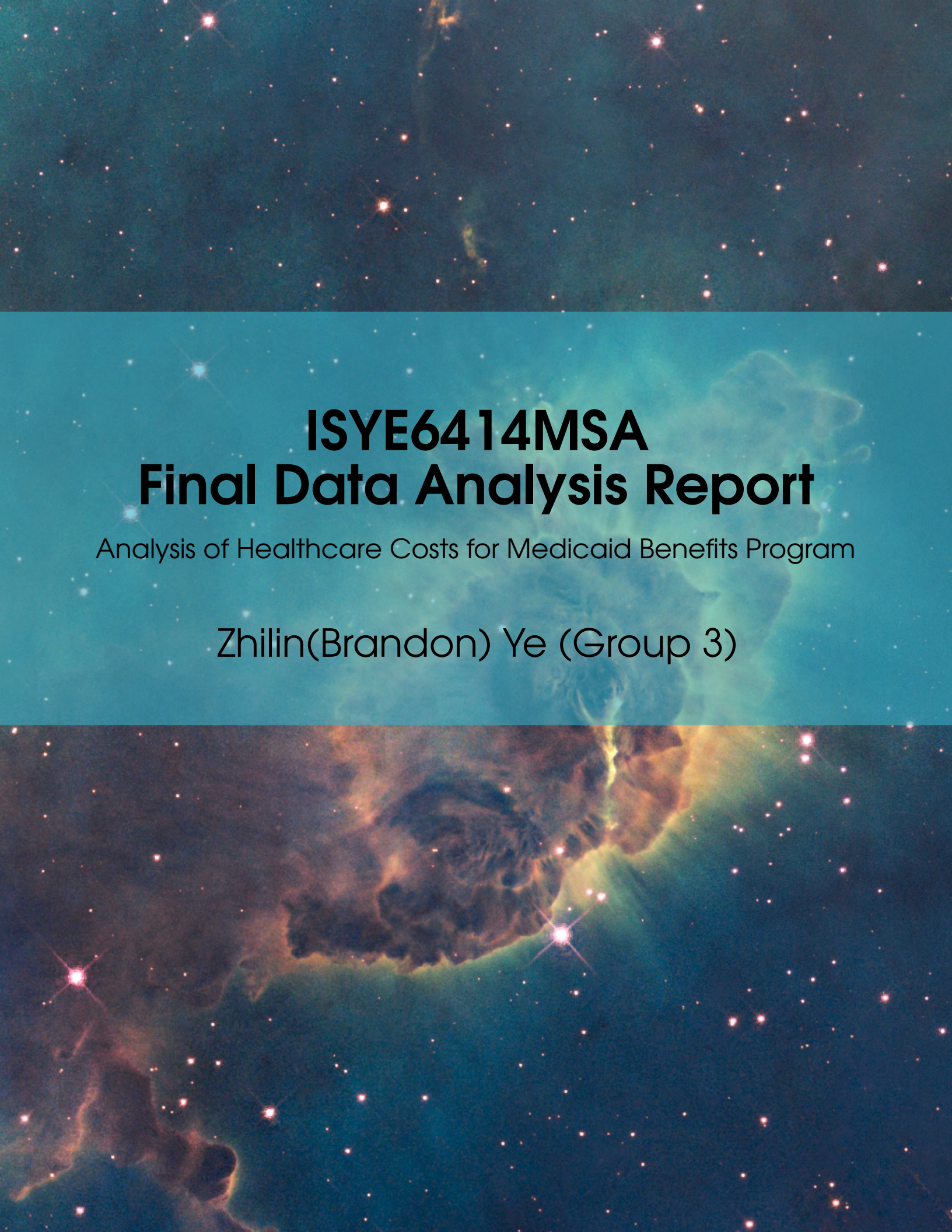


ISYE6414MSA
Final Data Analysis Report

Analysis of Healthcare Costs for Medicaid Benefits Program

Zhilin(Brandon) Ye (Group 3)

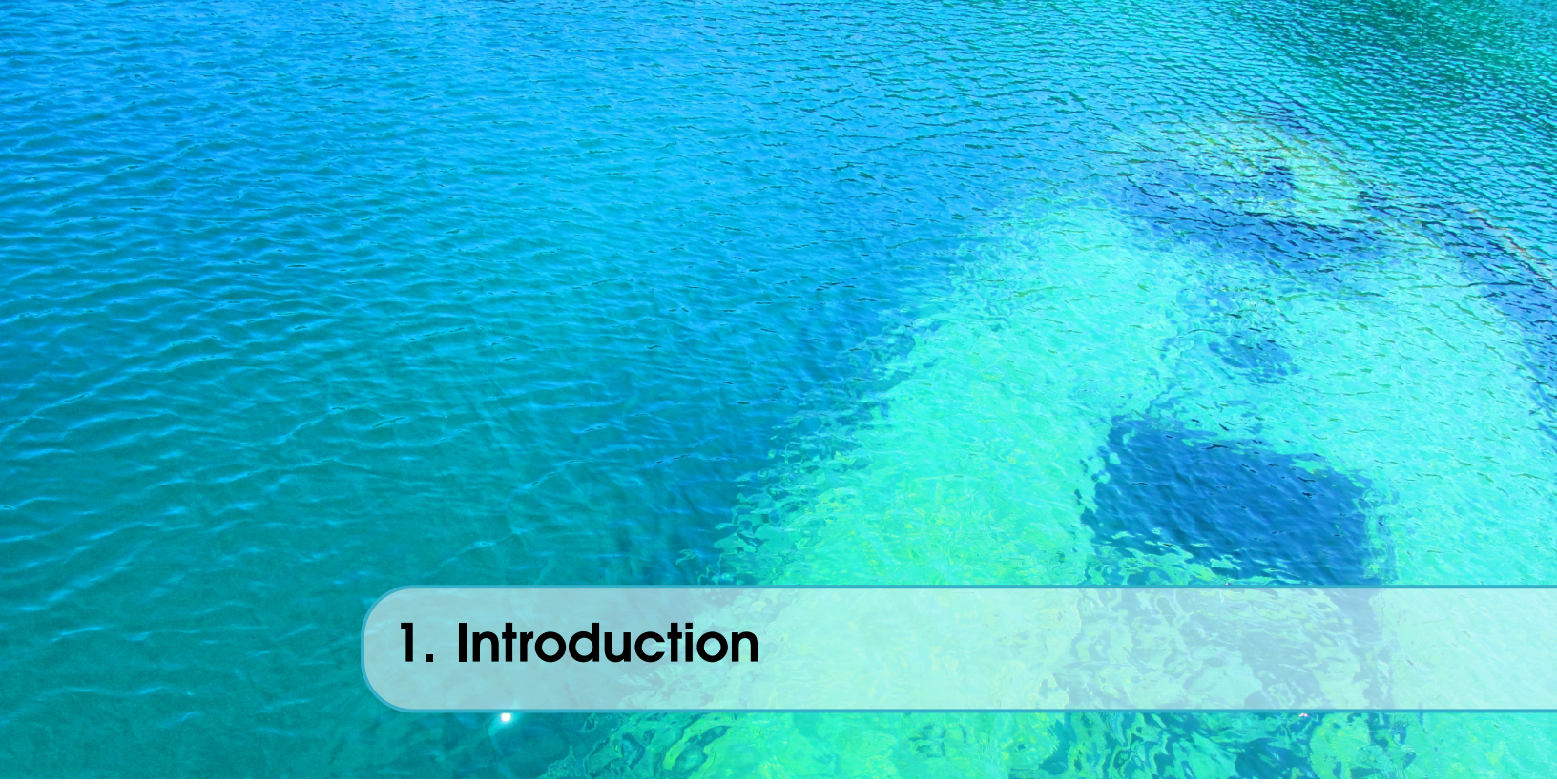
Dec.2016

MEDICAID BENEFITS PROGRAM in the UNITED STATES is a social health care program for families and individuals with limited resources.(https://en.wikipedia.org/wiki/Medicaid). A fact in this program is that about 5% of the population of Medicaid beneficiaries drive up to 50% of the total spending. This may imply a problem of abuse.

In this report, it aims to analyze data through regression models that could aid in investigating potential contributing factors to the cost of this program. Models are built for the cost for emergency department (ED) encounters and physician office (PO) visits separately. Those models contain: Both Stepwise Regression Model, Lasso Regression Model, Logistic Lasso Regression Models. The Both Stepwise Regression Models and Lasso Regression Models are built with interaction terms between State and several factors, while the Logistic Lasso Regression Models are built controlling the State effect. Since same methods are used to build models for both response(ED Cost and PO Cost), it is possible to compare them to find out major contributing factors. Such factors include variables like: White Medicaid-enrolled Adults, Healthy Medicaid-enrolled Adults, Unemployment rate.

With the identification of these important factors, it is possible to notice the existing operational issues of this program. Then a suggestion has been made to improve the Medicaid program, like reducing the cost and balancing the expenditure on different groups of people.

Contents



1. Introduction

1.1 Describe the data

The primary source of data for this analysis is the 2011 Medicaid analytic extract (MAX) file for the states of Alabama, Arkansas, Louisiana, and North Carolina. The description of the data is already in details in the assignment. This study focus on the cost measured only for Adults.

The Primary Outcome Variables of our study is ED Cost and PO Cost in 2011, and they need to be scaled by the total number of enrollment months in year 2011.

The Controlling and Explanatory Variables can be divided into four groups: *Location Information*, *Healthcare Utilization Factors*, *Study Population Characteristics*, *Socio-Economic and Health Environment Factors*. In total there are 27 predictors.

The last two group of variables can be redevided into the following 5 groups: *Population*, *Financial Factors*, *Social Factors*, *Health Environment*, *life*. It is shown in the Table 1.1.

Population	Financial Factors	Social Factors	Health Environment	life
WhitePop	Unemployment	Education	Accessibility	RankingsFood
BlackPop	Income	Urbanicity	Availability	RankingsHousing
OtherPop	Poverty		RankingsPCP	RankingsExercise
HealthyPop			ProvDensity	RankingsSocial
ChronicPop				
ComplexPop				
TotalChild				
TotalAdult				

Table 1.1: Regroup the Variables

1.2 Goal and Methodology

This study aims to analyze the operation situation of the Medicaid program and propose guidelines to reduce the Medicaid cost by performing a regression analysis that identifies the important factors which contribute to the cost. Therefore, the goal of this study is explanatory as opposed to predictive. Hence a larger model is allowed.

In order to accomplish the goal of this study, a number of regression models for ED Cost and PO Cost separately were built and compared. By comparing these models, the main influencing factors, which have a statistical significance, were found. The results were then interpreted accordingly.

The road map of how to reach this goal is as follow:

- Focus on building models for ED Cost at first.
- Explore the data. Do a quick and dirty model to notice if there are certain problems that need to be fixed at the very beginning. Then do Generalized Additive Models analysis to provide guideline for transformation of predictors. After that, do boxplot or scatterplot appropriately to suggest any kinds of transformation.
- Build a stepwise regression model using AIC and both direction. Seek for further improvement of the model. Do partial F-test if there are certain variables that may or may not be included in the model. Notice the outliers and seek for chance to get rid of them, and then come up with a final model. Interpret.
- Do a Lasso regression on the same maximum model as above. Identify the significant factors.
- Do a Logistic regression controlling the contribution of States. This may provide information of how certain tracts can make the cost lower than state average. Identify the significant factors.

- Compare the above three models, identify the most significant factors.
- Build the same set of models for PO Cost. Compare the above 6 models and identify the most significant factors.

2. ED Cost Models

All the coefficients of the models in this chapter are shown in Appendix.

2.1 Exploratory

This section aims to identify possible transformation on the response and the predictors.

2.1.1 Quick And Dirty Model

I began the exploratory analysis by building a quick and dirty model with scaled ED Cost as response against all the other variables except the GEOID and PO Cost. I decided to exclude GEOID because the number of the GEOID itself makes no sense to be included in the model. There may be some additional information on how this ID is constructed, yet it seems trivial and I decided to ignore that. And the reason why I exclude PO Cost is that after all I need to compare the models on ED Cost and models on PO Cost, therefore I don't want them to interfere each other.

Initially I noticed that there exists a long tail in the qq-norm plot. And in the residuals plot, the variance is not constant along the fitted value. To fix that, I took the logarithm of the response (scaled ED Cost). After taking the logarithm, I redid the quick and dirty model, and found out that the residuals pattern is much better than before, which can satisfy the constant variance assumption.

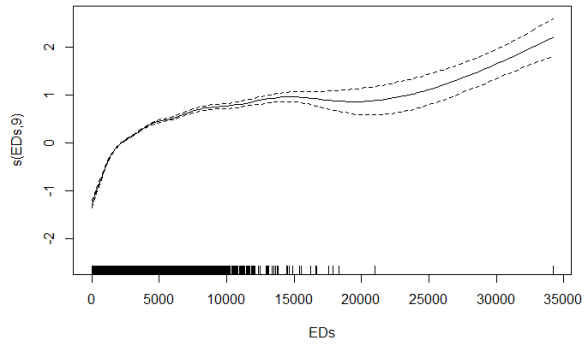
2.1.2 GAM

To provide some clues to transforming predictors, I decided to build a generalized additive model (GAM) with the logarithm of the response. Most of them are not very informative except two as shown in Figure 2.1: for EDs, it suggests a pattern of log-transformation, as Figure 2.1.b; for ComplexPop, the transformation is more complex – it seems like an inverse of logit transformation, as Figure 2.1.d.

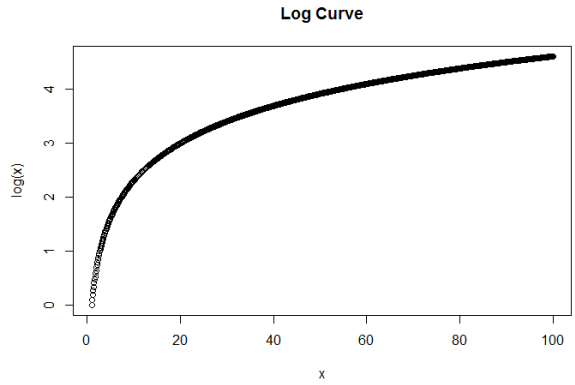
2.1.3 Transformation

After building GAM, I did boxplots or scatter plots to see if there are other possible transformations. It can be noticed that a log transformation can improve the linearity on EDs and POs. But the inverse of logit transformation seems not to change a lot on ComplexPop, but it still can be added to the model and let stepwise regression to choose.

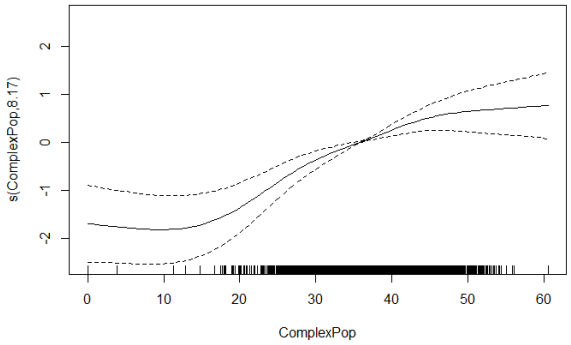
There are another two possible transformations worth being noticed. They are on Accessibility and Availability. Look at Figure 2.2 – both of them have a lot of data points with small values. Therefore, I decided to categorize them. For Accessibility, I transformed it into a 3-level factor variables – Low Congestion: Availability ≤ 0.15; Middle Congestion: $0.15 <$ Availability ≤ 0.5; High Congestion: Availability > 0.5 . For Availability, I transformed it into



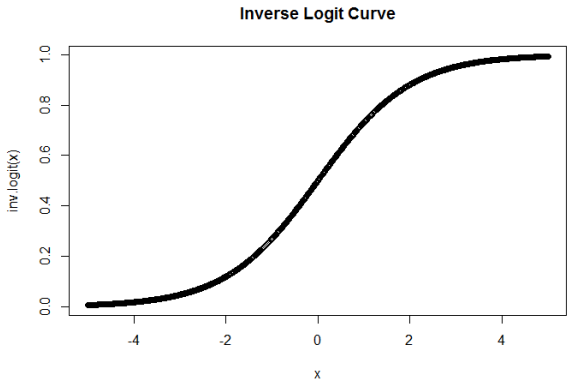
(a) EDs GAM Curve



(b) Log Curve



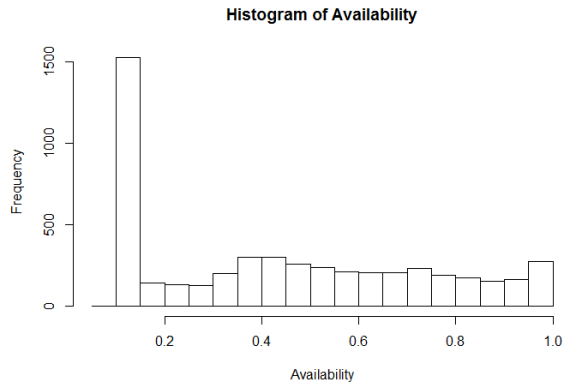
(c) ComplexPop GAM Curve



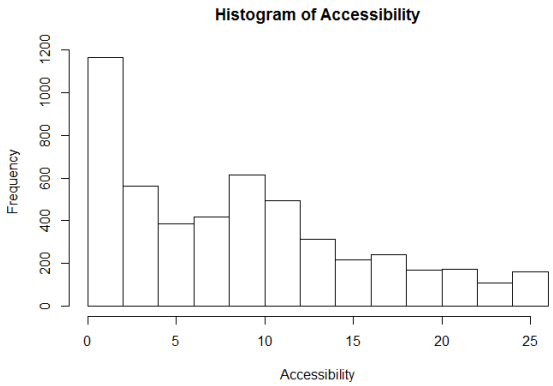
(d) Inverse Logit Curve

Figure 2.1: Smooth Curves Created by GAM on EDs and ComplexPop & the Log Curve and Inverse Logit Curve for Reference.

a 3-level factor variables too – Low Distance: Accessibility ≤ 4; Middle Distance: $4 <$ Accessibility≤ 10; High Distance: Accessibility > 10. I divided them mainly based on making the number of data points in each set equal.



(a) Histogram of Availability



(b) Histogram of Accessibility

Figure 2.2: Histogram of Availability and Accessibility. Both of them have a lot of data points with small values

2.2 Regression Models

In this section, a number of models were built for the goal of this study. They include: Linear Regression Model, Both Stepwise Variable Selection on a Linear Regression Model using AIC, Lasso Regression Model, Lasso Logistic Regression Model. Ridge Regression Model is omitted here, since the main purpose of this study is to identity important factors, and Ridge Regression fails to do that.

I did stepwise regression to identify what factors contributing to the explanatory power in this model. Noticing a strange gap in the residual plot, I added interaction terms of State with other predictors in the model and ended up with stronger explanatory power. The final R^2 is 0.9018 with 46 variables in the model.

After building the stepwise regression model, I built a Lasso regression model with the same maximum model as above and came up with a different set of variables selected. The overlapped predictors should be considered as the important factors.

Also, considering a strong effect from State and its interaction terms on the model, finally I built a Logistic Lasso model controlling the State effect. I categorized the cost into two group – those who reach the state average and those who do not. This interesting prospective provides us a new set of relevant variables which can deepen our understanding of the model.

2.2.1 Both Stepwise Regression Using AIC

I chose to do a Both stepwise variable selection model using the Akaike Information Criterion (AIC). AIC is favored over BIC since the purpose of this study is explanatory and a slightly larger model is allowed (which AIC normally favors). On the other hand, I chose "both stepwise" as opposed to "forward stepwise" and "backward stepwise" as forward and backward seem more likely to be stuck in some worse local optima. To build a 'Maximum Model', I added all the possible transformations into the model, and let stepwise regression to select them.

For more information about the 'Maximum Model', review Chapter 16: Selecting the Best Regression Equation, page 439, from **APPLIED REGRESSION ANALYSIS AND OTHER MULTIVARIABLE METHODS**, *KLEINBAUM | KUPPER | NIZAM | ROSENBERG* , FIFTH EDITION, 2014.

However, the fitted value of the model shows a strange gap as shown in Figure 2.3. The gap between 3.0 and 3.5 in the fitted value plot is obvious. Notice that each bulk of data points is from one state, therefore it is very likely that some interaction effects are ignored in the model.

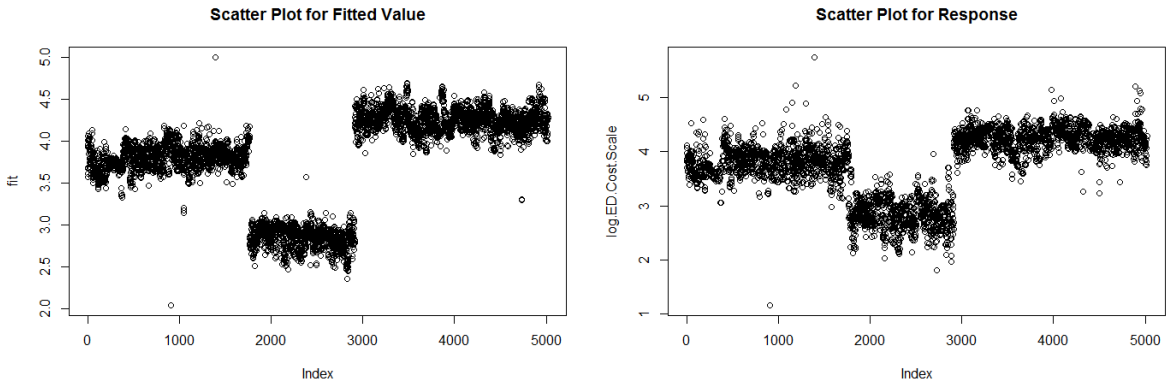


Figure 2.3: In the fitted value scatter plot, a clear gap between 3.0 and 3.5 is observed. Since each bulk of data points belong to one state, some interaction effects may be ignored in the model.

By using R's package `lattice` to do a Trellis Display on ChronicPop by State as shown in Figure 2.4, it can be seen that the pattern for response to ChronicPop is different for each state. Therefore, it is reasonable to add interaction term into the model.

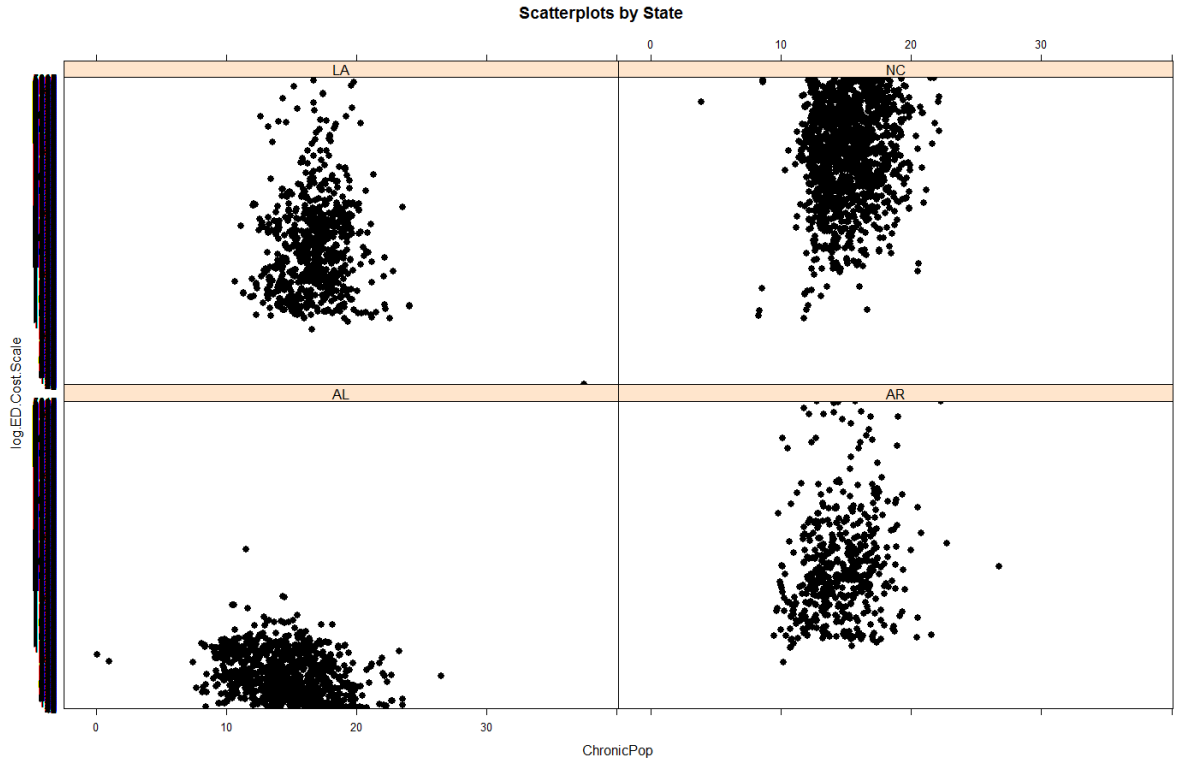


Figure 2.4: Trellis Display on Response vs ChronicPop by State

Using similar methodology as 'largest model', I added as much as interaction term as possible and let stepwise regression to choose them automatically. After that, I identified one potential outlier using Cook's distance plot (Figure 2.5). After removing that, the Cook's distance is within 0.12, and the residual plots (Figure 2.6) are satisfactory though not perfect. Meanwhile, the gap in fitted value plot(Figure 2.7) is much smaller compared to Figure 2.3. The R^2 for this model was around 0.902, indicating that the model explains around 90.2% of the variability in the data.

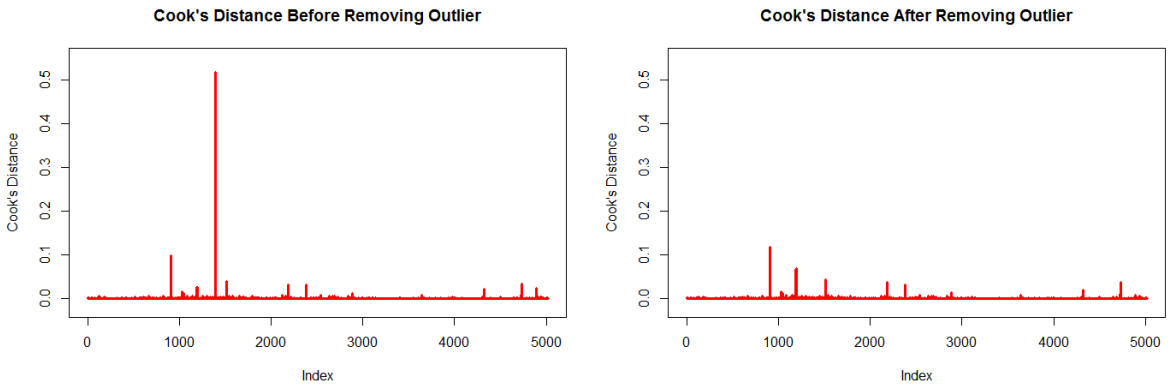


Figure 2.5: Cook's Distance Before and After Removing outliers

Some of the variables that are significant are: Availability, RankingsSocial, Unemployment, RankingsFood, HOs.

For RankingsSocial, I noticed that its coefficient is negative, indicating that controlling for all the other variables in the model, the higher the value for this variable, the lower the ED Cost PMPM is.

For Availability, since both the coefficients of Availability.NewLow and Availability.NewMiddle are negative,

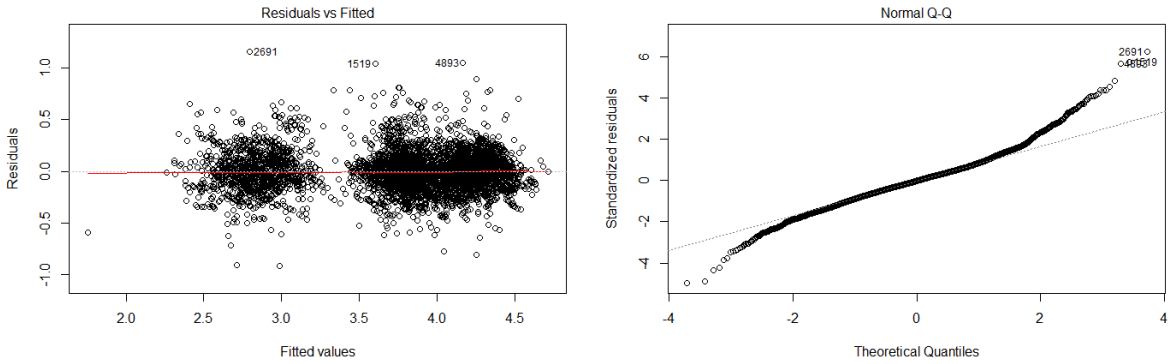


Figure 2.6: Residual plots after adding interaction terms and removing outliers

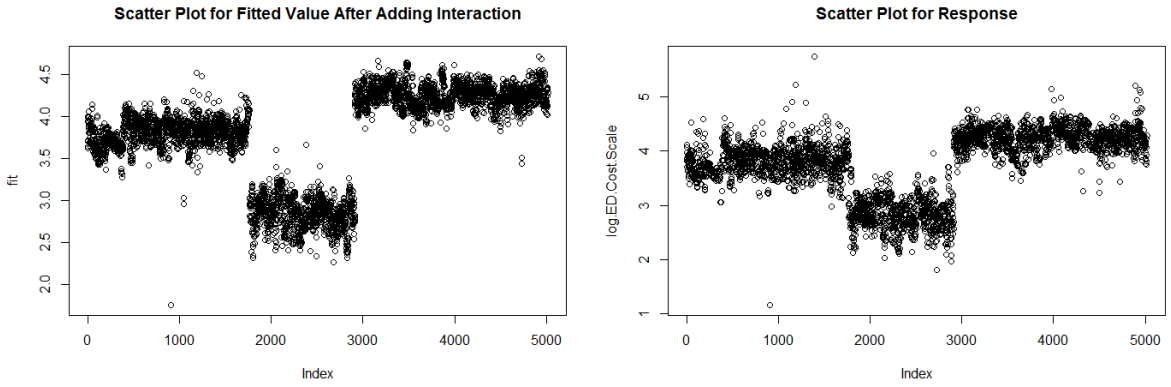


Figure 2.7: Scatter plots for fitted value and response after adding interaction terms. The gap is much less than before.

and the coefficient for Availability.NewLow is more negative than Availability.NewMiddle. This indicates that the lower Availability can lead to lower ED Cost PMPM, controlling for all the other variables in the model, which means a positive effect.

For Unemployment, for the base state AL, the higher the unemployment rate is, the lower the ED Cost PMPM is, controlling for all the other variables in the model. The same for state AR, since it is not statistically significantly different than AL. But for state LA and NC, since their coefficients in the interaction terms are positive, and the absolute value of them is greater than the coefficient of the base, therefore, greater unemployment rate will lead to greater ED Cost PMPM, controlling for all the other variables in the model. This difference may stem from the difference of the policies toward unemployed people between each state.

For HOs, for the base state AL, higher HO can lead to lower ED Cost PMPM, controlling for all the other variables in the model. And for the interaction terms, since all of the coefficients are either not significant or have a positive sign but less absolute value, HOs in general can lead to lower ED Cost PMPM for each state, controlling for all the other variables in the model. And for State LA, compared to AL, the same increase in Hos will lead to less decrease in ED Cost PMPM, controlling for all the other variables in the model.

One interesting fact worth being notice is the coefficient of RankingsFood – it is positive! That means that controlling for all the other variables in the model, the better access to healthy food or food insecurity, the higher ED Cost PMPM is! But thinking about it for a moment, it is not necessarily impossible – with healthier food, people will be more confident with their health and it may be an indication of wealthy in that tracts. Therefore, if a person needs to be enrolled in this program, it may be an indication of worse financial and health condition.

Since I thought RankingsExercise should be important and included in the model, I also performed a partial

F-test to examine whether it should be included. For the test, it is assumed:

- H_0: $\beta_{RE} = 0$, where RE means RankingsExercise, and
- H_A: $\beta_{RE} \neq 0$

By using the ANOVA tables in R, I computed a test statistic, F* = 0.0091 corresponding to a 0.92 p-value. Based on the test statistic, I failed to reject the null hypothesis and concluded that the predictor of RankingsExercise adds no further information to the model.

2.2.2 Lasso

With the same maximum model as above, I did a Lasso Regression Analysis after scaling all the predictors within (0,1). I chose to do lasso using R's `glmnet` package to do the variable selection. The best value for the penalty 'lambda' was chosen using 10-fold cross validation. Then I came up with another set of variables selected which can be considered as important.

2.2.3 Logistic Lasso

From the previous study, it is easy to be noticed that State has a great impact on the ED Cost PMPM, however it doesn't provide much valuable information on how to reduce the cost. Therefore, an intuition is to control the State effect and study the remaining.

Then the residuals after controlling the State effect will become how far the ED Cost PMPM is from the State average for a certain tract. Now it is reasonable to divide the tracts into two groups: those who reach the State average and those do not.

Then the new response becomes: $TRUE$ — ReachedAverage; $FALSE$ — NotReachedAverage. Fitting a Lasso Logistic Model yielded the a new set of variables. I noted that since there is only have one replication (n=1), model diagnostics cannot be performed in this case. The Lasso Logistic Model chose a total of 24 predictors to include with nonzero coefficients. The worse case of the model can be calculated as following:

Let $p_1 = p(\text{observe a success})$, $p_2 = p(\text{predict a success})$,

$$Acc_{worst} = (1+0.25)(p_1 * p_2 + (1-p_1)(1-p_2)) = 62.1\%$$

The prediction accuracy is 62.5%, which is just slightly greater than the worse bound, which means that it is not a very good model. However we can still use it to identify some important factors.

Now we notice that the Unemployment has a positive coefficient while RankingsFood has a negative coefficient, controlling all the other predictors in the model. The positive coefficient for unemployment means the value increase in Unemployment will lead to a decrease of the log odds of ReachedAverage, and the negative coefficient for RankingsFood means the value increase in RankingsFood will lead to a decrease of the log odds of ReachedAverage, controlling all the other predictors in the model. Therefore, the contribution of this two variables is still open to debate.

2.3 PO Cost Model

Using the almost same methods as previously described in section 2, a set of models containing Stepwise Regression Models, Lasso Regression Models, and Lasso Logistic Models were built, with sets of variables selected accordingly.

Up to now, in total, there are 6 set of variables. We can compare them in the next chapter.

3. Findings and Conclusions

3.1 Comparison of two set of models and Discovery

Comparing the above six models, I concluded the effect of each variables in the Table 3.1. '+' means positive effect, '-' means negative effect, '?' means this variable appear in the model several times with different signs, and it is hard to tell its effect based on the coefficient, '/' means this variable doesn't appear in the model. All above are considered the situation of 'controlling all the other variables in the model'.

From the table, we can notice several facts:

a) The linear terms of WhitePop, BlackPop, OtherPop will never appear in the model at the same time, because they are linearly dependent, with a sum of 100. And the same for HealthyPop, ChronicPop, ComplexPop.

b) When focusing on ED Cost, the EDs has positive effect while the remaining 2 Healthcare Utilization Factors – HOs and POs will have negative effect, controlling all other predictors in the model. The same for PO Cost – the remaining 2 Healthcare Utilization Factors will have negative effect. This may indicate the fact that if a census tract has more PO or HO claims within the year, the Medicaid cost on ED PMPM will be less. If a census tract has more ED or HO claims within the year, the Medicaid cost on PO PMPM will be less. These facts may due to the fact that each person has limited time to go to the health care service.

c) The white population percentage has a positive effect on both the cost PMPM, given all the other predictors in the model. This is an indication of that White people tend to utilize this insurance plan more. Also notice that the black population percentage tends to have a negative effect on the cost PMPM, given all the other variables in the model. Therefore, to make the expenditure on different races more balanced, the government should spend more resource to propagate more about this plan, like the benefits, how it works, among the other two groups of people. By doing this, more of them will use the plan more, therefore the expenditure will be more balanced.

d) The Healthy population percentage has a positive effect on both the cost PMPM given all the other variables in the model, which is quite surprising. This may be an indication of the abuse of the plan. To prevent this, the policy of this Insurance Plan may need to be re-made to increase the use of preventative services, which are helpful but cost less, and decrease the unnecessary use of emergency rooms.

e) The Unemployment rate has a positive effect on both of the cost PMPM, given all other predictors in the model. This may lead to more people with limited resources, therefore more people utilize this plan. On the other hand, the unemployed people may have a worse health condition because they may live an irregular life schedule. Also, they may utilize the health care resource more because they are idle at home and need some place to go. Therefore, to reduce the healthcare cost, government should seek to decrease the unemployment rate, which is an action outside the Medicaid Insurance Plan.

f) Income has a positive effect on both the cost PMPM, given all other predictors in the model. It may result from the fact that place where people have more income will be expensive, therefore people need to pay more for

		StepED	LassoED	LogLassoED	OverallED	StepPO	LassoPO	LogLassoPO	OverallPO	Overall
	State									
Healthcare	EDs	+	+	+	+	-	-	-	-	
Utilization	HOs	-	-	-	-	-	-	-	-	
Factors	POs	-	-	-	-	+	+	+	+	
Population	WhitePop	?	+	+	+	?	+	/	+	+
	BlackPop	/	/	/		-	-	-	-	-
	OtherPop	/	/	-		/	+	/		
	HealthyPop	+	/	+	+	?	+	/	+	+
	ChronicPop	+	+	-		-	-	+	-	-
	ComplexPop	/	+	/		/	/	-		
	TotalChild	-	-	-	-	/	-	/	-	-
	TotalAdult	-	-	-	-	-	-	-	-	-
Financial	Unemployment	?	+	+	+	+	+	/	+	+
	Income	+	+	/	+	+	+	+	+	+
	Poverty	/	/	/		/	/	/		
Social	Unemployment	/	/	-		/	-	-	-	-
	Urbanicity	/	/	/		+	?	?		
Health Environment	Accessibility	+	+	+	+	?	-	-	-	
	Availability	+	+	+	+	?	?	/		
	RankingsPCP	+	+	+	+	/	/	/		
	ProvDensity	/	-	+		-	-	-	-	
Life	RankingsFood	+	+	-		+	+	+	+	+
	RankingsHousing	+	+	-		-	-	+		
	RankingsExercise	/	/	-		+	+	/	+	
	RankingsSocial	-	+	+		-	-	-	-	

Table 3.1: The Effect of Each Variables To the Response for Each Model

the same health care service. Hence the Insurance Plan need to cover more too. It is the same reason for the positive effect of RankingsFood.

g) Education has a negative effect on both the cost PMPM in general, given all other predictors in the model. The reason of it may be that more educated a person is, more knowledge to become healthy he may have. Hence it is not a bad idea to have more people with a bachelor's degree.

h) Accessibility and Availability to pediatric primary care both have positive effect on ED Cost PMPM, given all other predictors in the model. High accessibility means high congestion, and high availability means long distance. Hence having both of them high means the children who utilize the emergency rooms must be in a very bad condition, which leads to a higher cost per person. To reduce this problem, government should seek possibility to increase the density of health care resource, which can enable more people to get treated when their health condition is not that bad. This opinion can be further confirmed by the coefficients of ProvDensity, since it seems to have a negative effect on the cost, given all other predictors fixed in the model.

3.2 Guideline to improve the Medicaid Program

From the previous study, it is reasonable to come up with the following suggestions:

a) Spend more resource on informing the people except the white to have them involved into the plan.

b) Re-make the insurance policy to prevent the unnecessary use of emergency rooms and increase the use of preventative services, which are helpful but cost less.

c) Decrease the unemployment rate and have more people with bachelor's degree.

d) Increase the density of health care resource, in order to make more people possible to get medical care when their health condition is not that bad.

3.3 Discussion(Limitation and Issues notices)

During this study a number of issues were observed that will be discussed here.

One issue is that we have always observed that there are outliers or influential points in the models. Even after removing some of them, we still see slight non-conformance to the assumptions of linear regression.

The second issue is that the models seem to be too big. But I don't think over-fitting is an issue here, since after model-selection, the explanatory power of the model doesn't increase much. Moreover, there are over 5,000 data points in the data sets. Therefore, it may still be possible to add more predictors into this model.

Another issue is that I didn't perform group Lasso for model selection. This may lead to an inaccurate model since a subset of categorical variables will be deleted from a certain group, which makes no sense. However, since this study mainly focus on finding the important factors, this issue is ignorable.

3.4 Future Study

One aspect for the future work is about the models. Ridge regression may be worthy to perform to gain a better fit of the model and compare the importance of the factors. Group Lasso and Elastic Net should be performed to improve the result of model selection.

On the other hand, for the Medicaid Program, there are several things worth further study.

The first thing is the difference between people in different colors in utilization of the plan. It is not good if the insurance plan only benefits a certain group of people.

The second thing is the whether more coverage of healthcare service can lead to a lower cost. This may or may not be true. On the one hand, higher the density is, more the Medicaid claims will be because more people will go to get the service. However, people will be treated when their health condition is still not bad, which will reduce the cost per claim. Therefore, this aspect is very interesting to be noticed.

The last thing is about the policy of this plan. Since from the data analysis above, the plan seems to be abused by the labeled 'healthy' enrollees. Maybe some restrictions should be set to prevent the overuse of healthcare resources, especially the emergency rooms.

Thank you for this semester Professor Serban, and wish you all the best. Zhilin Ye