USING LINEAR REGRESSION AND MIXED MODELS TO PREDICT HEALTH CARE COSTS AFTER AN INPATIENT EVENT

by

Christopher W Freyder

BS in Industrial Math and Statistics, West Virginia University, 2014

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Christopher W Freyder

It was defended on

June 1, 2016

and approved by

Jeanine Buchanich, PhD, Research Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Ada O. Youk, PhD, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Thesis Advisor: Eleanor Feingold, PhD, Senior Associate Dean, Professor, Department of Human Genetics, Graduate School of Public Health , University of Pittsburgh

Eleanor Feingold, PhD

USING LINEAR REGRESSION AND MIXED MODELS TO PREDICT HEALTH CARE COSTS AFTER AN INPATIENT EVENT

Christopher W Freyder, MS

University of Pittsburgh, 2016

ABSTRACT

Gateway Health Plan® wanted to compare the before and after costs of a member who had an inpatient stay in a hospital which will allow them to evaluate costs in comparison trials. As part of my internship with Gateway Health Plan®, I was able to estimate a formula to evaluate difference in costs.

Using Gateway Health Plan's® internal data from the past three years, I used regression to evaluate the difference in cost for members before and after an inpatient event. I ran a simple linear regression model as well as a mixed effects model in order to look at the comparison of the before and after costs. Age and gender were also considered at as possible covariates in the prediction process because both of those factors are known to be associated with healthcare costs. The results showed that average cost before an inpatient event as well as gender were significant in estimating the average cost after an inpatient event. I found that females tend to cost less than males, and female patients cost less after the inpatient event compared to before the inpatient event, while men cost more after the event.

Public health significance: This research will help Gateway Health Plan to evaluate interventions to assess whether they lower health care costs. Being able to evaluate if interventions are cost efficient will improve healthcare leading to an improvement in population health.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to thank Dr. Eleanor Feingold for her guidance throughout this project, as well as Dr. Ada Youk and Dr. Jeanine Buchanich for help during this project and guidance throughout my time in graduate school.

I would also like to thank Gateway Health Plan® for allowing me to work with them, especially Fernando Arbelaez. Lastly I want to thank my family for their continued support during my educational career.

1.0 INTRODUCTION

The purpose of this study was to estimate a formula Gateway that Health Plan® could use to compare the average monthly cost of a member before an inpatient event to the average monthly cost after an inpatient event. In the first chapter of this paper, I will discuss Gateway Health Plan and why there is a need for them to have a formula that can be used to assess cost. In the next chapter, I will give a brief introduction of the different types of regression that will be used to formulate this equation. In the third chapter I will summarize the statistical methods that were used to create the formulas. In the fourth chapter, the results of my analysis will be presented. Finally, in the last chapter I will discuss the results and possible limitations of the calculated models, as well as future directions that should be explored.

2.0 GATEWAY HEALTH PLAN®

Gateway Health Plan® is a healthcare company in southwestern Pennsylvania that deals primarily with individuals who are enrolled in Medicaid or Medicare. Medicare and Medicaid are both governmental programs, which work differently than privatized health insurance.

Covering these special populations provides specific obstacles for Gateway. For example Gateway Health Plan® is paid a flat fee for each Medicare and Medicaid participant. [4] This fee is determined by the government based on the comorbidities of each individual in yearly evaluations. Because this is assessed yearly, it is extremely important that individuals enrolled in these programs have at least an annual checkup, because if a member has conditions that go undetected and are not reported the cost Gateway could be large.

Claims are designated as one of four types: inpatient event, outpatient event, emergency department, or pharmacy. Inpatient events are known to be the most expensive type of claim for a member. A patient becomes an “inpatient” when they are formally admitted to a hospital with an order from a doctor. [2] This means that simply spending a day or two at a hospital does not classify you as an inpatient if the doctor has not formally written an order to admit you. Many times, once a patient is admitted as an inpatient, doctors or health care companies will try to implement some intervention to improve that individual’s health. This could be doctor follow-ups, prescribed medication, or some other intervention used to make the individual more self-aware of the daily choices they make that affect their health.

2.1 MEDICAID/MEDICARE

Medicaid was created in 1965 by Congress in order to provide health coverage to low income families and individuals. It is funded jointly by the federal and state governments meaning that every state can have different qualifications for members to qualify for Medicaid. Medicaid serves a vast number of individuals, and in 2014 alone, over 80 million Americans used Medicaid services. [8]

Medicare provides health coverage to individuals that are over the age of 65, persons entitled to Social Security or Railroad Retirement disability benefits for at least 24 months, members with end-stage renal disease, as well as some other specific clientele. Medicare consists of multiple parts commonly referred to as Part A, B, C, and D. Medicare Part A, also known as Hospital Insurance pays for inpatient hospitals stays, skilled nursing facilities, home health agencies and hospice care. Part B, or Supplemental Medical Insurance helps pay for outpatient hospital visits, physicians, as well as other services. Part C, the Medicare Advantage program, allows individuals to have options in the private sector health plans. Part C plans include HMOs and PPOs, which must cover everything parts A and B do, but may choose to charge different copays. They can also restrict which doctors a member can see in network under this part of Medicare. Lastly part D helps pay for prescription drugs not covered by the other parts. [4, 12] Gateway provides part C and part D Medicare to qualifying members.

2.2 VENDORS

Because Gateway is paid a flat fee and cannot change their fee for each patient, it is important that they lower costs as much as possible. While private insurance can raise premiums if costs are higher than revenue, Gateway cannot. Therefore, Gateway investigates interventions from different vendors to assess possible ways to lower per member per month costs. However, because vendors are trying to sell their product, it is not uncommon for a vendor to display data in a way that makes their product look better. Because of this, Gateway tries to validate the results a vendor presents with their own data.

One complexity in these types of analyses is that it is believed that individuals have different costs associated with leading up to an inpatient event and after an inpatient event. Vendors that come to Gateway will not take this difference into consideration and will just report raw numbers.

2.3 POSSIBLE PREDICTORS

According to the Centers for Medicare and Medicaid services, in 2010 overall spending for women was 29% higher than for men, but when looked at on a per enrollee basis, men spent 54% more than females. [11] It has also been shown that aging increases the cost of healthcare. [6] Along with gender and aging, a discrepancy in health and healthcare costs has been shown to be associated with race. [1] Therefore race, age, and gender were considered as possible covariates in my analysis.

3.0 REGRESSION

For this analysis I will use two different types of regression to come up with different possible equations that can be used for the estimation of cost after an inpatient event. The following is a brief description of the two methods, linear regression and mixed model regression

3.1 LINEAR REGRESSION

Linear regression analysis is used to investigate the relationship between a dependent variable and an independent variable. This is done by obtaining data on variables that are believed to affect the relationships the investigator is interested in and estimating the equation:

$\hat{Y}_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + \epsilon_t$

In this formula, β_0 represents the intercept of the regression equation, β_1 to β_k represent the slopes associated with each variable, and X_{it} represents the data value of each variable i for each subject t. Lastly, ϵ_t is the error term associated with individuals t. The point of this regression is to find the best equation that uses the X variables to predict the Y variable. In order to use linear regression the following things are assumed [7]:

1. The effects of the predictor variables, x_i, on the outcome variable, y_i, are linear and additive
2. The predictor variables are independent of each other

3. There is homoscedasticity of the errors
4. The errors are normally distributed

In certain situations data measurements, will be correlated. This occurs when subjects are grouped or there are repeated measures on a subject. In these situations, it is important to account for this correlation in the model because independence is an assumption of this model. When this type of correlation occurs, a mixed model can be used instead .

3.2 MIXED MODELS

A mixed model is a model which allows for both fixed and random effects. A fixed effect is an effect that comes from a variable where the only levels in question are levels that are of interest. For example, gender and age are fixed effects because our sample will reflect all gender and ages that are of interest. A random effect is an effect that comes about from a random sample of variables from a larger set of possible choices for that variable. There are different ways an effect can be considered random. One example is if data are collected from five hospitals, but the results are to be generalized to all hospitals, a random effect for the hospital will be added. Another example can be if there are repeated measures on patients. There will be correlation between the measurements based on the patient, so a random patient effect can be added to account for the correlation. [3, 5] These types of models are useful in healthcare analysis when a randomized control study is not an option, and “patient as their own control” methodology is used instead.

The formula for a mixed model for subject j at time i is given as follows:

- Y_{ij} = response of subject i at the j^{th} time measurement where i = 1, 2, …, m and j= 1, 2, … n_i
- n_i = number of measurements for subject i
- m = number of subjects
- β = matrix of fixed effect parameter
- x_{ij} = covariate vector at the j^{th} measurement for the i^{th} subject for fixed effects $\beta \in \mathbb{R}^p$
- γ = matrix of random effect parameter
- z_{ij} = covariate vector at the j^{th} measurement of the i^{th} subject for random effects $\gamma \in \mathbb{R}^q$
- Final model $\mathbf{Y_{ij} = X_{ij}\beta + Z_{ij}\gamma + \epsilon_{ij}}$

I will also assume that $\boldsymbol{\gamma \sim N_q (0, G)}$ where **G** is the covariance matrix of the random effects. Also $\epsilon_{ij} \sim N_{n_i}(0, \mathbf{R_i})$ where $\mathbf{R_i}$ = covariance matrix of error vector in cluster i. This leads to the variance of **Y** being **V= ZGZ' + R** which can be modeled by specifying a covariance structure. [3, 5] This model can be generalized to handle different types of data.

When using a random variable in a mixed model, there are many different covariance structures that can be assigned for different situations. The most commonly used structure, and the default in SAS, is the variance components structure. In this structure, the variances for each parameter estimate are allowed to differ, and the covariance estimates are 0. The other commonly used method is the unstructured covariance structure. In unstructured, each variance and covariance has a different estimate. The makeup of these two covariance structures can be seen in figure 1. In my analysis the variance component structure was used.

$$Varaince\ Component: \begin{bmatrix} {\sigma_A}^2 & 0 & 0 & 0 \\ 0 & {\sigma_B}^2 & 0 & 0 \\ 0 & 0 & {\sigma_C}^2 & 0 \\ 0 & 0 & 0 & {\sigma_D}^2 \end{bmatrix} \qquad Unstructured: \begin{bmatrix} {\sigma_1}^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & {\sigma_2}^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & {\sigma_3}^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & {\sigma_4}^2 \end{bmatrix}$$

Figure 1: Common covariance structures used in mixed models

4.0 METHODS

The purpose of this study was to estimate a formula that Gateway Health Plan® could use to compare the average monthly cost in the six months before an inpatient event to the average monthly cost six months after an inpatient event. This measure allows for a two month grace period after the inpatient event, which was excluded from the calculations for the individual to regain normal health. Therefore the six months used to calculate the average monthly cost after the inpatient stay started 60 days after the last day of the inpatient stay. Gateway was interested in a formula that could be used so when a new member comes to Gateway after an inpatient event, they could estimate the cost based on the patient costs before the inpatient event. Both a linear regression and a mixed model regression are reported and compared. For this analysis only the data for Medicaid in PA was used because members in different states and who have different coverage behave differently

Gateway's initial attempt at solving this problem was to divide the average monthly cost after an inpatient event by the average monthly cost before that event. While this gave them a ratio for before and after costs, it did not take into account differences related to other measureable factors, i.e. covariates. Therefore, we decided that a model was needed to achieve a better estimate of the cost difference. Initially I ran a linear regression. Next, I log transformed the outcome variable because it was heavily skewed. There were almost 3000 patients with no claims in the outcome period. Because these values were undefined when the log transformation

was applied, I added $1 to all zero amounts before applying the log transformation. Because the results were very similar to the linear regression, and because Gateway wanted a “simple” model, we decided not to report the model with the log transformation. Lastly because the data were “patient as their own control” measurements, I decided to run a mixed model as well to allow the correlation between measurements to be handled.

4.1 DATA

Gateway data is stored in SQL databases. In order to extract the data, first I created a table in SQL that contained anyone that had an inpatient event since January 1st, 2013. Then, I deleted anyone who was not a member for at least 6 months before and 8 months after the date of their inpatient event. These individuals had to be removed because if they did not have data for those 14 months, they would not have enough data to contribute to the study. Then I used the member numbers of these individuals to pull any claims they had from 6 months before to 8 months after the date of their inpatient stay. I also pulled age, race, and gender for these individuals. At this point I moved the data to SAS using proc SQL to carry out further data manipulation.

Once the data were in SAS, anyone with more than one inpatient event was dropped, and I calculated the total costs in the six months before and the six months after a two month grace period. Individuals with more than one inpatient event were dropped because inpatient events are the most expensive type, and the members with more than one significantly change the results. The final dataset contained the variables member ID, average monthly cost before, average monthly cost after, gender, race, and age. For the mixed model portion I had to manipulate the

data set so that the before and after costs were in one column with a time and patient identifier for each measurement.

After talking to an expert on these types of events, patients who were under the age of 18, over the age of 65, or who had inpatient events due to pregnancy were then removed from the dataset. Patients under the age of 18 were removed because children are known to behave differently than adults when it comes to insurance costs. Patients over the age of 65 were removed because once a member turns 65 they become eligible for Medicare, and their Medicaid benefits change. Patients whose inpatient stay was due to pregnancy were removed because pregnancy is different than other inpatient stays. In the six months before giving birth an individual will likely have many appointments and checkups, whereas for the time after a pregnancy, most checkups are filed under the child's insurance. This would likely be a different behavior than most other inpatient events.

4.2 ANALYSIS PLAN

Once the dataset was finalized, I first ran descriptive statistics on the variables that were used to create the models. I calculated the medians and interquartile range of average cost before, average cost after, and age, as well as how they varied between the genders. I also calculated the count of males and females and races in the data. I plotted a histogram of the ages in order to look at the age distribution of the data. Reason for hospitalization was also looked at but most code numbers had less than 5 observations, so it was omitted from the regressions.

After the descriptive statistics were analyzed, I used SAS to create linear regression models. Average monthly cost after the inpatient event was the outcome variable, and I used

combinations of age, gender, race, and average monthly cost before as predictors. Because the main purpose of this analysis was to see how cost before affected cost after, the average monthly cost before variable was forced into the model. I used backwards selection to create a model of the significant predictors. Then all combinations of models were run with those predictors, and the models were compared using AIC and R^2 criteria. For AIC, lower values are better, and values within 2 of each other show there is no difference between the models. For R^2 the model with the higher value is considered to be a better fit, however, as covariates are added the R^2 will go up which must be taken into consideration when comparing models. The best fit regression line, based on these criteria, can then be used as the formula for predicting average monthly costs of a member after an inpatient event. I then calculated residuals to asses model fit on the final chosen model.

Next I reshaped the data so that I could run a mixed model. This allowed me to have two values of cost for each individual, one associated with before the IP, one after the IP. For this model, I used cost as the outcome; time, gender, race, and age were the fixed effects, and patient subject number was used as the random effect to allow for a random slope in the model and to allow the correlation within patient measurements to be addressed. A random intercept was also used in this model to allow for difference in patients baseline cost. For the mixed model I used backwards elimination on the fixed effects to get to a final model. I chose to use variance components as the correlation structure for this analysis. Studentized conditional residuals were calculated and plotted in order to asses model fit of the final model.

After running these calculations, I decided to run a second set of models that would take out individuals who had extreme costs in one time period but not the other. I decided to look at a histogram of the difference in costs and to pick cutoffs that contained most of the data. This

method was chosen because it would eliminate some special cases that are not like the average healthcare member. For example, some members might not go see a doctor until they have an inpatient event, meaning even if they were sick and should have sought out care, their cost before the event would be $0 and if the patient needs regular medical attention because of the event it could result in high discrepancy of costs. Also this would eliminate individuals who had some major cost before or after the event that is an unusual occurrence that greatly skewed their cost for one period. After I used this method to delete members who I considered “outliers” the same methods stated above were repeated to come up with an addition linear regression model and an additional mixed model.

5.0 RESULTS

5.1 DESCRIPTIVE STATISTICS

Initially, 21811 patients were included. After I finalized the dataset by removing individuals who did not fit the inclusion criteria, I was left with 17320 patients aged 18 to 65, who had one inpatient event in the last three years not related to pregnancy, and were also a member of Gateway Health Plan® for six month before and eight months after their inpatient event. Then after looking at the difference between the before and after costs, I decided that a cutoff of ± $30,000 would be used to get rid of outliers and create the second dataset, of 17214 members. The median and IQR of average monthly cost before, average monthly cost after, and age for all patients and stratified by gender are displayed in Table 1 for the whole dataset and Table 2 for the subset of data where outliers were taken out of the sample.

Table 1: Descriptive Statisitcs for all patients

Variable	Median (IQR)		
	Total n=17320	Male n=3938	Female n=13382
Average Monthly Cost Before ($)	277.67 (396.70)	197.91 (491.85)	291.53 (366.74)
Average Monthly Cost After ($)	104.96 (329.74)	189.35 (541.05)	87.80 (275.67)
Age (Years)	35.0 (25.5)	35.8 (25.6)	34.9 (25.5)

Table 2: Descriptive Statistics for data excluding outliers

Variable	Median (IQR)		
	Total n=17214	Male n=3887	Female n=13327
Average Monthly Cost Before ($)	276.33 (392.17)	193.60 (478.51)	290.90 (363.38)
Average Monthly Cost After ($)	103.53 (323.70)	183.29 (521.44)	87.18 (273.628)
Age (Years)	35.0 (25.5)	35.8 (25.7)	34.8 (25.5)

As seen in Table 1 and Table 2, there is a difference in costs between males and females, and also a difference in before/after costs. The median average monthly cost for males only goes down about $10 a month (if the mean is looked at the cost actually goes up) but for females the cost goes down over $200 per month. The larger IQR in the male categories can be explained by the small sample size for men compared to women in this data. The cost is also much higher in the female before cost but much lower in the female after cost compared too males. These results led to inclusion of an interaction term between gender and time in the mixed model building process.

Women account for about 78% of the individuals in this study. Women are known to be more likely to qualify for Medicaid; however, this is still an extremely large difference. 51 men and 55 females were dropped when the outliers were removed. Because there were so many more females than males in this study, proportionally more men were removed.

In figure 2, the distribution of ages can be seen. There is a bimodal distribution: a spike in individuals between the ages of 20-30 as well as a smaller spike with individuals between the ages of 49-55.

Table 3 shows the number of members identifying with each race classification. Over half of the individuals identify as white, with another quarter identifying as African American. The remaining individuals are made up of Native Americans, Asians, other, or did not answer.

Figures 3 and 4 show scatterplots of the cost after an inpatient event vs the cost before an inpatient event for the full data and the data with the outliers removed. From these scatterplots we can see that the data that doesn't fit the idea of the before cost predicting the after cost is removed.

Reason for hospitalization was also looked at for these individuals; however most of the codes had between 1 and 5 members in that code, and therefore was not used because the sample size per code was too small.

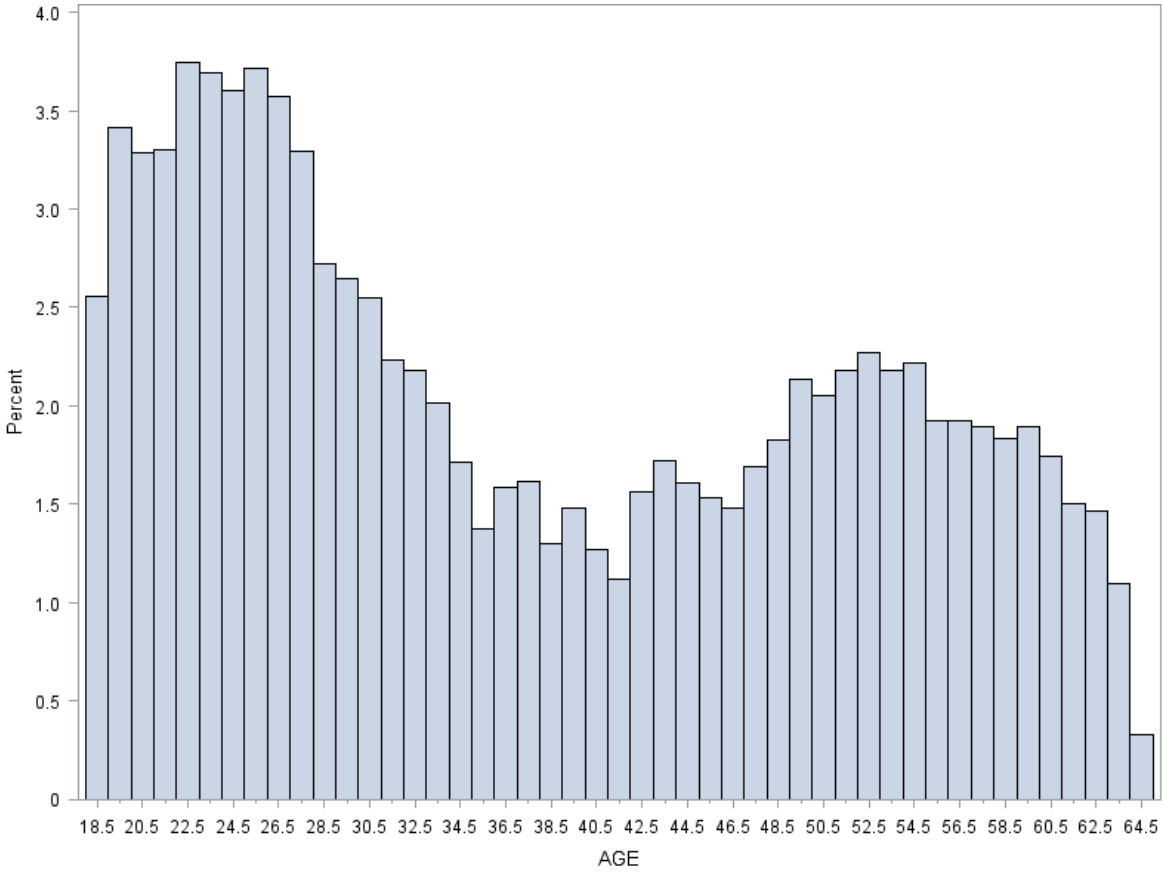

Figure 2: Distribution of age

Table 3: Self-Identified Race Counts

Race	Frequency	Percentage (%)
Native American	48	0.3
African American	4087	23.6
Asian	206	1.2
White	10870	62.8
Other	1826	10.5
Unknown	284	1.6

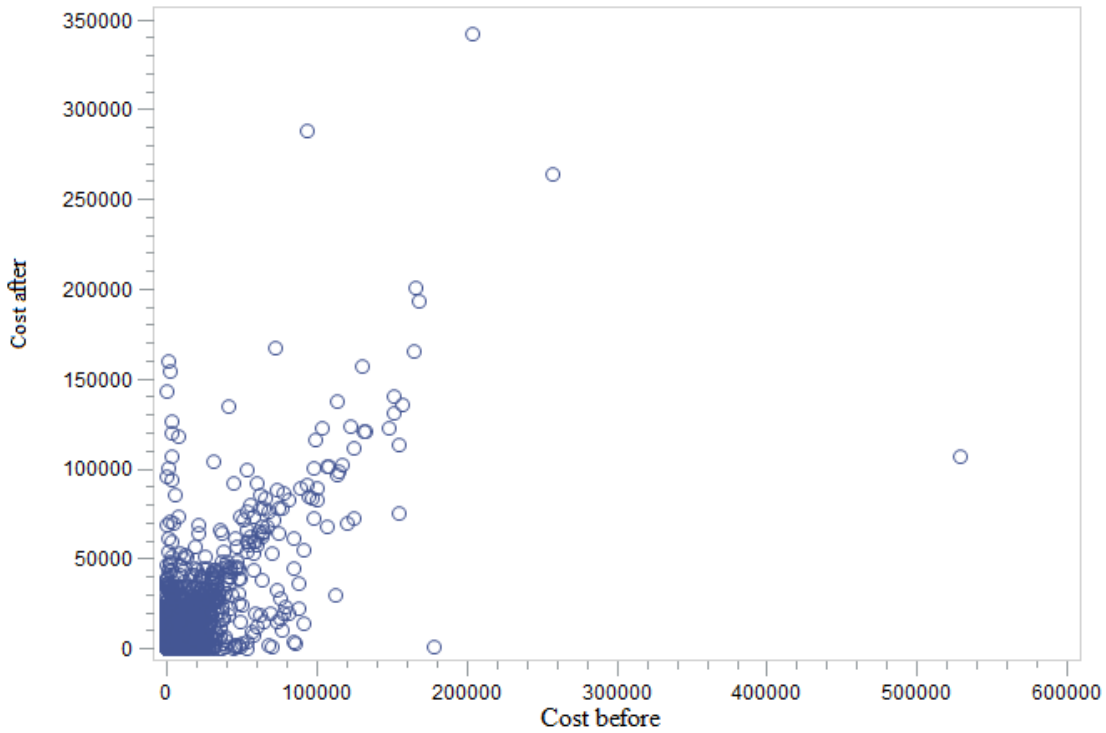

Figure 3: Scatterplot of after cost vs before cost for full data

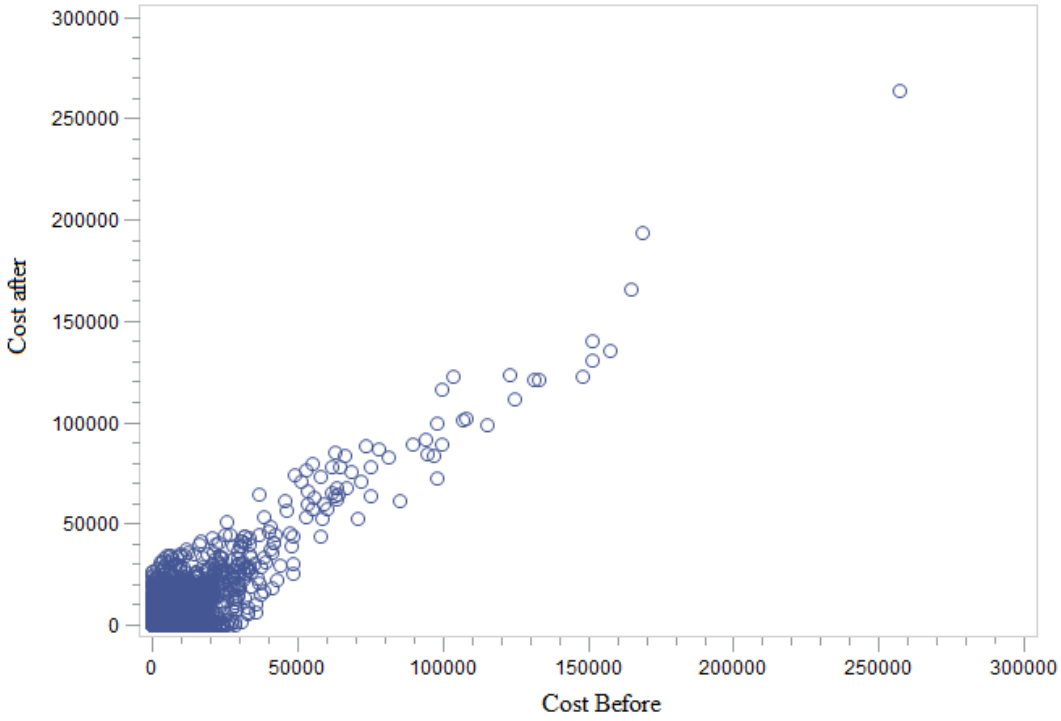

Figure 4: Scatterplot of after cost vs before cost for data without outliers

5.2 LINEAR REGRESSION

The outcome variable for the linear regression was average monthly cost after the inpatient event. Possible predictors were in this model included gender, age, race, and average monthly cost before the inpatient event. First I ran linear regression model with all of the variables, and race was seen to be not statistically significant (p-value =.3517) so it was removed. The, because age and gender were significant, models with just age, just gender, and both were run, and AIC and R^2 values were calculated to see which model was the best fit. As seen by table 4, model 2 and model 4 have AIC's that are less than 2 apart as well as the same R^2 value. These models also have the lowest AIC's and the highest R^2, so I concluded that these models are the best two models, and there is no evidence that one model is better than the other. The model with just cost before and gender was chosen since it will be a simpler model than the full model, yet just as effective. After the final model was selected, studentized residuals were calculated, plotted and can be seen in figures 5 and 6. In figure 5, it is seen that the homoscedasticity assumption is violated. In figure 6 it is seen that the normality assumption is violated. The final model looks as follows;

Average Cost After = 254.41+ 0.703 * Average Cost Before – 252.26 * Female

Table 4: Linear regression model comparisons for full dataset

Model	Covariates	AIC	R^2
1	Cost Before	240640.79	.507
2	Cost Before + Female	240463.19	.513
3	Cost Before + Age	240642.78	.507
4	Cost Before + Age + Female	240465.11	.513

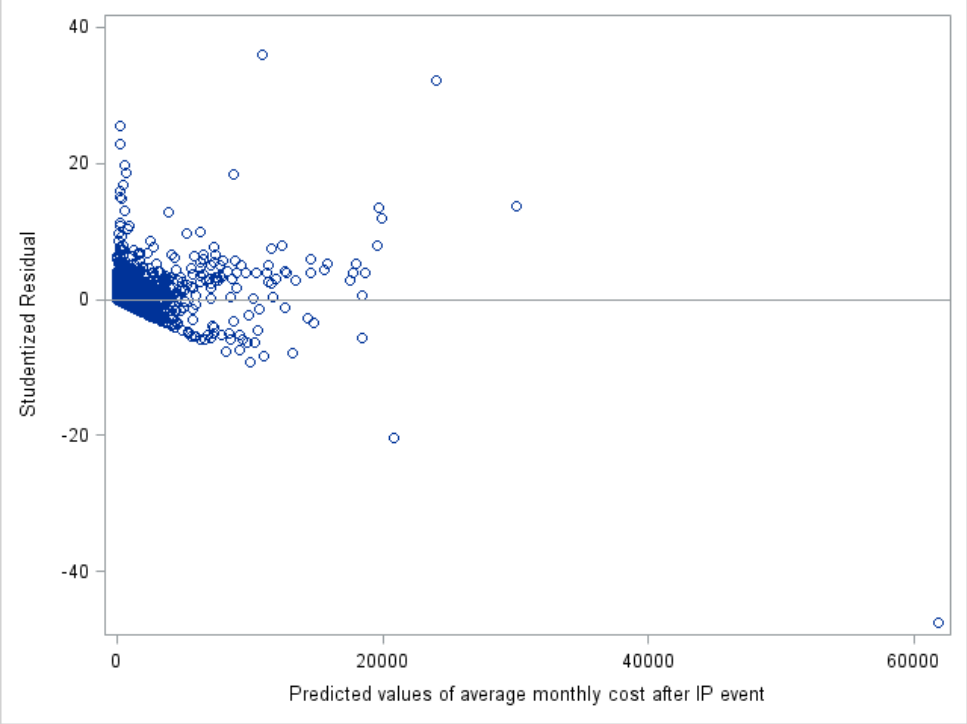

Figure 5: Residual vs fitted for model 2

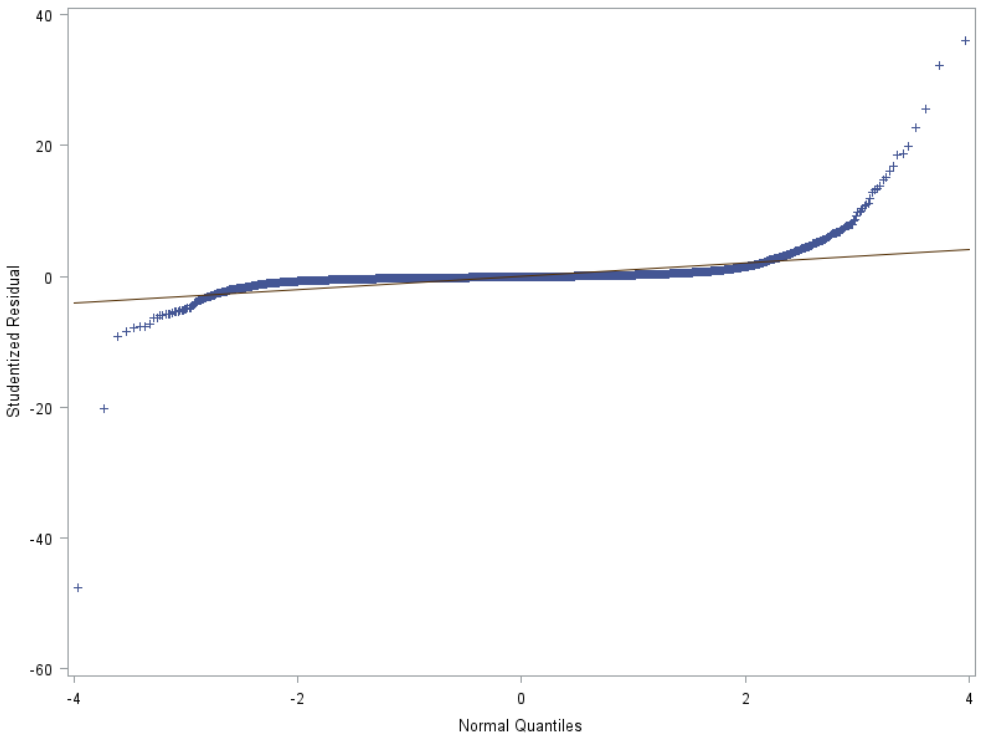

Figure 6: QQ plot for model 2

Next I wanted to look at a model excluding any patients who had an extreme cost because this would be less likely to represent the general population. For this the second dataset was used. Once again the models with average monthly cost before and female, along with the full model were the best models based on AIC and R^2 values. I chose model 6 with average monthly cost before and female again due to the simplicity and so that I have the same model to compare between the datasets. Residuals were run on model six to look at the model fit and can be seen in figures 7 and 8 and it can be seen that the assumptions of homoscedasticity and normal errors are violated again. I got the final model that looks as follows;

Average Cost After = 79.80+ 0.901 * Average Cost Before – 189.68 * Female

Table 5: Linear regression model comparisons for dataset without outliers

Model	Covariates	AIC	R^2
5	Cost Before	278655.10	.771
6	Cost Before + Female	278290.254	.776
7	Cost Before + Age	278654.81	.771
8	Cost Before + Female+ Age	278290.62	.776

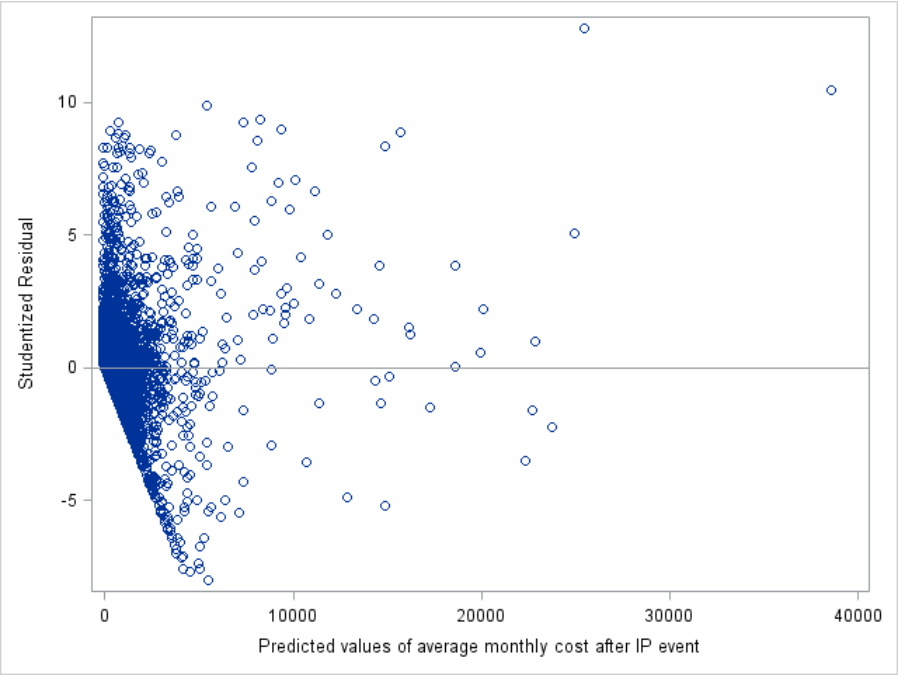

Figure 7: Residuals vs fitted for model 6

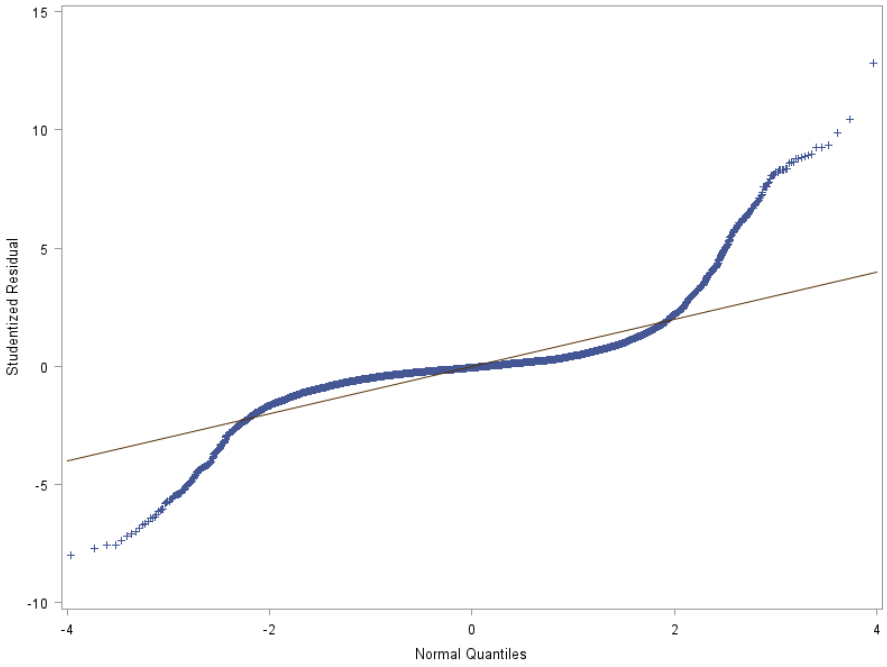

Figure 8: QQ plot for model 6

5.3 MIXED MODELS

For the mixed models, I used cost as the outcome variable. Time, gender, race and age were used as possible fixed effects, and subject number was used as the random effect. I also used a random intercept in the model to allow for different baseline costs for the members.

I used backwards elimination to come up with a final mixed model. Starting with the full model including age, gender, race, and time, the least statistically significant variable, age (p =.576) was removed and the model was rerun. The next least statistically significant variable, race (p=.317), was removed and the model was rerun. Removing age and race did not change the other coefficients in the model much, and the parameter estimate of $0.50 per month for age and around $20 per month for each race were so little compared to the other variables estimates, it

was decided that age and race could safely be removed. I then added a time*gender interaction because the descriptive statistics showed that there was a large difference in the genders. The interaction term was statistically significant so it was kept in the model. The results from this model can be seen in table 6. Marginalized residuals for the model were also calculated and can be seen in figure 9. From these residuals it can be see that the homoscedasticity assumption as well as the normality of errors assumption are both violated.

From this model I can get our estimates for the price difference of average monthly cost before and average monthly cost after for each gender. For males the average monthly cost in the six months before an inpatient event was $678.19, and the cost per month for the six months after the inpatient event was $722.02. This shows that men cost an average of $43.83 per month more after an inpatient event as before the inpatient event. For females the average monthly cost before the inpatient event was $499.63 while the cost after the inpatient event was 344.48. This yields that after an inpatient event females cost an average of $155.15 less per month than they cost before the event.

Table 6: Coefficients and p-values of variables in full data mixed model

Variable	Coefficient	P-value
Time	43.83	0.0146
Female	-178.56	<.0001
Time*Female	-198.98	<.0001
Intercept	678.19	<.0001

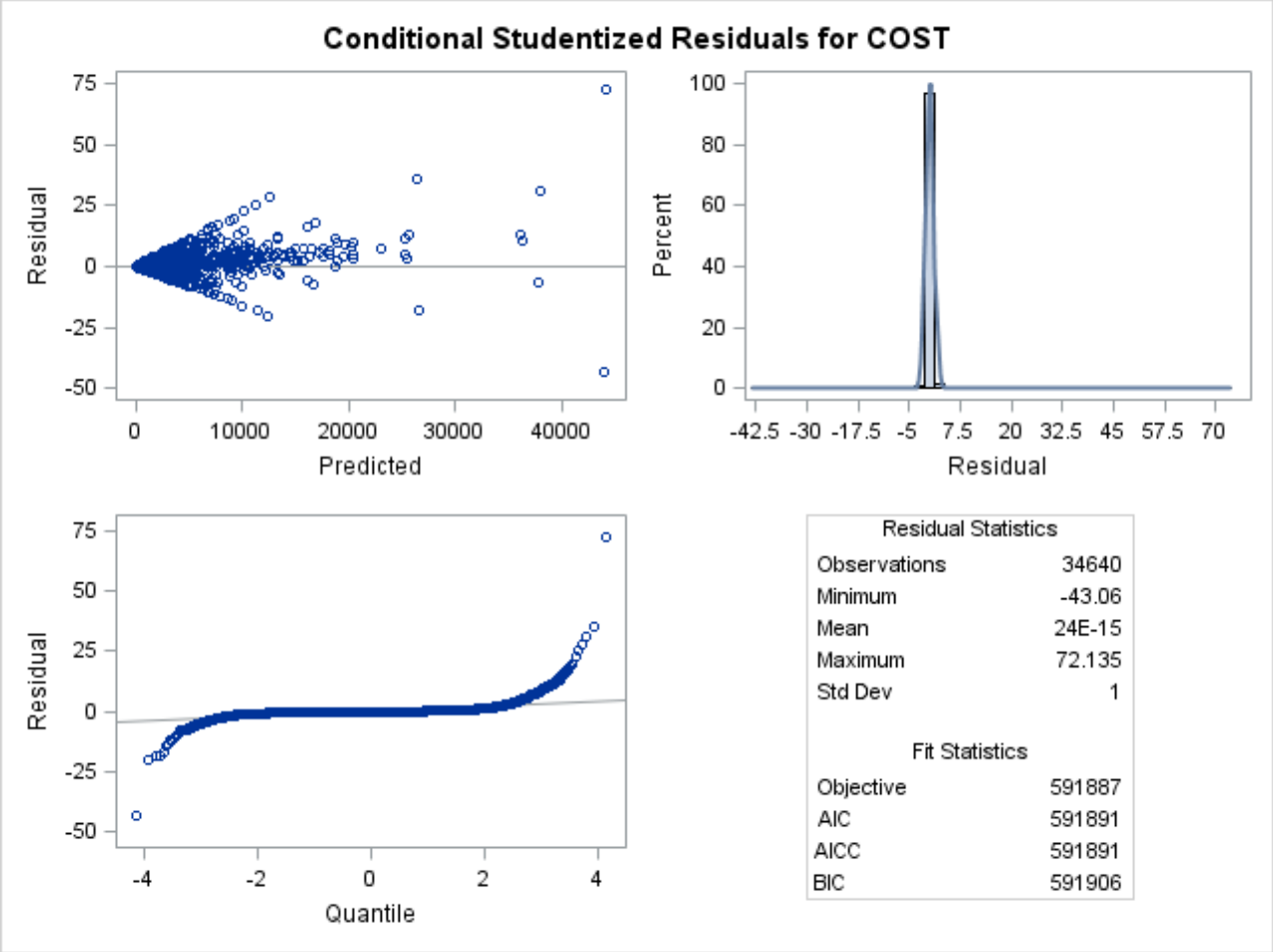

Figure 9: Conditional residuals for mixed model on full data

The same outliers were removed for the mixed model as the linear regression. The model was rerun and the results can be seen in table 7, as well as the marginalized residuals in figure 10. The homoscedasticity assumption and the normality of residuals assumption look to be violated. From these results it can be seen that males cost on average $586.69 per month before an inpatient event and $609.90 per month after an inpatient event. This means males cost on average $23.21 more after an inpatient event than before that same event. Females cost on average $470.82 per month before an inpatient event, and $314.49 per month after the inpatient

stay. Therefore, women cost on average $156.33 less per month after the inpatient event compared to before the event.

Table 7: Coefficients and p-values of variables in the subset of data mixed model

Variable	Coefficient	P-value
Time	23.21	0.0086
Female	-115.87	<.0001
Time*Female	-179.54	<.0001
Intercept	586.69	<.0001

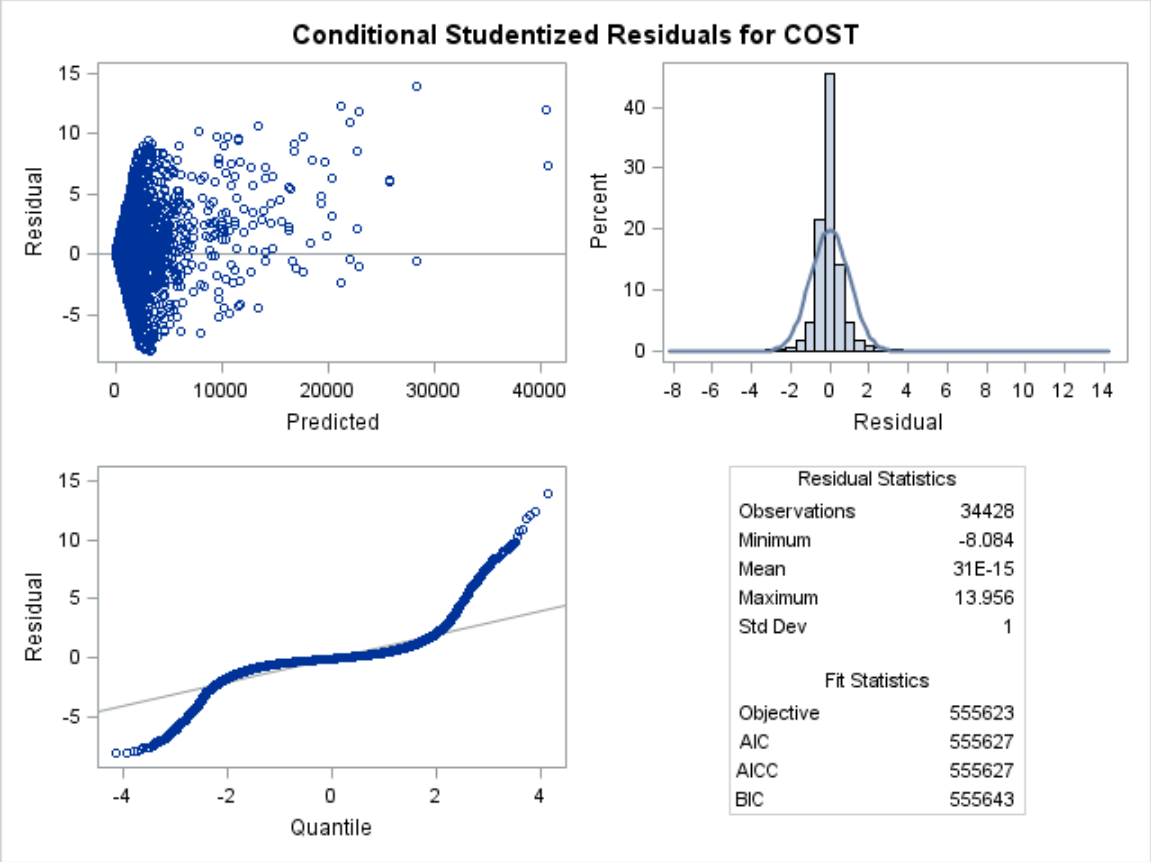

Figure 10: Conditional residuals for mixed model without outliers

6.0 DISCUSSION

It can be seen from the linear regression equations that after deleting 106 of the extreme observations, there is a substantial change in the coefficients. This change in coefficients greatly affects how much gateway can expect to pay in certain situations. The comparison of coefficients can be seen in table 8.

Table 9 shows the different payments Gateway can expect for patients with different costs before the inpatient. Looking at the different starting costs, the comparison between the end cost changes substantially between the models. Model 2 has higher expected costs for lower initial costs, however model 6 has higher expected costs when the initial cost is higher. This can be explained looking at the comparison of coefficients in table 8.

Because the members who were deleted were individuals with high costs in one time period can explain the change in models. The members with the high cost after the event would be outliers in the Y, and when removed would shift the whole regression line up, explaining the larger intercept in model 2. The members with high cost before would be outliers in the X, and would flatten the line out explaining the lower coefficient associated with cost before covariate when removed.

Table 8: Comparison of coefficients for linear regression models

	Model 2		Model 6	
Covariate	Coefficient	p-value	Coefficient	p-value
Intercept	254.41	<.001	70.80	<.001
Female	-252.26	<.001	-189.68	<.001
Cost Before	.703	<.001	.901	<.001

Table 9: Comparisons of average monthly costs after an inpatient event between models

Gender	Male		Female	
Average cost before	Model 2 (full)	Model 6 (subset)	Model 2 (full)	Model 6 (subset)
$500	$605.91	$530.30	$353.65	$340.62
$1000	$957.41	$971.80	$705.15	$782.12
$5000	$3769.41	$4584.80	$3517.15	$4395.12

Table 10: Comparison of coefficients between mixed models

Coefficient	Model on full data	Model on subset of data
Time	43.83	23.21
Female	-178.56	-115.87
Time*Female	-198.98	-179.54
Intercept	679.19	586.69

When looking at the mixed models the fixed effect coefficients, as seen in table 10, the estimated parameters sizes are larger in the model on the full data than in the model on the data without the outliers. This is easily explained by the fact that member with extremely high costs were deleted. Deleting the high costs led to the model shifting down.

In the mixed models it can be seen that the time variable and the time female interaction are significant for both models. This is important because it tells us that there is a difference in before and after costs. However, looking at the parameter estimate, the interaction has a much larger estimate than the time variable. This was expected when looking at the summary statistics and the difference in the costs related to the genders. Because men did not change cost much, the time variable does not have a large coefficient, but the interaction term have large coefficients because females cost changes more between the different time points.

The results in this paper show that the cost before an inpatient event can be used as a significant predictor of cost after an inpatient event. It also agrees with previous literature by showing that males cost more individually than females, however females make up more of the overall costs due to the high percentage of members being females. [11]

I would advise Gateway Health Plan® to use the mixed model on the subset of the data. This model allows them to take into account correlation involved in a “patient as their own control” study. I believe the subset of data are more appropriate to use since very few individuals (0.6%) have extreme costs in one time period but not in the other.

6.1 LIMITATIONS

One possible problem with using these models is the range of values the outcome variable can have depending on the inputs. In model 2 the intercept is larger than the estimated female parameter; therefore there would never be a negative prediction for monthly cost after. For model 6 however, it can be seen that the estimated female parameter is much larger than the intercept, making it possible to obtain a negative value. Solving the equation for total after cost of $0, any female with an average monthly cost before that is less than $121.95 will have an expected monthly cost after of less than $0. Because the goal was to use this model to predict an after cost, one option would be to assume that patients with a before cost less than this will just have an average monthly cost after the inpatient event of $0, however this is not practical and a solution to this problem should be found.

Another problem with these models can be seen when looking at the residual plots. In both models it can be seen that there is a violation of the homoscedasticity assumption. There is a pattern of the magnitude of the maximum negative residuals being proportional to the predicted value. However, once the values get larger, the variances seem to be more equal. In the QQ plots, the residuals look to have a heavy tailed distribution rather than a normal distribution. This means that the normality of residuals assumption is also violated.

Looking at the residuals for the mixed models, the conditional residuals will be used. These residuals are the difference between the observed and fitted values. These residuals take into account the known information in the random variable. [9] From the residual panels, it can be seen in the model run on the full dataset, that there is a clear fanning pattern throughout predicted values, violating the heteroscedasticity assumption. The QQ plot for this model shows that the residuals are once again heavy tailed.

In the model run on subset of the data, like with the linear regression, the residuals have a fanning pattern at the beginning but even off as the predicted values get higher. The QQ plot associated with this model shows that the normality of the residuals are violated. With all of the bad diagnostics, none of these models seem to be a good fit. Therefore when using any of these models, results should only be provisionary.

6.2 FUTURE DIRECTIONS

The next step with the models developed in this paper would be to use other available data to validate the models. Gateway has data from other states that could be used to validate the models I developed. Gateway could also look at incoming patients and track their costs in the future to see how the model works at predicting in the intended situation.

While this model can be used to predict costs of patients who had an inpatient event, the lack of model fit is concerning. There are many possibilities that could contribute to finding a better model. First, other regression methods should be considered in order to address the specific difficulties that these models have in fitting the given data. Possible solutions would be polynomial models, tobit models, or other models that are nonparametric.

Another solution could be transforming the data. Because the nature of healthcare costs are skewed, transformations can be run on the cost variable to see if a better model fit can be obtained. Also, other covariates could be considered. The difference in gender may be due to some other confounding variable. I would like to investigate the reason for hospitalization in order to see if the genders are equally represented in each code. I believe that different reasons for hospitalization could explain the cost discrepancy between genders. Because there are so

many hospitalization codes that have few observations, they would need to be grouped by similar codes to run this analysis.

Along with reason for hospitalization, I would like to look into the comorbidity loads of the individuals in this study. This could be another confounder for the interaction term. The number of comorbidities an individual can change, so I could look at the average comorbidities before and after the inpatient event by gender. Even if this term doesn't contribute to the interaction, it could be statistically significant in predicting costs.

Similarly to the problem seen with reason for hospitalization, I saw that race had most individuals grouped into White or African American and not many individuals in the other four categories. I would like to collapse the race categories to White, African American, or other and rerun the analysis to see if anything would change.

Another possibility to consider while moving forward is that we want to consider the average healthcare member for these predictions. Since most individuals have relatively low cost, I could also fit a model on just members with costs less than a certain amount. This would allow us to have another model to be used on individuals who come into Gateway with a low cost before an inpatient event.

There are still a lot of possibilities to be considered for this type of analysis. While this paper explored some solutions to cost analysis in healthcare, I believe this problem needs to be analyzed more in order to come up with a better way to estimate health care costs for members after they have an inpatient event.

6.3 PUBLIC HEALTH SIGNIFICANCE

This research will help Gateway Health Plan® to evaluate research interventions to assess whether they lower health care costs. Being able to evaluate if interventions are cost efficient will improve healthcare leading to an improvement in population health.

BIBLIOGRAPHY

1. Ayanian, John Z. "The Costs of Racial Disparities in Health Care." *Harvard Business Review*. N.p., 01 Oct. 2015. Web. 25 May 2016.

2. "Are you a Hospital Inpatient or Outpatient?" *The Official U.S. Government Site for Medicare.* Medicare.gov, n.d. Web. 25 Apr. 2016.

3. Brown H and Prescott. Applied Mixed Models in Medicine, 2nd edition. Chichester, England: John Wiley, 2006. Print

4. "How do Medicare Advantage Plans Work?" *The Official U.S. Government Site for Medicare.* Medicare.gov, n.d. Web. 21 May 2016.

5. Littell, Ramon C., George A. Milliken, Walter W. Stroup, Russell D. Wolfinger, and Oliver Schabenberger. 2006. SAS® for Mixed Models, Second Edition. Cary, NC: SAS Institute Inc.

6. Mendelson, D. N., and W. B. Schwarts "The Effects of Aging and Population Growth on Health Care Costs." *Health Affairs* 12.1 (1993): 119-25.

7. Osborne, Jason. "Four Assumptions of Multiple Regression That Researchers Should Always Test." Practical Assessment, Research & Evaluation, 2002. Web. 14 Apr. 2016.

8. "Policy Basics: Introduction to Medicaid." *Center on Budget and Policy Priorities.* N.p., 19 June 2015. Web. 13 Mar. 2016.

9. Schabenberger, O. (2004) Mixed Model Influence Diagnostics, SUGI 29 – Statistics and Data Analysis, Paper 189- 29.

10. SAS Institute Inc., SAS 9.1.3 Help and Documentation, Cary, NC: SAS Institute Inc., 2002-2004.

11. "U.S. Personal Health Care Spending by Age and Gender." *Centers for Medicare and Medicaid Services* N.p., 2010. Web. 25 May 2016.

12. "What is Medicare?" *The Official U.S. Government Site for Medicare.* Medicare.gov, n.d. Web. 18 Apr. 2016.