Assignment 4

Sarah Kim

I. Introduction

Healthcare costs can be difficult to predict, as they are influenced by many different socioeconomic and health factors. In this report, I analyze the relationship between healthcare costs and a range of health and demographic characteristics, specifically smoking status, sex, age, and cardiac disease status. I find that smoking status, age, and cardiac disease are correlated with higher healthcare costs, whereas being female is associated with lower costs.

Section II presents summary statistics of the cohort data, describing the composition of the sample and showing the distribution of the outcome. In Section III, I discuss the methods used for analysis, and Section IV follows with regression results and visual aids to help illustrate the relationship between healthcare costs and the contributing factors included in the cohort dataset. Section V concludes with a brief discussion of limitations and future direction of research.

II. Data

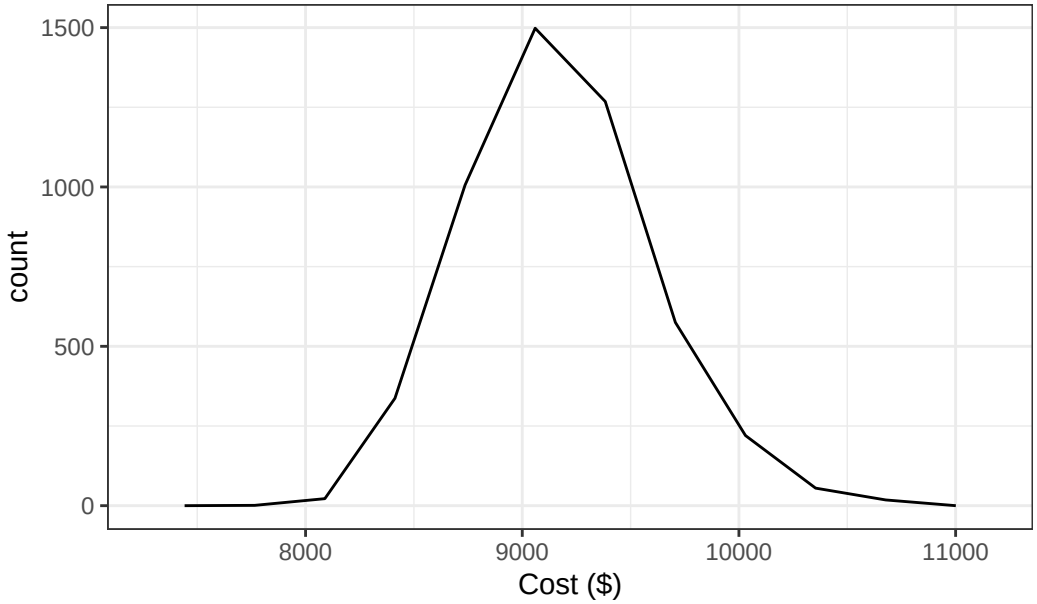
The cohort dataset consists of individual-level information on healthcare costs, age, sex, smoking status, and cardiac disease status. The table below provides basic summary statistics of each variable in the sample:

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
smoke	1	5000	0.16	0.36	0	0.07	0.00	0	1	1	1.88	1.52	0.01
female	2	5000	0.55	0.50	1	0.57	0.00	0	1	1	-0.22	-1.95	0.01
age	3	5000	43.94	15.10	44	43.90	19.27	18	70	52	0.01	-1.22	0.21
cardiac	4	5000	0.05	0.22	0	0.00	0.00	0	1	1	4.13		
cost	5	5000	9165.73	420.80	9143	9149.78	409.20	7878	10790	2912	0.39		

```
cardiac    15.05 0.00
cost        0.21 5.95
```

Of the 5000 individuals in the sample, 16% are smokers, 55% are female, and 22% have a history of cardiac problems. Ages range from 18 to 70, with a median age of 44. The median healthcare cost is approximately $9165, and the mean is similarly close at around $9150, suggesting that the distribution is not heavily skewed. Figure 1, which shows the distribution of healthcare costs, further supports that the data is approximately normally distributed.

Figure 1. Cost Distribution



III. Methods

As Section II suggests that healthcare cost is not heavily skewed, a simple ordinary least squares (OLS) regression is used to estimate the relationship between cost and a set of demographic and health factors. The regression equation is specified as follows:

$$cost_i = \beta_0 + \beta_1 age_i + \beta_2 female_i + \beta_3 smoke_i + \beta_4 cardiac_i + \epsilon_i \tag{1}$$

where i indexes individuals. The outcome of interest is healthcare cost, and the predictors include age, gender, smoking status, and presence of cardiac disease. Age is a continuous integer variable, representing an individual's age in years. All other variables are binary indicators, taking the value of 1 when the corresponding condition is true, and 0 otherwise (e.g., $female = 1$ if the individual is female, and 0 if male). As the cohort data does not include

a time variable indicating when each observation was recorded, all variables are assumed to be measured at approximately the same point in time, which is reasonable given that the dataset includes information on age.

Additionally, to explore potential differences in how age relates to healthcare costs across subgroups, I conduct a heterogeneity analysis by examining the association between age and cost separately for males and females, smokers and non-smokers, and individuals with and without cardiovascular conditions.

IV. Results

The summary below shows the regression results for Equation 1:

```
Call:
lm(formula = cost ~ age + female + smoke + cardiac, data = cohort)

Residuals:
    Min      1Q  Median      3Q     Max
-830.33 -137.76    8.22  140.69  889.72

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8506.8687     9.8643  862.39   <2e-16 ***
age           15.7782     0.1953   80.81   <2e-16 ***
female      -252.9474     6.0622  -41.73   <2e-16 ***
smoke        541.9547     8.3271   65.08   <2e-16 ***
cardiac      408.2382    14.2247   28.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 208.3 on 4995 degrees of freedom
Multiple R-squared:  0.7551,    Adjusted R-squared:  0.7549
F-statistic:  3850 on 4 and 4995 DF,  p-value: < 2.2e-16
```

The regression results indicate that age, smoking status, and presence of cardiac conditions are positively correlated with healthcare costs. Being a female is associated with reduced costs. Accounting for gender, smoking status, and cardiac disease status, a person's age is associated with an average of $15.78 higher cost for every additional year of age. Similarly, holding all else constant, females are associated with an average of $252.95 lower healthcare costs, while smokers and patients with cardiac conditions incur on average $541.95 and $408.24 higher

costs, respectively. For all four factors, the estimates are highly significant, with p-values approximately equal to 0.

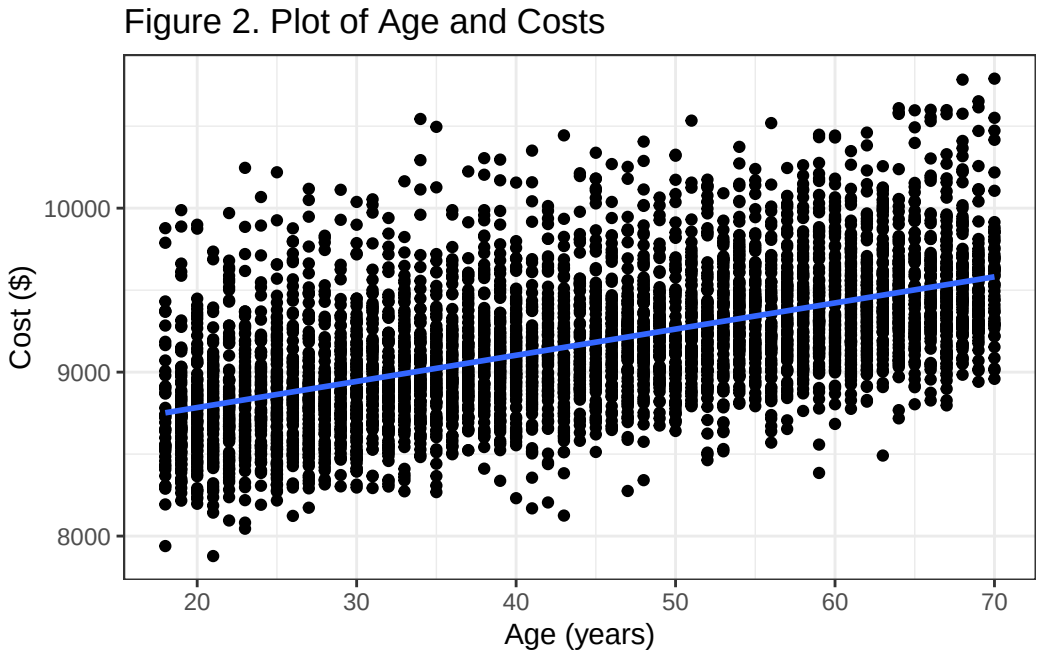
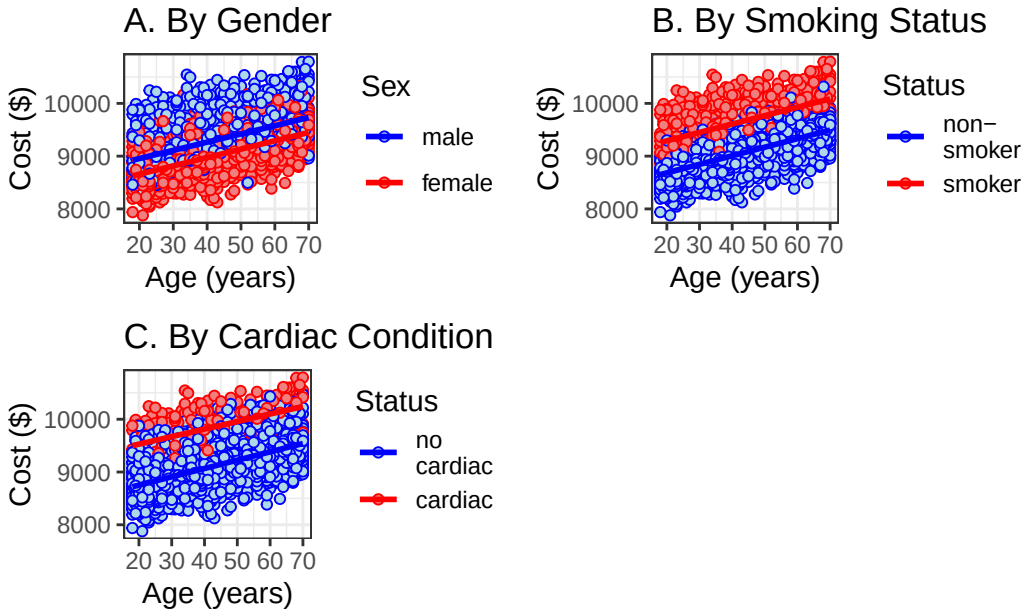


Figure 2 displays a scatter plot of age and cost, with a fitted line graph representing a positive relationship between the two variables. Figure 3 displays the relationship between age and cost, separately by gender, smoking status, and cardiac disease status. Across all three subgroup comparisons, the fitted lines appear to be parallel, suggesting that the cost differences between groups remain relatively constant across age.

Figure 3. Plot of Age and Costs: Subgroup Analysis



V. Discussion

In conclusion, holding all else constant, one additional year of age is associated with an extra $16 in healthcare costs. Ceteris paribus, females tend to spend $253 less than males in healthcare costs, while smokers and patients with cardiac disease incur $542 and $408 more in costs, respectively.

Additional data would help clarify the underlying mechanisms. Since total healthcare costs are a function of both price and quantity (cost = price * utilization), the observed gender difference could reflect several possibilities. Females may generally be healthier than their male counterparts and therefore require fewer healthcare services, or they may tend to receive care for conditions that are less costly to treat, even if they utilize services more frequently. It would be informative to have data on healthcare utilization, diagnoses, and treatment-specific costs to better understand the sources of these differences. Similarly, smokers and individuals with cardiac conditions may differ from their counterparts in ways not fully captured by the current dataset, such as unobserved health behaviors, access to care, or other demand- and supply-side factors influencing healthcare spending.

A simple OLS regression allows for the exploration of associations but does not provide evidence of a causal relationship. OLS estimates may be biased due to confounding variables, reverse causality, or omitted variable bias. Establishing causality would require a study design that better accounts for these issues, such as randomized control trials, instrumental variables, or difference-in-differences methods to track changes over time. Estimating the causal

effects of demographic or behavioral factors on healthcare costs would yield more insights for policymakers and healthcare providers to help reduce healthcare expenditures.

I did not use generative AI technology (e.g., ChatGPT) to complete any portion of the work.