Medicare PUF Analysis

Arjun Srinivasan

March 5, 2019

Introduction

- Dataset: the Basic Stand Alone (BSA) Medicare Claims Public Use Files (PUFs), a 2010 dataset that contains information from Medicare Carrier claims
- Research Question: Infer the relationship between a variety of claim-related factors-from patient sex and age group to diagnoses to type of provider-and the cost of any given claim in the dataset

Literature Review

- Prior analysis of Medicare PUFs:
- Found beneficiaries diagnosed with cancer-even without any other chronic conditions-cost five times more than beneficiaries with no chronic conditions (Prada 2014) [1]
- Looked at factors related to Alzheimer's disease (Prada 2013).[2]

[1] Prada, S. I. (2014), Quantifying the effect of a cancer diagnosis on Medicare payments and use according to new public use files. Cancer, 120: 158-162. doi:10.1002/cncr.28409

[2] Prada, S. I. (2013), Medicare Public Use Files and Alzheimer's disease factors in 2008 and 2010. Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 9(4): 472-474. doi:10.1016/j.jalz.2013.04.509

Dataset Preparation

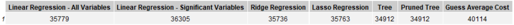
- Transform all relevant categorical and multi-level factors into quantitative binary factors
- Take two small samples (n = 20,000) from the large dataset (n = 70,052,393; for df1 loaded from .csv, n = 2,801,660) as training and test datasets.

Models

- Linear regression with all 66 factors.
- Linear regression w/ 10 significantly related factors
- Ridge regression
- Lasso regression
- Tree
- Pruned Tree

Mean Squared Errors

- Test models on training set sampled from excess observations not in training sample.



	Linear Regression - All Variables	Linear Regression - Significant Variables	Ridge Regression	Lasso Regression	Tree	Pruned Tree	Guess Average Cost
1	35779	36305	35736	35763	34912	34912	40114

Figure 1: Mean Squared Errors

Decision Tree

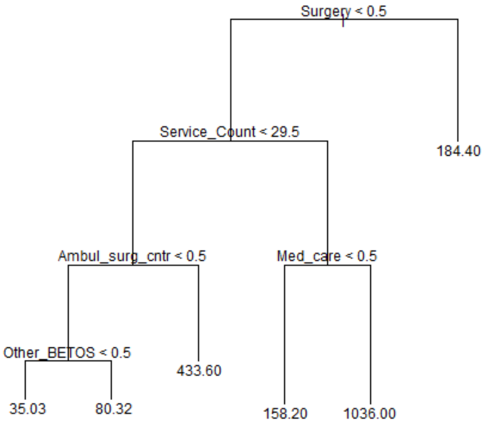


Figure 2: Decision Tree

Conclusion

- Key factors: whether the claim's type of service was surgery, whether the service count was greater than 29.5, whether type of service was medical care (if not surgery), whether the type of service was facility usage of an ambulatory surgical center, and whether the Berenson-Eggers Type of Service code was not one of the 9 most frequent codes (i.e. was label Other_BETOS) were the most determinative factors in inferring the cost of a claim
- Comparing the last mean squared error (mse7) to the mse of the trees shows that trees moderately improve modeling of the relationship between a claim's cost and other factors compared to simply guessing the cost to be the mean of the claims' costs in the sample set
- Focus on the costs and usage of ambulatory surgical centers and costs and incidence of surgery