Alternative Predictive Modeling for Medicare Patient Costs

Jordan Jang, Samuel O'Neill, and Ming Yi
Faculty Advisors: Ian Duncan and Xiyue Liao
Department of Statistics and Applied Probability - University of California, Santa Barbara
Partially supported by Society of Actuaries Centers of Actuarial Excellence 2015 Education Grant

Abstract

As health care expenditures increase, patient cost mitigation becomes more essential. Cost mitigation programs such as Accountable Care Organizations rely on the ability to accurately predict patient risk, which is difficult because of highly-skewed data. We examine Medicare public use data that includes demographics, costs, and health conditions. We first consider the Centers for Medicare and Medicaid Services' currently-used linear model and then implement more complex generalized linear and additive models to predict patient costs in a future year based on current year data. We find that the latter models more accurately predict the entire distribution of Medicare patient costs and, thus, can improve the existing cost mitigation frameworks.

Introduction

Medicare is a significant part of government spending, accounting for about 15% of U.S. government spending in 2017. As Medicare expenditure continues to strain federal spending, it becomes more essential to find ways to mitigate health care costs. One initiative involves patient interventions where a future high-cost patient is identified ahead of time so a hospital or insurance organization can reach out and recommend appropriate preventative care. Another cost mitigation technique involves Accountable Care Organizations ($\mathbf{ACO}$'s) that are financially incentivized to mitigate care costs. The first initiative requires an effective predictive model for high cost patients and the second one requires accurate cost modeling across the entire distribution of patients since ACO's receive money based on their health cost savings adjusted for group patient risk (expected expensiveness without cost-effective care).

The Centers for Medicare & Medicaid Services ($\mathbf{CMS}$) administers Medicare and currently uses a hierarchical condition categories ($\mathbf{HCC}$'s) model for risk assessment. HCC's are health condition indicator variables that are extracted from more than 14,000 International Classification of Diseases (ICD-10) categories. HCC's represent the most severe condition a patient has among a group of similar conditions. CMS currently uses a linear regression model to assess patient risk based on HCC codes and patient sex and age. Patient **risk scores** are constructed based off of this model, which are a predefined linear combination of demographic and HCC factors. This model is effective for assessing group risk, but it is not a sufficient model for individual cost predictions, especially for high-cost patients [1]. Therefore, we want to explore other models that can better predict the entire distribution of Medicare patient costs.

Data

Quantitative Variable Summary Statistics

Predictor	Mean	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.95}$	Max
Age	72	66	72	80	89	99
Risk score	3.3	0.8	2.3	4.7	9.7	23.1
Sum of HCC's	6	1	5	9	17	34
Inpatient costs	3,008	0	0	0	19,076	367,176
Outpatient costs	1,001	0	160	1,100	4,210	64,180
Carrier costs	1,825	280	1,180	2,570	6,000	28,920
Drug costs	178	0	40	280	740	3,180
2008 total cost	6,013	620	2,060	5,390	26,728	367,176
2009 total cost	6,273	950	2,890	6,770	24,581	201,406

q_α represents the α^{th} quantile of the variable. Cost categories are for 2008 cost data. All cost are reported in US dollars.

We use Medicare public use file data from 2008 and 2009 to train and evaluate our models. This data includes ICD-10 condition categories, patient demographics (sex, age, race, and state), and patient costs by category (inpatient, outpatient, carrier, and drug). Our cleaning process first involved creating HCC's from ICD-10 codes and then constructing patient risk scores according to the CMS HCC model formula. This allowed us to test the currently-used CMS linear model where a patient's risk score is the only predictor for the next year's cost. Most of our variables are highly skewed due to the fact that the majority of patients are low risk except for a small group of high cost patients.

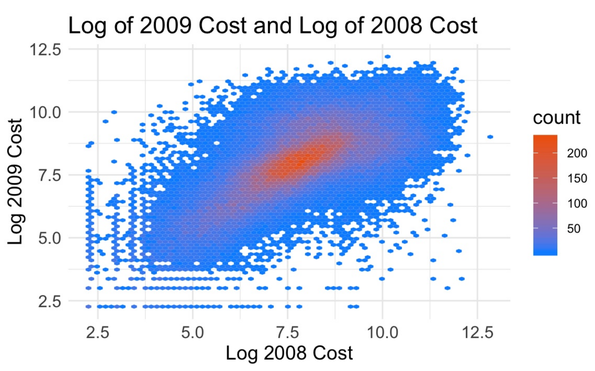


Figure 1: Although untransformed cost data is highly skewed, there is a linear relationship between a patients' current year costs and next year costs on the logarithmic scale.

Methods

All of our models use covariates based on 2008 data to predict 2009 individual patient costs. The CMS linear model is the simplest, which uses only risk score as a covariate. Our other models use risk scores, sum of HCC's, inpatient costs, outpatient costs, carrier costs, and drug costs as covariates to predict total 2009 patient costs. Using additional cost predictor variables can significantly improve model accuracy as shown by Figure 1 and Duncan et al. [2]. We test a variety of frameworks including generalize linear ($\mathbf{GLM}$), generalized additive ($\mathbf{GAM}$), and random forest ($\mathbf{RF}$) models. GLM and GAM models can be written in the following form where Y_i is the total 2009 cost for the i^{th} patient and $x_{m,i}$ is the m^{th} predictor variable for the i^{th} patient.

$$\begin{aligned} \text{GLM:} \quad & g(E[Y_i]) = \beta_0 + \beta_1 \quad x_{1,i} + \beta_2 \quad x_{2,i} + \ldots + \beta_m \quad x_{m,i} \\ \text{GAM:} \quad & g(E[Y_i]) = \beta_0 + f_1(x_{1,i}) + f_2(x_{2,i}) + \ldots + f_m(x_{m,i}) \end{aligned} \tag{1}$$

Both GLM's and GAM's express some (link) function of the expected 2009 cost as a linear combination of functions of the predictor variables. The key difference between GLM's and GAM's is that GLM's only uses linear functions of predictor variables where GAM's can use any arbitrary nonlinear function of predictor variables. In the R package `mgcv`, these arbitrary predictor functions are fit to the data using penalized regression splines (Figure 2).

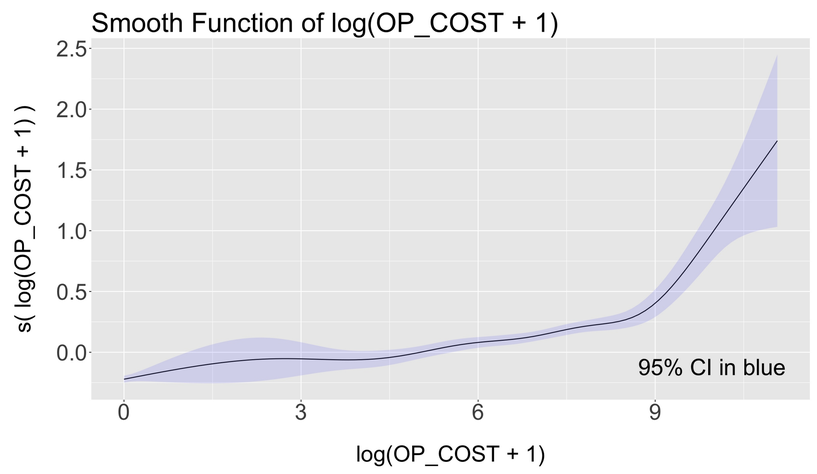


Figure 2: This is an example of a smooth predictor function used in one of our general additive models. This function is nonlinear and would be difficult to parameterize in a generalized linear model but can be automatically fit using a GAM.

Random forest models are nonparametric models that make predictions based on the average predictions of many simple, moderately uncorrelated regression trees. These models are advocated in previous research by Duncan et al. [2] from an accuracy standpoint. However, they can be difficult to interpret.

Our generalized additive models also use a 2-part framework which has been advocated in previous healthcare cost research by Frees et al. [3]. The first stage in our 2-part framework uses logistic regression to separate zero and nonzero cost patients. The second stage is used to make conditional cost predictions for predicted nonzero cost patients.

We evaluate all models using a variety of error metrics as defined in Equation 2 where predicted costs, $\hat{\mathbf{Y}}$ are compared to actual costs, $\mathbf{Y}$. Root mean square error ($\mathbf{RMSE}$) is a commonly used metric, but in the case of health care data mean absolute error ($\mathbf{MAE}$) and mean absolute proportional error ($\mathbf{MAPE}$) can be more useful since they are less sensitive to outliers (high cost patients) in the data. Quantile-truncated metrics do not account for prediction errors for very high cost patients and by design are robust to outliers.

$$\begin{aligned} RMSE &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2} = ||\mathbf{Y} - \hat{\mathbf{Y}}||_2 \qquad & qRMSE_\alpha &= \sqrt{\frac{1}{(1-\alpha)N}\sum_{i:Y_i<q_{1-\alpha}}(Y_i - \hat{Y}_i)^2} \\ MAE &= \frac{1}{N}\sum_{i=1}^{N}|Y_i - \hat{Y}_i| = ||\mathbf{Y} - \hat{\mathbf{Y}}||_1 & qMAE_\alpha &= \frac{1}{(1-\alpha)N}\sum_{i:Y_i<q_{1-\alpha}}|Y_i - \hat{Y}_i| \\ MAPE &= \frac{1}{N}\sum_{i:Y_i>0}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \end{aligned} \tag{2}$$

All computational analysis was performed in RStudio using the following packages: `dplyr`, `Hmisc`, `mgcv`, `ranger`, `ROCR`, and `tidyverse`.

Results

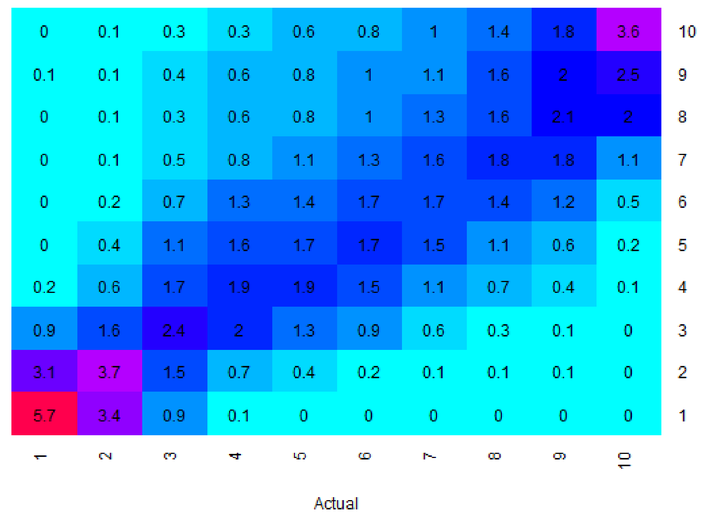
We use 10-fold cross-validation to assess each model's performance across all error metrics. For all 2-part models, we also check the accuracy of the first stage logistic model, which has an area under the receiver operating curve of 99%. This means the first stage model is extremely accurate for predicting nonzero cost patients. It is imperative that the 2-part models have an accurate first stage model as all conditional coast predictions are contingent on the first stage predictions. We confirm that this is indeed the case due to a natural separability in the data between zero and nonzero cost patients.

10-Fold Cross-Validation

Model	R^2	$RMSE$	$qRMSE_\alpha$	MAE	$qMAE_\alpha$	$MAPE$
CMS Linear Model	0.23	9,187.14	4,883.49	4,823.37	3,500.03	3.28
	(0.02)	(304.26)	(91.58)	(90.11)	(66.75)	(0.19)
Full Linear Model	0.29	8,839.77	4,673.08	4,552.23	3,290.28	2.34
	(0.02)	(315.14)	(80.93)	(72.19)	(49.10)	(0.09)
Tweedie Model	0.27	8,938.99	4,730.10	4,607.58	3,322.60	1.97
	(0.02)	(337.57)	(148.88)	(106.68)	(72.43)	(0.09)
2-Part Linear GAM	0.29	8,806.76	4,624.28	4,510.73	3,240.05	2.05
	(0.01)	(402.00)	(109.83)	(116.15)	(54.28)	(0.10)
2-Part Gamma GAM	0.26	8,997.26	4,768.15	4,612.88	3,323.96	2.1
	(0.03)	(360.15)	(145.88)	(116.24)	(65.38)	(0.11)
2-Part Lognorm GAM	0.24	9,133.36	4,167.11	3,875.35	2,407.55	0.97
	(0.02)	(358.93)	(132.81)	(104.08)	(54.23)	(0.03)
2-Part Random Forest	0.29	8,799.57	4,633.97	4,513.05	3,247.79	1.93
	(0.02)	(277.63)	(67.18)	(69.34)	(47.38)	(0.11)

Standard deviation of metrics are shown in parenthesis. Quantile-truncated error metrics use $\alpha = 0.05$. GAM indicates that a model is a generalized additive model. "CMS Linear" denotes the linear risk score model used by CMS. The Tweedie and Gamma models use non-gaussian link functions.

A quick comparison of the error metrics shows that the 2-part lognormal generalized additive model performs best on average. This model uses log 2009 expenditures as the response variables and log-transformed predictor variables. We closely examine the predictions of this model compared to the currently implemented CMS linear model in Figure 3.



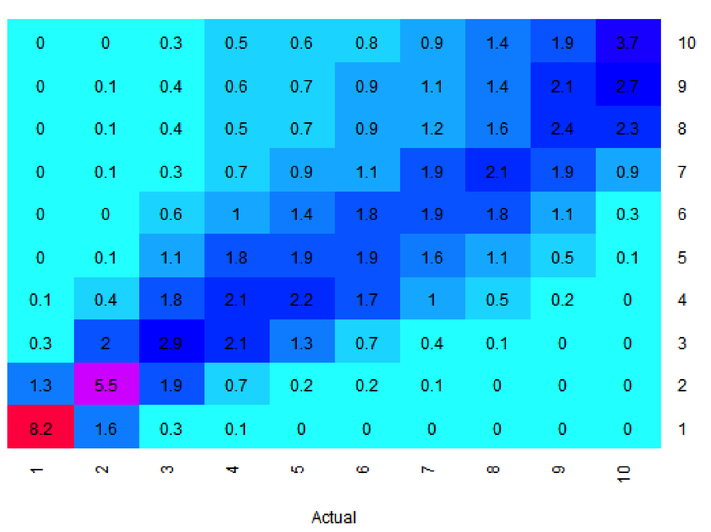


Figure 3: The decile-decile plots compare predicted patient cost deciles to actual cost deciles. Ideally all mass should lie along the diagonal line. The 2-part lognormal GAM performs is significantly more accurate than the CMS linear model.

Discussion

Based on our analysis, we propose the 2-part lognormal general additive model as the optimal model. There are likely more accurate models among the plethora of complex modeling techniques available today, but these models tend to lose interpretability as their complexity and accuracy increase. A 2-part model has a much more interpretable context in terms of separating zero and nonzero cost patients. The generalized additive framework also allows better fitting of nonlinear relationships in the data. Parameterizing a model may improve interpretability, but it assumes we understand more knowledge about the underlying process than we truly have. Healthcare costs are the results of complex processes, and by utilizing a semi-parametric method such as general additive models, we avoid assuming too much about the underlying expenditure process. We only assume there is some relationship between each predictor in the model, which is fitted based on the data itself, not our assumptions. Due to the 2-part lognormal GAM's improved accuracy, it has the potential to improve Medicare patient risk assessment which can in turn improve healthcare cost mitigation.

References

[1] A. Pope G C. Kautter J. Ingber M J. Freeman S. Sekar R. Newhart C. Evans, M. Evaluation of the cms-hcc risk adjustment model. 2011.

[2] M. Ludkovski M Duncan, I. Loginov. Testing alternative regression frameworks for predictive modeling of health care costs. *North American Actuarial Journal*, 20(1):65 – 87, 2016.

[3] W. Gao J. Rosenberg M A. Frees, E. Predicting the frequency and amount of health care expenditures. *North American Actuarial Journal*, 2011.

Acknowledgements

We thank Professors Ian Duncan and Xiyue Liao (UCSB) for their exceptional guidance through the entire course of this research and Nhan Huynh (UCSB) for providing code to construct the HCC's from ICD-10 codes. This research was partially supported by the Society of Actuaries Centers of Actuarial Excellence 2015 Education Grant Project-Based Research Training in Actuarial Science.

Contact Information

- Email: sroneill@pstat.ucsb.edu