HRP 203 Final Project

Genna Campain

2024-05-28

Introduction

In the United States, healthcare spending makes up a large portion of GDP, equal to about 17.3% in 2022 (Statista, 2024). For the majority of individuals, these expenditures are covered at least partly by some type of public or private insurance which (loosely) allows individuals to pay monthly premiums in exchange for coverage of medical expenses (KFF, 2023). Coverage of this form gives individuals a buffer against the potentially ruinous costs of a serious medical condition and allows risk-sharing among all insured individuals on the same plan (or insured by the same entity in some cases) (American Academy of Actuaries, 2024). However, this risk-sharing relies on the plan (or entity) budget balancing over time (so that the money coming in from premiums must equal the money paid out for medical expenses). With many insurers, plan enrollment and premium setting occur at the beginning of a plan year, so insurers must balance the budget based on *predicted* expenditures. In this paper, I consider one model that predicts individual health expenditures based on individuals' characteristics and pre-existing conditions.

Methods

For the data analysis, I used the `cohort` simulated data set provided through the course GitHub repository. A copy of this data set can be found in the `Data` folder in the GitHub repository for this project. The table below shows the first six rows of the data set.

	smoke	female	cardiac	age	cost
1	1	0	0	44	10566
2	0	1	0	46	9668
3	0	0	0	56	9889
4	0	0	0	35	9780
5	0	0	0	49	10200
6	0	0	0	64	10082

I included all of the variables from the data set in my analysis. Since documentation was not provided, I made assumptions about the meanings of the variables. These assumptions are outlined in Table 1.

Table 1: Definitions for five variables included in analysis

Variable Name	Description
`smoke`	indicator variable, equal to 1 if the patient smokes regularly (more than three times per week) and 0 if not
`female`	indicator variable, equal to 1 if the patient's sex is designated as female and 0 if not
`cardiac`	indicator variable, equal to 1 if the patient has a diagnosis of a cardiac-related problem and 0 if not
`age`	numeric variable, indicates the patient's age in years
`cost`	numeric variable, indicates the cost (in USD) of all healthcare visits for a given patient in a given year

I started the analysis with descriptive statistics for the five variables of interest. I examined the mean, standard deviation, minimum and maximum values, and skewness of each variable to get an idea of the type of variation available in the data set.

For the next part of the analysis, I ran a regression to examine how an individual's smoking habits, sex, cardiac history, and age can be used to predict their yearly healthcare costs. I used the `lm()` function with the following equation:

$$cost = \beta_0 + \beta_1 smoke + \beta_2 female + \beta_3 cardiac + \beta_4 age$$

Finally, I generated predicted spending amounts for each individual using the model and calculated the residuals to assess the model fit:

$$residuals = cost_{observed} - cost_{predicted}$$

.

Results

Data Summary Table

The table below shows the summary statistics for the five variables included in the model. In the sample of 5,000 individuals, approximately 10% of individuals smoke, 49% are female, and 4% have previous history of a cardiac problem. The average age of the sampled individuals is 41, although adults of all ages who might have private insurance (18 to 65) are included.

The mean spending is $9,672, with a maximum spending amount of $11,326 and a minimum spending amount of $8,478.

While the characteristics of the individuals in the data set seem fairly representative of the privately insured population with respect to age and sex, the fact that no individuals have $0 spending in the data set might limit the applicability of the model to populations with low levels of spending (i.e. insurers with fairly young insured populations).

	Variable	Min	Max	Mean	SD	Skewness
1	smoke	0	1	0.1016	0.30215	2.63656
2	female	0	1	0.487	0.49988	0.052
3	cardiac	0	1	0.038	0.19122	4.83128
4	age	18	65	41.4702	13.5407	0.01173
5	cost	8478	11326	9672.2744	402.63168	0.32417

Figure 1

The regression output table and coefficient plot below show the coefficients and associated standard errors for the four predictors in the model. Overall, the model performance seems adequate for the limited amount of information available. All of the included variables are significant predictors of cost at the 95% level, with very small standard errors compared to the magnitude of the coefficient. The estimated coefficients represent meaningful associations (i.e the change in cost associated with the change in a predictor is by a nontrivial amount) and the model itself explains a large portion (76%) of the overall variation in cost.

Regarding the individual coefficients, smoking is associated with the largest increase in cost of $592, following by being male ($294) and then having a cardiac condition ($289). The age variable has the smallest magnitude relationship, as an additional year of age is associated with only an $18 increase in costs. However, since this $18 is per year, the cumulative effect as individuals age is far greater.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8988.79814	9.5391709	942.30392	0.000000e+00
cardiac	289.22361	15.2188580	19.00429	8.261916e-78
smoke	592.75834	9.5149068	62.29786	0.000000e+00
age	18.21239	0.2081448	87.49867	0.000000e+00
female	-293.65483	5.7040625	-51.48170	0.000000e+00

```
[1] "Adjusted R^2: 0.755"
```

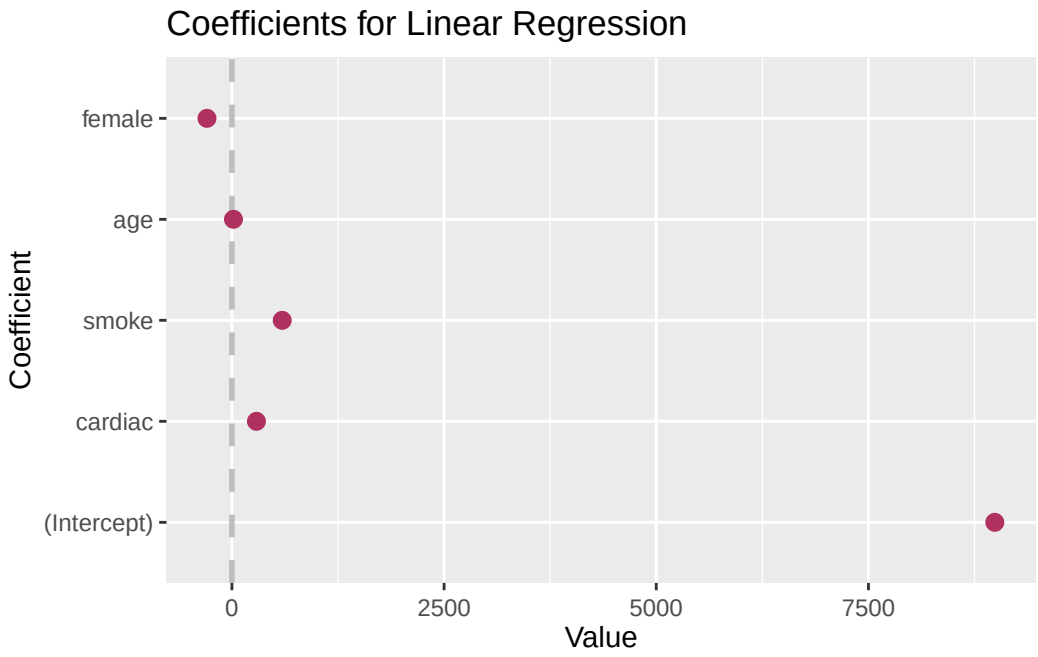
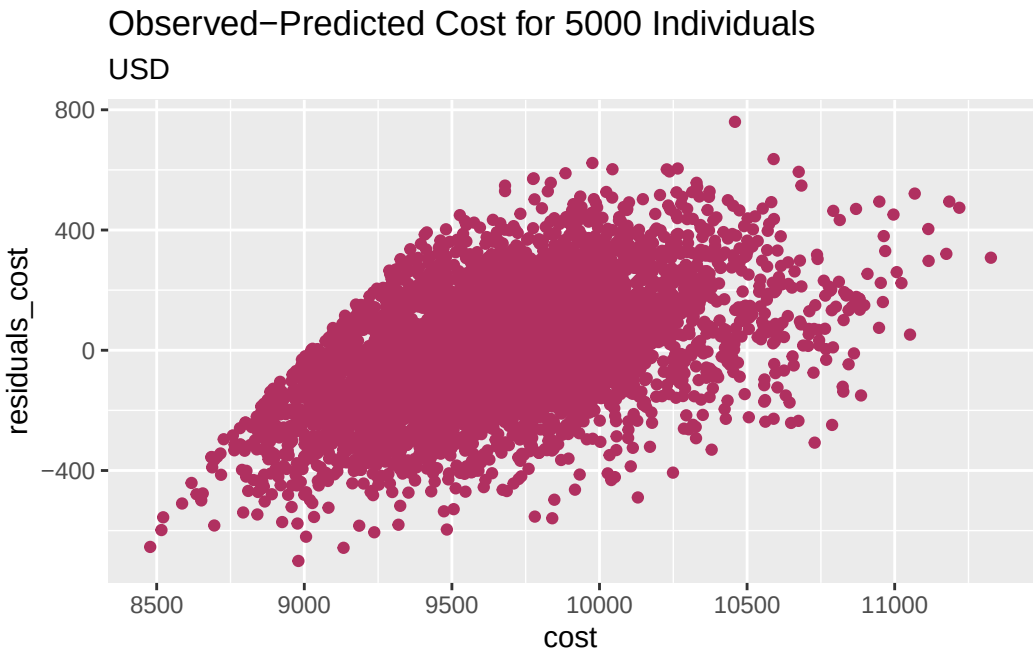


Figure 2

The scatterplot below shows the residuals for the model, plotted against the observed cost values. Ideally, the negative and positive residuals would be equally distributed at all values of observed cost. However, the residuals for this model have a clear upward trend as cost increases, indicating that the model is underpredicting high cost values and overpredicting low cost values. Allowing for a more flexible model fit might help resolve this problem, although it comes with the risk of overfitting.



Discussion

In this paper, I have demonstrated one potential model that can be used to predict patient health expenditures based on patient characteristics and pre-existing health conditions. This type of model could be of use to insurers seeking to predict future expenditures for budgetary purposes, or for individuals trying to predict their own health expenditures when choosing between plan types. The model has decent predictive power even though it only includes four variables and a very simple specification. This simplicity has both benefits and limitations. The main benefits are that the simple specification makes the model very easy to interpret and understand, and the included variables are easy for insurers or patients to collect or observe. The main drawback is that the model fit is adequate but may not be good enough for some applications. It is not hard to imagine that a much better fit could be achieved by adding a few additional variables, especially with the significant amount of data on healthcare use that is available to insurers. Future work could focus on expanding the number of variables included, or on incorporating more advanced techniques for fitting the model.