Modeling Health Expenditures Using Generalized Linear Models

Ziyu Ge, August 2024

1 Introduction

Healthcare costs have surged dramatically: from $74.1 billion in 1970 to $1.4 trillion in 2000, and reaching $4.5 trillion by 2022(McGough et al. 2023). This escalation puts immense pressure on both systems and individuals, highlighting the need for advanced statistical models to inform policies, set insurance premiums, and guide financial planning. Our focus is on the 2003 Cohorts 7 and 8 from the Medical Expenditure Panel Survey (MEPS), where data shows many zeros and a long-tailed distribution for non-zero expenditures. Inspired by my mentor, Professor Yang's work on generalized linear models (GLMs) for cost estimation, this study applies a Two-Part Model (TPM) and Tweedie GLM, incorporating AIC and Lasso for variable selection. Previous research by Frees, Gao, and Rosenberg (2011) extended the TPM to predict healthcare costs(Edward W. Frees and Rosenberg 2011), while other studies have compared GLMs with Tweedie GLMs for aggregate claims(Quijano Xacur and Garrido 2015).

2 Modeling Method

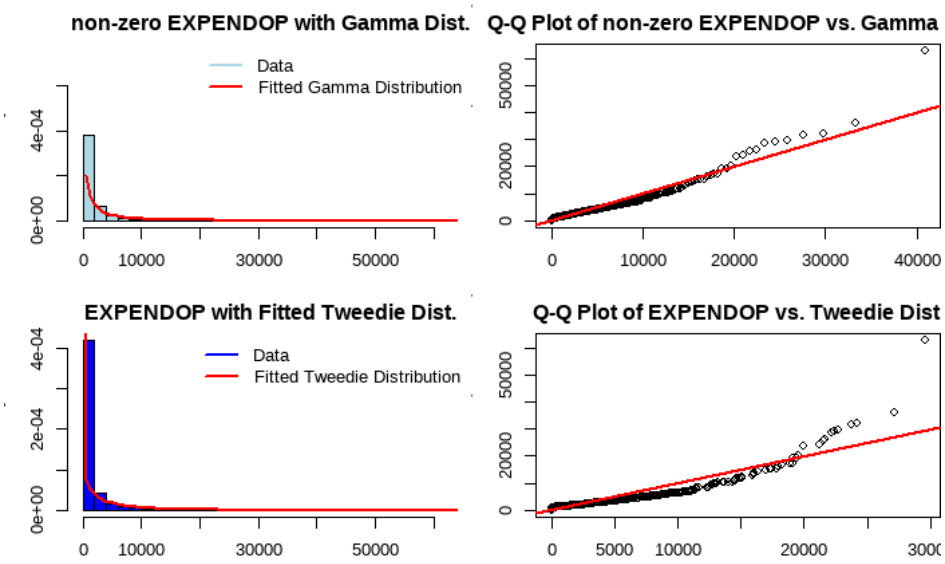

Figure 1: Expenditure Distribution

Our exploratory data analysis (EDA) showed that the outpatient expenditure data has many zero values, with the non-zero expenditures fitting a Gamma distribution (Figure 1, top plots). Given this distribution, we chose a two-part model: logistic regression for predicting the presence of expenditures (i.e., zeros vs. non-zeros) and Gamma regression for modeling the amount of non-zero expenditures.

Given the high-dimensional nature of our data, which includes 16 independent variables, we employed two methods for variable selection: Akaike Information Criterion (AIC) and Lasso regression. AIC is a method used to compare models, where a lower AIC value indicates a better fit. We used AIC for stepwise selection, Using AIC, starting with all variables and then removed the less important ones, ending up with 10 variables for the logistic model and 8 for the Gamma model (as shown in Table 1).

For Lasso regression, which is a technique that adds a penalty to the model based on the absolute value of the coefficients, we used 5-fold cross-validation to select variables. This method helps prevent overfitting by shrinking some coefficients to zero, effectively selecting a simpler model. In the first part, we used logistic

Lasso to model the probability of non-zero outcomes, selecting 14 out of the 16 variables. For the second part, we used Lasso on the Gamma model, resulting in 14 selected variables.

We also used a Tweedie model, which is ideal for data with many zeros and a long tail(Figure 1). Unlike the two-part model, the Tweedie model handles all data in one step. We used both AIC and Lasso to fit and select the Tweedie model.

To ensure robust model evaluation and prediction accuracy, we split the dataset into 80% for training and 20% for validation.

Table 1: Combined Results from TPM and Tweedie GLM

Parameter	Two-Part Model				Tweedie GLM			
	Logistic(AIC)		Gamma(AIC)		Lasso		AIC	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
(Intercept)	-0.801	0.010	5.054	0.001	3.985	0.001	3.930	0.001
GENDER	0.837	0.001	—	—	0.287	0.001	0.314	0.001
USC	1.111	0.001	0.356	0.007	0.741	0.001	0.735	0.001
EDUCHIGHSCH	-0.285	0.083	—	—	-0.113	0.231	-0.125	0.187
EDUCLHIGHSC	-0.434	0.028	—	—	-0.288	0.015	-0.314	0.008
MARISTATMARRIED	0.281	0.196	—	—	—	—	0.341	0.005
MARISTATNEVMAR	-0.347	0.121	—	—	—	—	-0.001	0.995
MARISTATWIDOWED	-0.330	0.487	—	—	—	—	-0.460	0.074
famsize	-0.202	0.001	—	—	-0.066	0.010	-0.110	0.001
INCOMELINCOME	-0.423	0.048	—	—	-0.056	0.666	0.031	0.812
INCOMEMINCOME	-0.382	0.024	—	—	-0.142	0.161	-0.073	0.472
INCOMENPOOR	-0.302	0.307	—	—	-0.097	0.614	-0.033	0.862
INCOMEPOOR	0.177	0.440	—	—	0.210	0.114	0.321	0.018
PHSTATFAIR	0.853	0.001	0.614	0.002	0.873	0.001	0.910	0.001
PHSTATGOOD	0.683	0.001	0.311	0.026	0.499	0.001	0.524	0.001
PHSTATPOOR	16.543	0.971	0.966	0.001	1.241	0.001	1.251	0.001
PHSTATVGOO	0.684	0.001	0.268	0.046	0.466	0.001	0.500	0.001
ANYLIMIT	1.288	0.001	0.549	0.001	0.716	0.001	0.756	0.001
insure	0.737	0.001	0.335	0.025	0.701	0.001	0.674	0.001
MANAGEDCARE	0.301	0.080	—	—	—	—	—	—
AGE	—	—	0.019	0.001	0.022	0.001	0.019	0.001
RACEBLACK	—	—	0.754	0.008	0.693	0.007	0.789	0.002
RACENATIV	—	—	2.122	0.001	2.132	0.001	2.251	0.001
RACEOTHER	—	—	0.883	0.035	1.002	0.006	1.106	0.003
RACEWHITE	—	—	0.811	0.002	0.789	0.001	0.821	0.001
REGIONNORTHEAST	—	—	-0.600	0.001	-0.747	0.001	-0.722	0.001
REGIONSOUTH	—	—	-0.526	0.001	-0.562	0.001	-0.561	0.001
REGIONWEST	—	—	-0.451	0.002	-0.489	0.001	-0.458	0.001
MNHPOOR	—	—	0.298	0.126	0.396	0.006	0.411	0.004

3 Data

The dataset includes 2,000 individuals aged 18 to 65 from the 2003 MEPS panels 7 and 8, with 1,352 reporting positive outpatient expenditures. Independent variables include demographic factors (age, sex, ethnicity), access (geographic region, usual source of care), socioeconomic status, health status, insurance coverage, employment, and industry classification.

4 Results

4.1 Modeling Results for the Two-Part Model (TPM)

Table 1 presents the modeling results for outpatient expenditures using the Two-Part Model (TPM). We focused on the AIC-selected model because its results were similar to those from Lasso selection, offering clearer interpretation.

For predicting whether or not individuals incur expenditures (the frequency of expenditures), demographic factors like gender, access to care factors such as dissatisfaction with the usual source of care, and socioeconomic factors like family size and income were all significant predictors. Health status, particularly self-reported physical health and any physical activity limitations, also played a key role, with insurance coverage being a significant factor associated with more hospital admissions.

When predicting the amount of expenditures, age emerged as a significant predictor, with older individuals generally incurring higher costs. Racial categories, such as Native American status, were linked to higher expenditures. Dissatisfaction with the usual source of care (USC) was associated with increased costs, highlighting the importance of continuous care. Poor physical health status and any functional activity limitations were strong predictors of higher costs, as was having insurance coverage. Additionally, geographic region influenced costs, with individuals from certain regions, like the Middle, incurring higher expenditures.

4.2 Modeling Results for the Tweedie GLM

The Tweedie Generalized Linear Model (GLM), using both Lasso and AIC for variable selection, produced similar results. Since this model accounts for the entire dataset, including zeros (no expenditures), it effectively combines the information about both the likelihood of incurring costs and the amount spent. This makes it comparable to the results from both parts of the TPM.

4.3 Prediction Results

To evaluate the accuracy of our models, we used two common metrics: Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE). Lower values for these metrics indicate more accurate predictions. Table 2 shows the MAPE and MAE values for predicting outpatient expenditures across the different models. For the TPM, the AIC-selected model produced the lowest MAPE and MAE values, indicating the best performance within this framework. For the Tweedie GLM, the Lasso-selected model achieved the lowest error metrics. Overall, when comparing the two models, the Lasso-selected Tweedie GLM shows a clear advantage in prediction accuracy based on the MAPE, with a difference of 14.3% ($14.3\% = (1.75 - 1.50) \div 1.75$). However, the AIC-selected TPM performs slightly better in terms of MAE, showing a 2.8% improvement ($2.8\% = (1264.37 - 1229.32) \div 1264.37$). This suggests that while the Lasso-selected Tweedie GLM excels in overall prediction, the AIC-selected TPM may provide better insight when separating the analysis into the likelihood of incurring expenditures and the amount spent.

	Two-Part Model			Tweedie GLM		
	Full	Lasso	AIC	Full	Lasso	AIC
MAPE	1.85	2.02	1.75	1.61	1.50	1.59
MAE	1246.43	1280.79	1229.32	1287.39	1264.37	1289.63

Table 2: Comparison of Models using MAPE and MAE metrics.

5 Evaluation and Reflective

I am very grateful to my advisor, Prof. Yang Lu, who guided me in expanding my knowledge at my level, and at the same time was very attentive to help me find out the mistakes I made in my project, so that I know that research is a process of continually revisiting and refining ideas, seeking out what I don't know, and finding new ways to solve problems. I also appreciate the opportunity provided by the UROP, which has fueled my motivation to pursue further scientific research.

References

Edward W. Frees, Jie Gao and Marjorie A. Rosenberg (2011). "Predicting the Frequency and Amount of Health Care Expenditures". In: North American Actuarial Journal 15.3, pp. 377–392. DOI: 10.1080/10920277.2011.10597626. URL: https://doi.org/10.1080/10920277.2011.10597626.

McGough, Matthew et al. (Dec. 2023). "How has U.S. spending on healthcare changed over time?" In: Kaiser Family Foundation. Accessed on August 29, 2024.

Quijano Xacur, Oscar Alberto and José Garrido (2015). "Generalised linear models for aggregate claims: to Tweedie or not?" In: European Actuarial Journal 5.1, pp. 181–202.