Predicting State Medicare Spending Per Beneficiary

Xuefei Qiao

Abstract

This is the report to explore what features may influence the State Medicare Spending Per Beneficiary (MSPB) and how they influence it through picking the best-fit Regression model among one pruned tree, simple linear regression, bagged trees and random forests to forecast the value.

Index Terms

medical spending, simple linear regression, one pruned tree, bagged trees, random forests

---✦---

1. Introduction

The State Medicare Spending Per Beneficiary (MSPB) measure can be used to evaluate medical efficiency and the cost of services performed by hospitals and other healthcare providers. According to the report from KFF (Kaiser Family Foundation), Medicare spending was 15 percent of total federal spending in 2018, and is projected to rise to 18 percent by 2029, and was 21% of total health spending in 2018 in the United States.

It can be seen from the above that MSPB is a huge cost, and the large number of enrollees (92.9 million) also shows that the importance of the Medicare Plan, so I think it is necessary to learn how to forecast the MSPB accurately based on some features' information to better control the medical budget.

2. The Dataset

This is the nationwide small dataset with 51 observations and 12 variables provided by KFF. There is no missing value in the dataset, and 10 out of 12 variables will be used as predictors, and 'MSPB' will be used as a response variable. Figure 1 shows four main categories for predictors, and the description of them.

Dependent Variable - MSPB

State Medicare Spending Per Beneficiary, 2018

Independent Variables

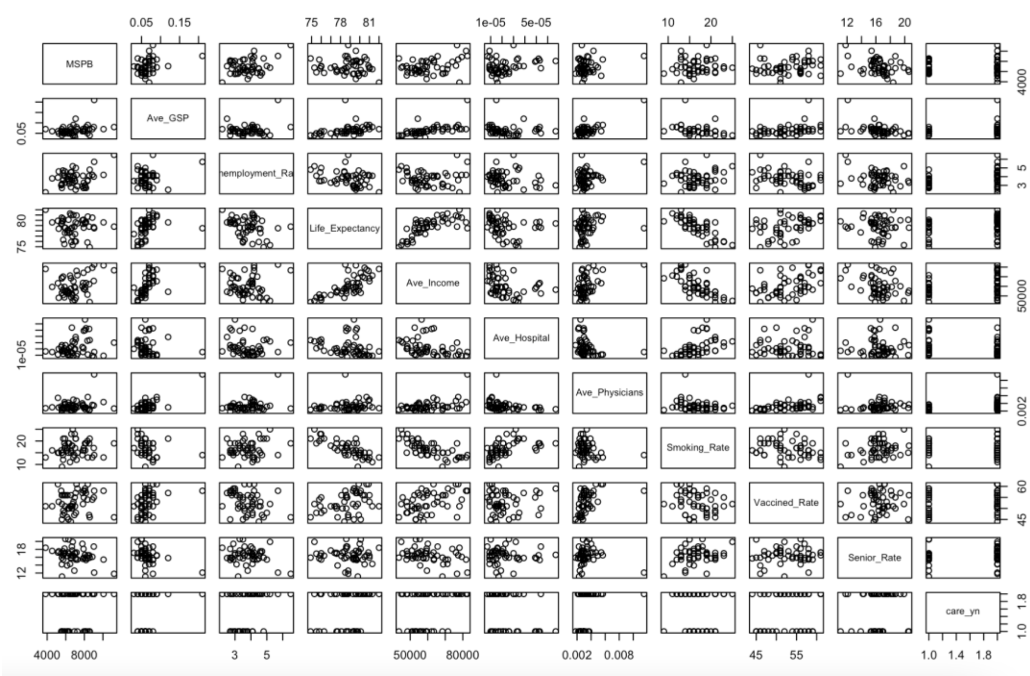
	Independent Variables	Type	Description
Policy	Obamacare	Qualitative	Whether the State Adopts Obamacare or Not, 2018
Economic Status	Ave_GSP	Quantitative	Gross State Product (in million)/Person, 2018
	Ave_Income	Quantitative	2014-2018 Median Annual Household Income (in 2018 dollars)
	Unemployment_Rate	Quantitative	Unemployment Rate, Sept 2018
Medical Condition	Ave_Hospital	Quantitative	Number of Hospitals per Person, 2018
	Ave_Physicians	Quantitative	Number of Professionally Active Physicians per Person, September 2020
Health Status	Life_Expectancy	Quantitative	Life Expectancy at Birth (in years), 2010-2015
	Smoking_Rate	Quantitative	Percentage of Cigarette Use, 2018
	Vaccined_Rate	Quantitative	Vaccination Rate, 2018-2019
	Senior_Rate	Quantitative	Persons Age 65 and Older as a Percentage of Total Population, 2018

Source: Kaiser Family Foundation

(figure 1)

3. Exploratory Data Analysis

Based on figure 2, Ave_Income and Life_Expectancy is highly positively correlated, Smoking_Rate and Life_Expentancy are highly negatively correlated, Smoking_Rate and Ave_Income is highly negatively correlated, Smoking_Rate and Ave_Hospital is positively correlated, so Ave_Income, Life_Expectancy, Smoking_Rate may cause the problem of multicollinearity in the models that I'll create later.



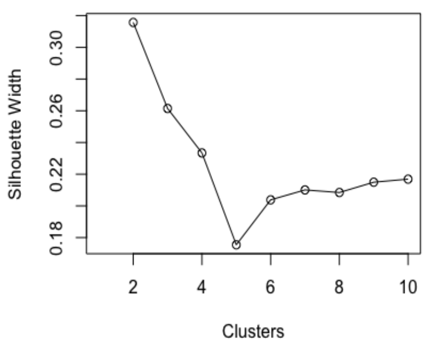
(figure 2)

Given that my predictors include both qualitative and quantitative variables, I'll first use the Gower+PAM and MCA+KMeans method to do clustering for 51 observations, then I'll remove the only qualitative variable 'Obamacare' to do PCA+KMenas hybrid cluster approach for the remaining quantitative variables. Lastly, I'll use the rand index and an unsupervised tree to check the agreement among those cluster methods.

- **Gower+PAM**

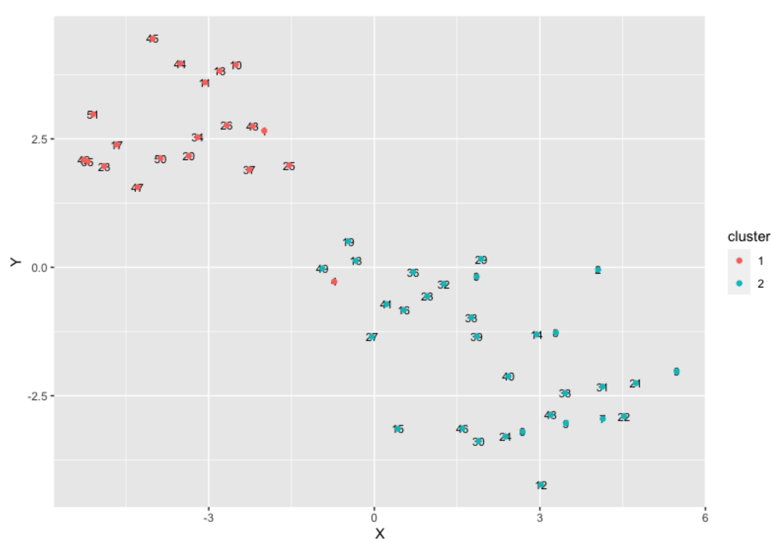
Figure 3 shows k=2 is the best number of clusters for the Gower+PAM method, because it corresponds to the biggest silhouette width, and figure 4 shows the first cluster has 20 observations, and the second cluster has 31 observations. #26 Missouri and #14 Illinois are representatives of cluster 1 and 2 respectively, the first obvious difference between them is whether this state adopts Obamacare, and the second one is the difference on Ave_Hospital.

Figure 5 shows how those two clusters look like in the scatter plot, #2 Alaska, #4 Arkansas, and #12 Hawaii are a little far from their clusters, maybe they are outliers in their respective clusters.




```
> pam_medical=pam(transformed.medical, diss = TRUE, k = 2)
> pam_medical$clusinfo
     size  max_diss   av_diss  diameter separation
[1,]   20 0.2218777 0.1453149 0.3998756 0.07679236
[2,]   31 0.3245904 0.1506053 0.5246600 0.07679236
> gowerPAMclustering=pam_medical$clustering
> data[pam_medical$medoids, ]
      MSPB Ave_GSP Unemployment_Rate Life_Expectancy Ave_Income Ave_Hospital Ave_Physicians
26 6378.24    0.05               3.2            77.6      53560      2.1e-05       0.003592
14 7851.69    0.07               4.1            79.3      63575      1.5e-05       0.003638
   Smoking_Rate Vaccined_Rate Senior_Rate care_yn
26           19            50        16.9      No
14           16            52        15.6     Yes
```

(figure 3) **(figure 4)**

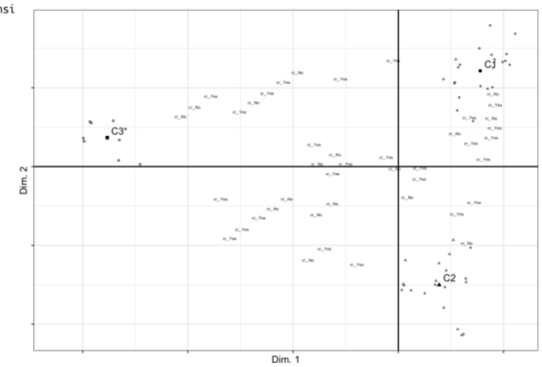


(figure 5)

- **MCA+KMeans**

Furthermore, I use the MCA+KMeans method to determine the number of clusters. Figure 6 shows the best combination is 3-cluster in 2-dimension. Figure 7 presents how the three clusters look in two dimensions.

```
The best solution was obtained for 3 clusters of sizes 20 (39.2%), 18 (35.3%), 13 (25.5%)
 in 2 dimensions, for an average Silhouette width value of -0.027.

Cluster quality criterion values across the specified range of clusters (rows) and dimensi
ons (columns):
      X2     X3     X4     X5
3 -0.027
4  -0.07 -0.058
5 -0.072  -0.09 -0.094
6 -0.085 -0.119 -0.108 -0.087
7  -0.15 -0.118  -0.13 -0.153

The average Silhouette width values of each cluster are:
[1] -0.03 -0.03 -0.02

Cluster centroids:
            Dim.1   Dim.2
Cluster 1  0.1635 -0.0380
Cluster 2 -0.1401 -0.1134
Cluster 3 -0.0577  0.2155

Within cluster sum of squares by cluster:
[1] 0.0369 0.0380 0.0209
 (between_SS / total_SS =  94.94 %)

Objective criterion value: 12.7818
```



(figure 6)

(figure 7)

- **PCA+KMeans**

To use this cluster method, I removed the ‘Obamacare’ qualitative variable, and figure 8 shows 6-cluster in 4-dimension is the best PCA+KMeans combination cluster method.

```
The best solution was obtained for 6 clusters of sizes 15 (29.4%), 13 (25.5%), 8 (15.7%), 7 (13.
7%), 7 (13.7%), 1 (2%) in 4 dimensions, for an average Silhouette width value of 0.218. Variables
 were mean centered and standardized.

Cluster quality criterion values across the specified range of clusters (rows) and dimensions (col
umns):
     X2    X3    X4    X5
3 0.211
4 0.147 0.171
5 0.128   0.2 0.155
6 0.122 0.152 0.218 0.211
7 0.121 0.158 0.206 0.205

The average Silhouette width values of each cluster are:
[1]  0.21  0.18  0.28 -0.02  0.34  0.00

Cluster centroids:
            Dim.1   Dim.2   Dim.3   Dim.4
Cluster 1 -0.8914  0.5099 -0.7402 -0.4042
Cluster 2  1.5383  1.0000  0.2604 -0.3572
Cluster 3 -2.5187 -1.2844 -0.0717 -0.4306
Cluster 4  1.5243 -0.5189 -0.9047  1.3453
Cluster 5 -0.5420 -0.2060  2.0945  0.9784
Cluster 6  6.6453 -5.2999 -0.0369 -2.1153

Within cluster sum of squares by cluster:
[1] 29.4310 32.4484 14.6903 33.2316 11.3779  0.0000
```

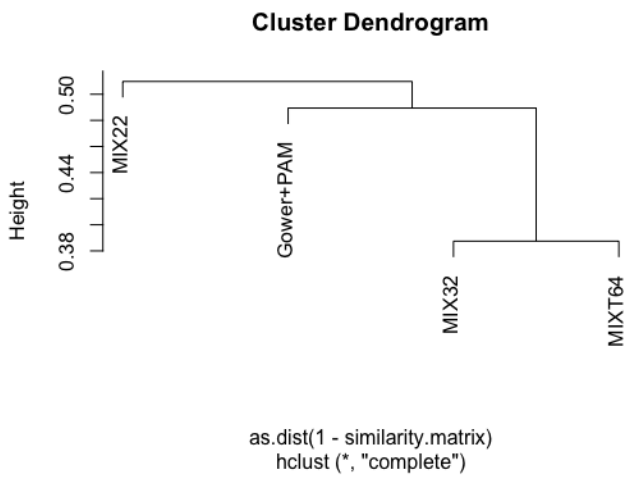
(figure 8)

- **Rand Index & Unsupervised Tree**

I use the rand index and the unsupervised tree to check the agreement among all above cluster methods. Both figure 9 and 10 shows that the most two similar cluster methods are MCA+KMeans (3-cluster in 2-dimension) and PCA+KMeans (6-cluster in 4-dimension). Plus, the small difference among all cluster methods means the dataset can be clustered well.

```
> ##Comparison
> rand.index(gowerPAMclustering,as.vector(mix32.clustering)) #0.52
[1] 0.5152941
> rand.index(gowerPAMclustering,as.vector(mix22.clustering)) #0.49
[1] 0.4901961
> rand.index(as.vector(mix32.clustering),as.vector(mix22.clustering)) #0.50
[1] 0.5011765
> rand.index(as.vector(mix32.clustering),as.vector(mixt64.clustering)) #0.61
[1] 0.612549
```

(figure 9)



(figure 10)

4. Theory and Methods

I'll mainly use Regression models including simple linear regression model, one pruned tree, bagged tree, and random forests.

- **Linear Regression**
  - It is created based on the rule of minimizing the sum of squared of residuals, the beta parameter is used to control how y_hat changes on average. In the model, there are estimates of parameters, and the p-value of each estimate determines whether the parameter has influence on y_hat.
  - The assumption includes those errors are independent of each other, the error term is normally distributed, no heteroscedastic problem (variance of error keeps constant) and the mean of error is equal to zero.
  - My hypothesis is each parameter has an influence on y.
- **One Pruned Tree**

- The parameter is a complex parameter. In the Regression problem, the one that can reduce the variance of data points best in the node will be used to split the node. In a classification problem, the one that can reduce entropy best in the node will be used to split the node.
- No assumption needed.

- **Bagged Trees**

- It doesn't have a parameter. The forecasted value in the Regression problem is the average of all predicted results of trees and the majority in the Classification problem. Each tree in bagged trees is like a single tree, and each tree is correlated because of the method that through comparing all features to select one to split the node.
- No assumption needed.

- **Random Forests**

- The parameter is the number of features. Through bootstrapping, each node determines how to split based on comparing with part of features, not all, and each tree in random forests is independent, so there is no correlation between each tree.
- No assumption needed.

5. Regression Modeling

Given that the data type of my response is quantitative, and my predictors have mixed data types, I'll firstly pick the winner among regression models including one pruned tree, bagged trees, random forests, and a simple straight line on cross-validation, then use the winner in practice and check its assumption.

- **Selection of the Best Fitted Model**

Based on figure 11, I think the single linear regression model is the winner with the highest R-squared (0.53) and the lowest RMSE (1234) among the four models on cross-validation. Figure 12 shows both RMSE (1689) and R-squared (0.47) for this model get worse in practice, it supposed to be fit better in the test dataset, I think a small size of the dataset with a large number of features may cause this situation.

Plus, from figures 11 and 12 we can see RMSE and R-squared on both bagged tree and the random forest did a better job in practice, I think the bagging method help solve high-variance problems to avoid them overfit on the validation dataset, so they can perform better on the test dataset. On the other hand, R-squared for the one-pruned tree on the test dataset doesn't improve, because it has a high-variance problem on the validation dataset, the overfitting on the validation dataset causes the bad performance on the test dataset.

```
> ##########
> #--One pruned tree--#
> ##########
> cv.tree$results
          cp     RMSE  Rsquared      MAE   RMSESD RsquaredSD    MAESD
1 0.03508263 1439.647 0.3553618 1170.089 676.3806  0.3150798 514.6158
2 0.15991014 1540.227 0.2197710 1231.838 605.6690  0.2155502 454.2051
3 0.19553531 1494.196 0.2496460 1218.408 609.9211  0.2624929 482.8930
> system.time(train(x=trans.tr.pred,y=tr.response,method='rpart',trControl=ctrl))
   user  system elapsed
  1.195   0.026   1.251


> ###################
> #--Next, a bagged tree---#
> ###################
> cv.bagged.tree$results
  parameter     RMSE  Rsquared      MAE   RMSESD RsquaredSD    MAESD
1      none 1264.524 0.4398577 1040.146 607.9136  0.3185713 480.1155
> system.time(train(x=trans.tr.pred,y=tr.response,method='treebag',trControl=ctrl))
   user  system elapsed
  4.718   0.047   4.823


> ###################
> #--Next, a random forest---#
> ###################

> cv.randomforest$results
  mtry     RMSE  Rsquared      MAE   RMSESD RsquaredSD    MAESD
1    2 1311.030 0.3361168 1086.176 544.1125  0.3003070 427.9107
2    6 1308.818 0.3370224 1086.339 549.3623  0.3107798 420.9639
3   10 1311.561 0.3392422 1086.420 559.8351  0.2959040 422.4668
> system.time(train(x=trans.tr.pred,y=tr.response,method='cforest',trControl=ctrl))
   user  system elapsed
  5.062   0.069   5.150


> ##################
> #---A simple straight line---#
> ##################
> cv.linearreg$results
  intercept     RMSE  Rsquared      MAE   RMSESD RsquaredSD    MAESD
1      TRUE 1234.477 0.5324037 1029.954 513.7078  0.3380201 389.9836
> system.time(train(x=trans.tr.pred,y=tr.response,method='lm',trControl=ctrl))
   user  system elapsed
  0.847   0.011   0.869
```

(figure 11)

```
> ##########
> #--One pruned tree--#
> ##########
> Acc.tree
      RMSE    Rsquare
1 1278.802 0.09585777

> ###################
> #--Next, a bagged tree---#
> ###################
> Acc.TreeBagged
      RMSE   Rsquare
1 915.2019 0.5182582

> ###################
> #--Next, a random forest---#
> ###################
> Acc.RF
      RMSE   Rsquare
1 931.3436 0.6442914

> ##################
> #---A simple straight line---#
> ##################
> Acc.Line
      RMSE   Rsquare
1 1689.069 0.4666997
```

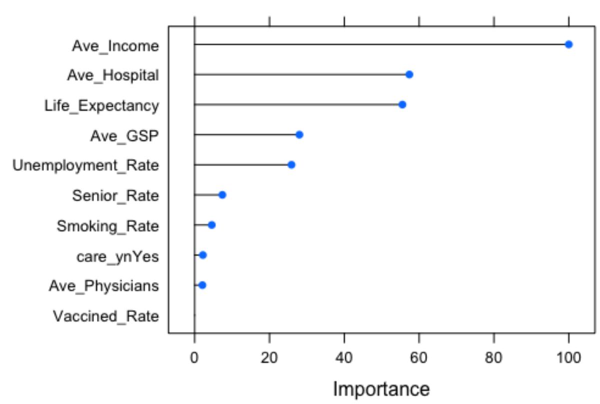
(figure 12)

- **Interpretation for the Best Fitted Model**

Based on figure 13 and 14, we can see Ave_Income, Ave_Hospital and Life_expctancy are three important variables to determine MSPB, and I can write the equation on figure 15 to conclude that when other variables keep fixed, every one unit on Ave_Income, the estimated MSPB will increase by $1509.02 on average; keep other variables keep fixed, every one unit on Ave_Hospital, the estimated MSPB will increase by $735.3 on average; keep other variables keep fixed, every one unit on Life_Expentancy, the estimated MSPB will decrease by $1290.95 on average.

```
> summary(cv.linearreg$finalModel)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-2546.1  -586.6   136.7   534.7  1913.0

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        7022.74     389.19  18.044   <2e-16 ***
Ave_GSP             335.01     265.69   1.261   0.2177
Unemployment_Rate   314.40     268.12   1.173   0.2508
Life_Expectancy   -1290.95     534.88  -2.414   0.0226 *
Ave_Income         1509.02     352.98   4.275   0.0002 ***
Ave_Hospital        735.30     295.15   2.491   0.0189 *
Ave_Physicians       51.07     290.72   0.176   0.8618
Smoking_Rate       -126.87     452.30  -0.281   0.7811
Vaccined_Rate        26.00     292.59   0.089   0.9298
Senior_Rate          98.91     248.11   0.399   0.6932
care_ynYes          -95.44     527.60  -0.181   0.8578
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1040 on 28 degrees of freedom
Multiple R-squared:  0.6179,	Adjusted R-squared:  0.4815
F-statistic: 4.528 on 10 and 28 DF,  p-value: 0.0007473

  0.814   0.009   0.831
```

(figure 13)



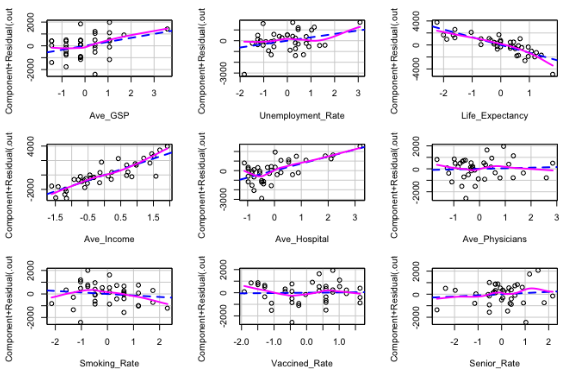
(figure 14)

$$\widehat{MSPB} = 7022.74 + 1509.02 \cdot Ave\_Income + 735.3 \cdot Ave\_Hospital - 1290.95 \cdot Life\_Expectancy$$

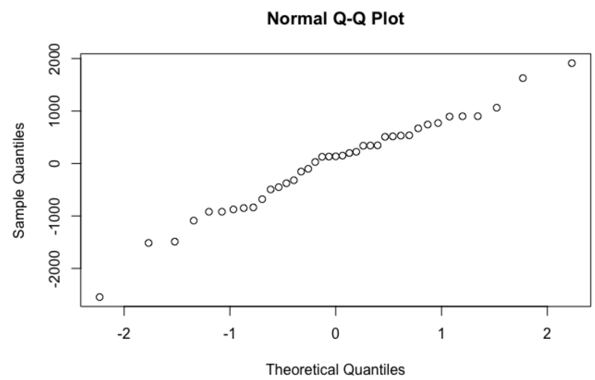
(figure 15)

- **Assumption Check**

Based on figure 16, there is a heteroscedastic problem on the partial residual plot for the variable 'Ave_Hospital', the variance of residuals in that plot is not constant and the mean of residuals is not equal to zero. Figures 17 and 18 don't show any serious problem with the linearity of residuals and independence among residuals (p-value is greater than 0.05).



(Figure 16)



(figure 17)

```
> dwt(cv.linearreg$finalModel)
 lag Autocorrelation D-W Statistic p-value
   1      -0.1729614      2.269449    0.45
 Alternative hypothesis: rho != 0
```

(figure 18)

6. Conclusions and Future Work

In conclusion, annual household income, number of hospitals per person, and life expectancy at birth may influence the MSPB, and the percentage of senior people does not have a significant effect on MSPB. Obamacare does not have a significant effect on MSPB; The best way to lower MSPB without harming the healthcare system and state economy is by improving overall health.

For further analysis in the future, I'll solve the multicollinearity problem based on the findings I have in the EDA process (Ave_Income, Life_Expectancy, Smoking_Rate) through removing one of them that isn't important (Smoking_Rate) and need to investigate outliers (#2 Alaska, #4 Arkansas, and #12 Hawaii) that I found in EDA process too to determine if I should keep them or remove them. In the process of modeling, I'll add Boosted model on cross-validation to check if it does better than the single linear model, and I may improve my current single linear model to an advanced linear regression model to solve the problem shown in the partial residual plot.

7. Reference

Juliette Cubanski, Tricia Neuman, and Meredith Freed, *The Facts on Medicare Spending and Financing*, KFF, Aug. 2019.