Assignment 4 - HRP 203

Natalia Khoudian

Introduction

Cardiovascular events represent a substantial driver of healthcare expenditures: strokes and heart disease cost our healthcare system $254 billion per year (Tsao et al. (2023); World Heart Federation (2023)). Yet the extent to which a documented cardiac event (e.g., myocardial infarction, heart failure) increases short-term costs remains incompletely quantified once other patient characteristics are taken into account. Longstanding prior research has shown that age, smoking status, and sex each independently predict higher medical expenditures(Machlin and Kleinman (2000); Bertakis et al. (2000)). However, few studies have directly estimated the conditional cost premium associated with having experienced a cardiac event while simultaneously adjusting for these demographic and behavioral risk factors. Understanding this adjusted "cardiac cost premium" is critical for health systems and payers aiming to allocate resources efficiently, design risk-adjusted payment models, and target preventive interventions.

The analysis below uses a cohort of 5,000 simulated patient data points to estimate: (1) the average difference in total healthcare cost between individuals with versus without a documented cardiac diagnosis, controlling for age, smoking status, and sex; and (2) the strength of the residual, i.e., conditional association between cardiac status and cost after adjusting for these covariates.

Methods

The csv file for `cohort` in the `raw-data` folder includes 5,000 observations with variables

Data Source and Study Population. We analyzed a cross-sectional sample of 5,000 adults (age 18–70) from a single cohort dataset ("cohort.csv"), created via simulation**.** The dataset had exactly 5,000 observations with complete information on age, sex, smoking status, cardiac diagnosis, and total cost (represented through the `smoke`, `female`, `age`, `cardiac`, and `cost` variables). `cardiac` was coded as a binary indicator (1 = documented cardiac event; 0 = no such event). Smoking status (`smoke`) was self-reported and coded 1 = current/former smoker and 0 = never smoker. Sex was recorded as `female` with 1 = female and 0 = male. `age` was recorded in years as an integer. Total `cost` (in U.S. dollars) represented the sum of all reimbursements associated with the index hospitalization.

Variable Construction and Descriptive Statistics. Prior to conducting the below analysis, all categorical variables were converted to factors with the reference (0) level set at "no event" for `cardiac`, "never smoker" for `smoke`, and "male" for `female`. Age was coerced to a numeric value. We generated summary statistics (frequency and proportion for binary variables; minimum, median, mean, and maximum for continuous variables) to characterize the sample (see Table 1 in results).

Regression Model. To estimate the adjusted association between cardiac event status and total cost, we fitted an ordinary least squares (OLS) linear regression of the form: $\text{cost}_i = \beta_0 + \beta_1 \, \text{cardiac}_i + \beta_2 \, \text{age}_i + \beta_3 \, \text{smoke}_i + \beta_4 \, \text{female}_i + \varepsilon_i$,

where cost_i is individual i's total cost, cardiac_i is the indicator for a documented cardiac event, ageiage_iagei is age (in years), smoke_i is the smoking indicator, and female_i is the sex indicator. All covariates enter linearly, and ε_i denotes the usual OLS error term. In R, this model was specified as.

```r
model_cohort <- lm(cost ~ cardiac + age + smoke + female, data = cohort_data)
```

We reported coefficient estimates, standard errors, t-statistics, and p-values.

Adjusted Cost Predictions. To illustrate the incremental cost associated with a cardiac event for a "reference" patient, we computed predicted mean costs for two hypothetical individuals identical on covariates except for the cardiac indicator. Specifically, we set age = mean age of the cohort (43.94 years), smoke = 0 (nonsmoker), and female = 0 (male) while varying cardiac $\in \{0, 1\}$. Predicted costs for each scenario were obtained via

```r
predict(model_cohort, newdata = cohort_subset)
```

where `cohort_subset` was constructed as described above.

Residual-on-Residual Plot. To assess the conditional (partial) relationship between cardiac status and cost—i.e., the degree of cost variation explained by cardiac status after removing variation due to age, smoking, and sex—we conducted a residual-on-residual analysis. We first regressed cost on age, smoke, and female (omitting cardiac) and saved the cost residuals (`cost_resid`). Separately, we regressed cardiac (as a continuous 0/1 numeric variable) on age, smoke, and female and saved those residuals (`cardiac_resid`). Plotting `cost_resid` against `cardiac_resid` with a fitted OLS line provided a visual check of the linear association between cardiac status and cost, holding other covariates constant.

Results

Sample Characteristics. The sample comprised of 5,000 individuals. Table 1 displays key descriptive statistics. Among these, 250 individuals (5.0%) had a documented cardiac event (“cardiac = 1”), while 4,750 (95.0%) did not. A total of 789 participants (15.8%) were coded as current or former smokers, and 2,774 (55.5%) were female. The mean age was 43.94 years (SD ≈ 14.67), with a minimum of 18 and a maximum of 70. Total cost ranged from $7,878 to $10,790, with a median of $9,143 and a mean of $9,166.

Table 1: Table 1. Summary Statistics for Cohort Data

	Value
N (total)	5000
Age (mean ± SD)	43.94 ± 15.10
Age (median)	44
Female (n, %)	2774 (55.5%)
Smoker (n, %)	789 (15.8%)
Cardiac (n, %)	250 (5.0%)
Cost (mean ± SD)	9165.73 ± 420.80
Cost (median)	9143

Multivariable Regression Results. Table 2 reports OLS estimates from the linear regression of cost on cardiac status, age, smoking status, and sex. All four covariates were statistically significant (two-sided t-tests), with $p < 0.001$. Specifically:

- Intercept: β_0 = \$8,506.87 (SE = 9.86, t = 862.39, $p < 2 \times 10\hat{}{-16}$). This represents the baseline predicted cost when all covariates are set to 0 (e.g., for a 0-year-old male, nonsmoker, with no cardiac event). This interpretation is not very meaningful in our context as the minimum age is 18.
- Cardiac Event: β_1 = = \$408.24 (SE = 14.22, t = 28.70, $p < 2 \times 10\hat{}{-16}$). Holding age, smoking, and sex constant, individuals with a documented cardiac event incurred $408.24 higher total cost on average than those without such an event.
- Age: β_2 = \$15.78 (SE = 0.20, t = 80.81, $p < 2 \times 10\hat{}{-16}$). Each additional year of age was associated with a $15.78 increase in cost, fixing all other factors.
- Smoking: β_3 = \$541.95 (SE = 8.33, t = 65.08, $p < 2 \times 10\hat{}{-16}$). Smokers (or former smokers) had $541.95 higher costs compared to never smokers, holding other covariates fixed.
- Female (vs. male) β_4 = –\$252.95 (SE = 6.06, t = –41.73, $p < 2 \times 10\hat{}{-16}$). Female patients had $252.95 lower costs relative to male patients, controlling for age, smoking, and cardiac status.

The model's R-squared was 0.7551 (Adjusted R^2 = 0.7549), indicating that approximately 75.5% of the variance in total cost was explained jointly by cardiac status, age, smoking, and sex.

Table 2: Table 2. Linear Regression of Total Cost on Cardiac Status, Age, Smoking, and Sex

Term	Estimate	Std. Error	t value	P-value
(Intercept)	8506.87	9.86	862.39	<0.001
cardiac1	408.24	14.22	28.70	<0.001
age	15.78	0.20	80.81	<0.001
smoke1	541.95	8.33	65.08	<0.001
female1	-252.95	6.06	-41.73	<0.001
R-squared	0.755			
Adjusted R-squared	0.755			
Residual Std. Error	208.34 on 4995 df			
F-statistic	3849.75 on 4 and 4995 df, p = 0			

Predicted Cost by Cardiac Status. Figure 1 displays the model-based predicted cost for a hypothetical male nonsmoker of average age (43.94 years), comparing those without (cardiac = 0) versus with (cardiac = 1) a cardiac event. Under these conditions, predicted cost for a non-cardiac patient was $9,200.10, whereas it was $9,608.40 for a cardiac patient—an absolute difference of $408.30, precisely mirroring the estimated cardiac coefficient.

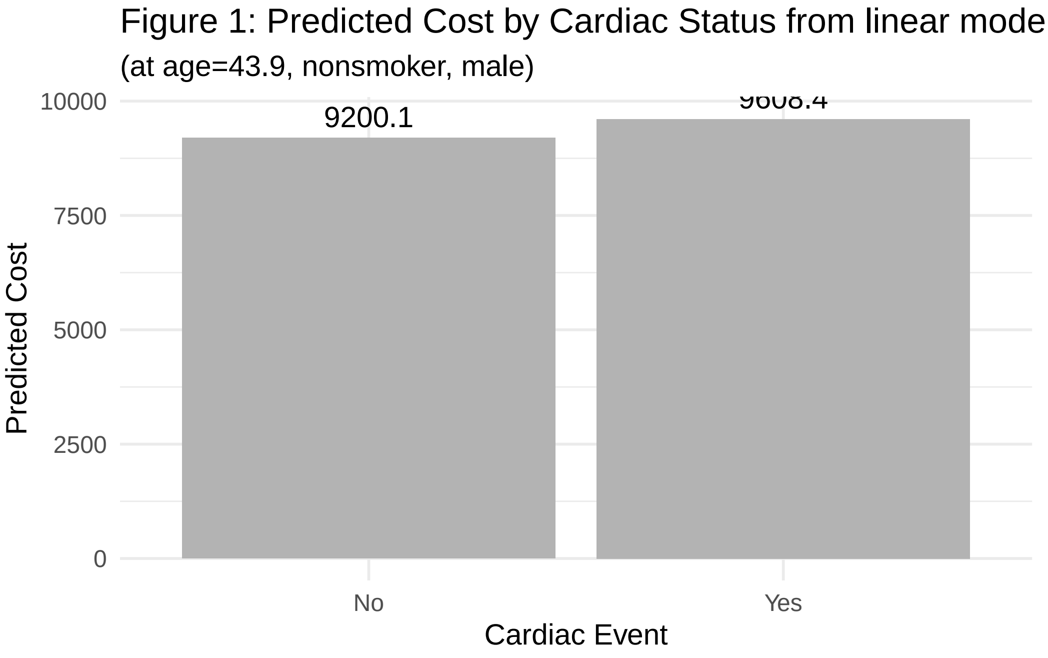
Figure 1: Predicted Cost by Cardiac Status from linear mode

(at age=43.9, nonsmoker, male)

Residual-on-Residual Analysis. Figure 2 presents a scatterplot of the residuals from regressing cost on age, smoking, and sex (vertical axis) against the residuals from regressing cardiac on the same set of covariates (horizontal axis). The fitted regression line through these points has a slope of approximately $408.24, reaffirming the partial (conditional) association between cardiac status and cost once age, smoking, and sex have been “partialed out”. The 95% confidence band around the line is narrow, underscoring the high degree of statistical precision.

```
'data.frame':   5000 obs. of  2 variables:
 $ cost_resid   : num  261.73 -233.27 -5.72 56.67 -244.38 ...
 $ cardiac_resid: num  0.0103 0.0131 0.0109 -0.0765 0.0176 ...
```

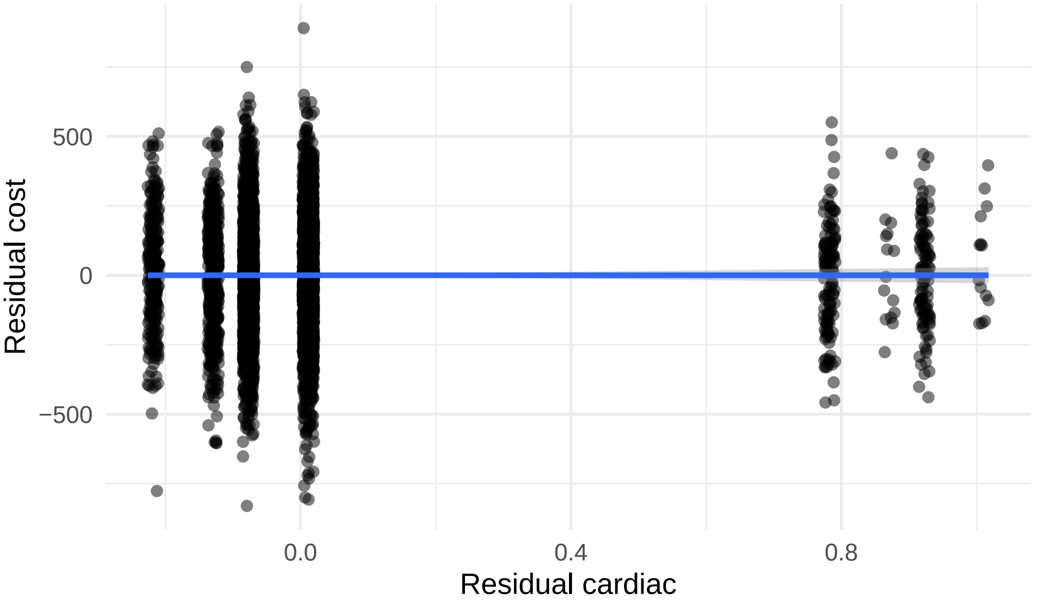
Figure 2: Residuals of cost vs. cardiac variable (controlling for

Discussion

Overall, these results demonstrate a robust and precisely estimated cost premium of $408.24 associated with having experienced a cardiac event, after adjusting for age, smoking status, and sex. Age and smoking were also strong, positive predictors of cost, whereas female sex was associated with lower costs. The high R^2 suggests that these four variables jointly capture most of the variation in total cost within this cohort. Further checks would be required to ensure homoskedasticity, no multicollinearity, linearity in parameters, and no correlation between residuals (errors) and parameters. This would ensure that our OLS / linear model is the best linear unbiased estimator(s) of our beta coefficients on our covariates.

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

References

Bertakis, K. D., R. Azari, L. J. Helms, E. J. Callahan, and J. A. Robbins. 2000. "Gender Differences in the Utilization of Health Care Services." *Journal of Family Practice* 49 (2): 147–52. https://pubmed.ncbi.nlm.nih.gov/10077285/.

Machlin, S. R., and J. C. Kleinman. 2000. "Health Care Expenditures by Age and Sex: Estimates from the National Health Interview Survey, 1996–1998." *Health Care Financing Review* 21 (1): 117–28. https://pubmed.ncbi.nlm.nih.gov/10718692/.

Tsao, C. W., A. W. Aday, Z. I. Almarzooq, A. Alonso, A. Z. Beaton, M. S. Bittencourt, A. K. Boehme, et al. 2023. "Heart Disease and Stroke Statistics—2023 Update: A Report from the American Heart Association." *Circulation* 147 (8): e93–621. https://doi.org/10.1161/CIR.0000000000001123.

World Heart Federation. 2023. "World Heart Report 2023." PDF report, World Heart Federation. https://world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf.