Z=1]	Pr(Y>0|Z=0)	E[Y|Y>0,Z=0]	E[Y|Z=0]	DAFE
45	$_{0.98}0.99_{1.00}$	$_{6.4}7.8_{9.4}$	$_{6.4}7.7_{9.3}$	$_{0.62}0.65_{0.68}$	$_{1.2}1.4_{1.7}$	$_{0.7}0.9_{1.1}$	$_{0.84}0.88_{0.91}$
55	$_{0.98}0.99_{1.00}$	$_{7.3}8.2_{9.3}$	$_{7.2}8.1_{9.2}$	$_{0.65}0.68_{0.71}$	$_{1.9}2.1_{2.3}$	$_{1.3}1.4_{1.6}$	$_{0.80}0.83_{0.85}$
65	$_{0.99}0.99_{1.00}$	$_{7.9}8.6_{9.5}$	$_{7.9}8.6_{9.4}$	$_{0.74}0.76_{0.78}$	$_{2.4}2.7_{3.0}$	$_{1.8}2.0_{2.3}$	$_{0.73}0.76_{0.80}$
75	$_{0.99}1.00_{1.00}$	$_{8.2}9.0_{9.8}$	$_{8.2}9.0_{9.8}$	$_{0.82}0.84_{0.86}$	$_{2.9}3.3_{3.7}$	$_{2.5}2.8_{3.2}$	$_{0.64}0.69_{0.74}$

Females

Conclusion

Expenditures increase with age for males and females with and without an MSCD, and expenditures are slightly higher for females than for males, particularly for subjects without an MSCD. The DAFE decreases

Age	Pr(Y>0|Z=1)	E[Y|Y>0,Z=1]	E[Y|Z=1]	Pr(Y>0|Z=0)	E[Y|Y>0,Z=0]	E[Y|Z=0]	DAFE
45	$_{0.99}1.00_{1.00}$	$_{6.4}7.8_{9.4}$	$_{6.4}7.8_{9.3}$	$_{0.77}0.79_{0.81}$	$_{1.3}1.5_{1.6}$	$_{1.0}1.2_{1.3}$	$_{0.82}0.85_{0.88}$
55	$_{0.99}1.00_{1.00}$	$_{7.3}8.2_{9.3}$	$_{7.3}8.2_{9.3}$	$_{0.79}0.81_{0.83}$	$_{2.1}2.2_{2.3}$	$_{1.7}1.8_{1.9}$	$_{0.75}0.78_{0.81}$
65	$_{0.99}1.00_{1.00}$	$_{7.9}8.6_{9.5}$	$_{7.9}8.6_{9.5}$	$_{0.85}0.87_{0.88}$	$_{2.7}2.9_{3.1}$	$_{2.4}2.5_{2.7}$	$_{0.67}0.71_{0.74}$
75	$_{1.00}1.00_{1.00}$	$_{8.2}9.0_{9.8}$	$_{8.2}9.0_{9.8}$	$_{0.90}0.91_{0.93}$	$_{3.4}3.7_{3.9}$	$_{3.1}3.3_{3.6}$	$_{0.58}0.63_{0.67}$

with age, ranging from 0.88 (0.84, 0.91) at age 45 to 0.69 (0.64, 0.74) at age 75 for males, and from 0.85 (0.82, 0.88) at age 45 to 0.63 (0.58, 0.67) at age 75 for females. Across ages and genders, well over 50% of expenditures can be attributed to having an MSCD.