WPI

Analyzing Reimbursement Ratios of Medicare Providers

A Study Using Generalized Linear Models

A Major Qualifying Project, submitted to the faculty of
Worcester Polytechnic Institute in partial fulfillment of the requirements for the
Degree of Bachelor of Science

Submitted by:

Marilda Bozdo

Blaine Bursey

Alberto Romo Herrera Ibarrola

Meghana Prakash

Submitted to:

Project Advisors:
Prof. Jon P. Abraham
Prof. Barry J. Posterro

Abstract

Generalized linear models are starting to gain popularity among actuaries in most countries for target marketing analysis. In order to better understand how these models work, a project was commissioned regarding medical providers and their reimbursement ratios. By using one-way analysis, several factors were selected to model the response variable and the factors' significance was determined by using an algorithm in the statistical software, SAS. Several general linear models were set up and tested to fit the reimbursement ratio. By calculating and analyzing each correlation, we were able to find a model that matched the reimbursement ratio with an 83 percent correlation.

Authorship

This project was completed through a combined effort from all group members. Each individual contributed equally throughout the project. Tasks, including researching methods, executing procedures, analyzing data, and writing the report were divided amongst the members.

Acknowledgements

Our MQP group would like to acknowledge and thank different individuals in the completion of our project.

We would like to acknowledge the following:

- Professor Abraham and Professor Posterro, our advisors, for their timely feedback as well as their input on our project. They constantly pushed us to do our best and helped us when we faced challenges and their expertise was very valuable in completing our project.
- The CMS Medicare Provider Data Team, part of the CMS Manual System Department of Health & Human Services, for their information on different variables used in our project.
- We would also like extend our thanks to Worcester Polytechnic Institute (WPI) for making this experience possible.

Table of Contents

Table of Figures

Table of Tables

Executive Summary

Generalized linear models (GLMs) are an extension of the linear modeling process that allows models to be fit to data. They have been in use for over thirty years but it is only recently that the level of interest and the rates of adoption have increased substantially. Our team looked into the GLM procedure for our project in order to gain a working knowledge of this method.

The goal of this project was to create a generalized linear model to determine what factors play a more significant role in fitting the Reimbursement Ratio (RR) of Medicare physicians. To meet this goal, we outlined the following objectives:

- Understanding the GLM procedure
- Finding appropriate data, cleaning, and analyzing the data
- Graphically examining factors for the model
- Utilizing the statistical software SAS to analyze and model the data using the significant factors

By following these objectives, we were able to explore the GLM procedure and indicate which factors were most useful in determining the fitted RR.

To pursue our objectives, we developed a methodology that consisted of four steps. The first step was cleaning and selecting the data that we would use. For instance, we limited our study to individual providers in the continental USA. These limitations were made so that no single factors such as being an organization or shipping costs due to being outside of the continental USA would be most significant compared to the other estimators of the RR. The second step was summarizing and analyzing all of our factors. Since we had several factors to

consider, we needed to select only the most significant ones so that our GLM was not over fitted to our data and could be used to estimate the RR in general. By performing one-way analysis to each factor, we were able to see the distributions of each factor, making it easier to select those that were statistically significant. Step three was selecting factors. After we compared these factors, we were able to evaluate which ones would be significant in calculating a fitted RR. The final step was the GLM analysis. In this step we took the remaining data and used SAS to develop a fitted model for the RR.

We utilized three different GLM procedures in SAS in order to find the best combination of distributions and link functions to build our model. Each test was conducted with the first half of the data, which calculated the model. From each model we were able to compute the fitted RR and compare it to the actual RR in order to find their correlations. We then chose the model that gave us the best correlation and used it to analyze the holdout data. Finally, we estimated the RR for the second half of the data using our model and calculated the correlation between this fitted RR and the actual RR.

Chapter 1: Introduction

Predictive modeling is an analytical method used to create statistical models that predict future behavior. A company can use predictive modeling to identify insurance risks, which can lead to improved underwriting and pricing. Traditional pricing methods in the United States are not statistically sophisticated. Claims for many lines of business are often analyzed using simple one-way and two-way analyses. Iterative methods known as minimum bias procedures, developed by actuaries in the 1960s, provide a significant improvement but are still only part way toward a full statistical framework (Anderson et al., 2007). A type of predictive modeling analysis method that has received widespread attention is the Generalized Linear Model (GLM).

The statistical framework of GLMs allows explicit assumptions to be made about the nature of the data and its relationship with predictive variables. The method of creating GLMs is more technically efficient than other standardized methods. Additionally, GLMs provide a statistical diagnosis which helps in selecting only significant variables and in validating model assumptions.

In order to learn more about GLMs and how they function, we conducted research and tested different distributions with a large amount of data, which we obtained from data.cms.gov, titled *"Medicare Physician and Other Supplier National Provider Identifier (NPI) Aggregate Report."* We analyzed this data and performed one-way analysis to determine which factors had the most impact on our response variable, which we called the Reimbursement Ratio (RR). This ratio accounted for the total amount paid to a Medicare provider divided by their total submitted charge. Once the most significant factors were selected, we split our data in half and used the first half to create a model that would estimate the RR. We used procedure GENMOD in SAS to

develop this model. Finally, we used our model to estimate the RR values for the second half of the data.

Chapter 2: Background

2.1. Linear Models

In order to fully understand the structure of GLMs, it is important to understand the classic linear model. The main purpose of the linear model is to express the relationship between an observed response variable (Y) and a number of predictor variables. GLMs observe this relationship. They are written in the form:

$$Y = \mu + \varepsilon$$

It is assumed that μ is the expected value of Y, and ε is the error term that is normally distributed with mean zero and variance σ^2.

Let us consider a simplified example of a private passenger auto classification system that has only two categorical rating variables: territory (urban or rural) and gender (male or female) (Anderson et al., 2007). The observed average claim severities are as follows:

	Urban	Rural
Male	800	500
Female	400	200

Table 1: Example 1 Average Claim Severity

In this example, the response variable, Y, is the average claim severity. The two factors, territory and gender, result in four different observed values: male (X_1), female (X_2), urban (X_3), and rural (X_4). These variables can either have a value of 0 or 1. In this case, the model would take the form:

$$Y = \beta_{male} X_1 + \beta_{female} X_2 + \beta_{urban} X_3 + \beta_{rural} X_4 + \varepsilon$$

However, this model has as many parameters as it does combinations of rating factor levels being considered, and there is a linear dependency between the four covariates X_1, X_2, X_3, and X_4. This means that the model is not uniquely defined - i.e. if any arbitrary value *k* is added to both β_{male} and β_{female}, and the same value *k* is subtracted from β_{urban} and β_{rural}, and the resulting model is equivalent. To make the model uniquely defined, we consider three variables instead of the four:

$$Y = \beta_{male} X_1 + \beta_{female} X_2 + \beta_{urban} X_3 + \varepsilon$$

This model assumes an average response for the *base case* of women in rural areas (β_{female}) with additional additive effects for being male ($\beta_{male} - \beta_{female}$) and for being in an urban area (β_{urban}). These observations can be expressed as the system of equations:

$$Y_1 = 800 = \beta_{male} + 0 + \beta_{urban} + \varepsilon_1$$

$$Y_2 = 500 = \beta_{male} + 0 + 0 + \varepsilon_2$$

$$Y_3 = 400 = 0 + \beta_{female} + \beta_{urban} + \varepsilon_3$$

$$Y_4 = 200 = 0 + \beta_{female} + 0 + \varepsilon_4$$

Next, we write out the sum of squared errors (SSE):

$$SSE = \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \varepsilon_4^2$$

$$= (800 - \beta_{male} - \beta_{urban})^2 + (500 - \beta_{male})^2 + (400 - \beta_{female} - \beta_{urban})^2 + (200 - \beta_{female})^2$$

We minimize these equations by taking the derivatives and setting them to zero:

$$\frac{\partial SSE}{\partial \beta \text{male}} = 0 \Rightarrow \beta_{\text{male}} + \beta_{\text{urban}} + \beta_{\text{male}} = 800 + 500 = 1300$$

$$\frac{\partial SSE}{\partial \beta \text{female}} = 0 \Rightarrow \beta_{\text{female}} + \beta_{\text{urban}} + \beta_{\text{female}} = 400 + 200 = 600$$

$$\frac{\partial SSE}{\partial \beta \text{urban}} = 0 \Rightarrow \beta_{\text{male}} + \beta_{\text{urban}} + \beta_{\text{female}} + \beta_{\text{urban}} = 800 + 400 = 1200$$

Solving these equations we get:

$$\beta_{\text{male}} = 525$$

$$\beta_{\text{female}} = 175$$

$$\beta_{\text{urban}} = 250$$

Using our equations we get the following predicted average claim severities:

$$Y_1 = \beta_{\text{male}} + 0 + \beta_{\text{urban}} = 525 + 0 + 250 = 775$$

$$Y_2 = \beta_{\text{male}} + 0 + 0 = 525 + 0 + 0 = 525$$

$$Y_3 = 0 + \beta_{\text{female}} + \beta_{\text{urban}} = 0 + 175 + 250 = 425$$

$$Y_4 = 0 + \beta_{\text{female}} + 0 = 0 + 175 + 0 = 175$$

Finally, we compare the fitted and the observed average claim severities in the following tables:

Fitted	Urban	Rural
Male	775	525
Female	425	175

Actual	Urban	Rural
Male	800	500
Female	400	200

Table 2: Fitted vs Actual Average Claim Severities

We can see that the four fitted values are close to the actual. The error in all cases is 25.

2.1.1. Linear Model Assumptions

The linear model assumes that all observations are independent and normally distributed. The linear model can be written in the following format:

$$Y = \hat{Y} + \varepsilon,$$

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n$$

Some other assumptions stated in *A Practitioner's Guide to Generalized Linear Models* are as follows:

- *Random Component:* Each component of Y is independent and normally distributed. The mean, μ_i, of each component is allowed to differ but the variance, σ^2, is the same.
- *Systematic Component:* Refers to the linear combination of explanatory variables that creates our predictor $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n$.
- *Link Function:* The relationship between the random and systematic components are defined by the link function. In a linear model, the link function is equal to the identity function so that:

$$Y = \hat{Y} + \varepsilon$$

2.1.2. Linear Model Limitations

Some limitations of the linear models stated in *A Practitioner's Guide to Generalized Linear Models* are as follows:

- It is difficult to insure that the response variables are normally distributed and that the variance is constant. Linear regression models transform data to fit the assumptions even when there is no reason for the transformation to exist.
- The values of the response variables may be restricted to be positive but the assumption of normality violates this restriction.
- If the response variable is strictly non-negative, then the variance of Y tends to zero as the mean of Y tends to zero. Therefore, the variance is a function of the mean.
- The additivity effects in the systematic component and the link function are not realistic because most of the time these predictor variables are entered multiplicatively in applications.

2.2. The Minimum Bias Procedure

Minimum bias procedures are iteratively standard univariate approaches. Each procedure involves the selection of a rating structure. These can be additive, multiplicative, or a combination of both. Additionally, there is a selection of a bias function which includes a balance principle, least squares, and maximum likelihood bias functions. The bias function is a way of comparing the procedure's observed loss statistics to the indicated loss statistics and measuring the error. Both sides of the equation must be weighted by the exposures to adjust for uneven mix of business.

For example, the balance principle applied to a multiplicative personal auto rating structure presented in the Casualty Actuarial Society's *Basic Ratemaking* is given below. This examples assumes two rating variables: gender and territory. Gender includes male (g_1) and female (g_2) and territory includes urban (t_1) and rural (t_2). We express female and rural as the base case (hence $g_2 = 1$ and $t_2 = 1$). The lost costs are given below:

	Urban	Rural	Total
Male	650	300	528
Female	350	240	244
Total	497	267	400

Table 3: Example 2 Loss Costs

The exposure distribution is as follows:

	Urban	Rural	Total
Male	170	90	260
Female	105	110	215
Total	275	200	475

Table 4: Example 2 Exposure Distribution

The balance principle requires that the exposure weighted observed loss costs equal the indicated exposure weighted loss cost of each rating variable. The four equations below show the observed weighted loss costs on the left and the indicated weighted loss costs on the right. The base case is assumed to be $100.

Males: 170 x \$650 + 90 x \$300 = \$100 x 170 x g_1 x t_1 + \$100 x 90 x g_1 x t_2

Females: 105 x \$250 + 110 x \$240 = \$100 x 105 x g_2 x t_1 + \$100 x 110 x g_2 x t_2

Urban: 170 x \$650 + 105 x \$250 = \$100 x 170 x g_1 x t_1 + \$100 x 105 x g_2 x t_1

Rural: 90 x \$300 + 110 x \$240 = \$100 x 90 x g_1 x t_2 + \$100 x 110 x g_2 x t_2

Next, we choose a seed for one of the rating variables. So the urban relativity is the total loss costs divided by the total rural loss costs:

t_1 = 1.86 = \$497/\$266

t_2 = 1.00

We substitute these seed values into the first two equations above and solve for the values of g_1 and g_2:

170 x \$650 + 90 x \$300 = (\$100 x 170 x g_1 x 1.86) + (\$100 x 90 x g_1 x 1.00)

\$137,500 = (\$31,620 x g_1) + (\$9,000 x g_1)

\$137,500 = \$40,620 x g_1

g_1 = 3.39

105 x \$250 + 110 x \$240 = (\$100 x 105 x g_2 x 1.86) + (\$100 x 105 x g_2 x 1.00)

\$52,650 = (\$19,530 x g_2) + (\$11,000 x g_2)

\$52,650 = \$30,530 x g_2

g_2 = 1.72

We now use these seed values for g_1 and g_2 and set up equations to solve for the new values of t_1 and t_2.

170 x $650 + 105 x $250 = ($100 x 170 x 3.39 x t_1) + ($100 x 105 x 1.72 x t_1)

$136,750 = ($57,630 x t_1) + (18,060 x t_1)

$136,750 = $75,690 x t_1

t_1 = 1.81

90 x $300 + 110 x $240 = ($100 x 90 x 3.39 x t_2) + ($100 x 110 x 1.72 x t_2)

$53,400 = ($30,510 x t_2) + ($18,920 x t_2)

$53,400 = $49,430 x t_2)

t_2 = 1.08

This procedure is repeated until there is no significant change in any of the values of g_1, g_2, t_1, and t_2. At this point, it is common to normalize the base case (g_2) relativities to 1.00.

g_1 = 3.39/1.72 = 1.97

g_2 =1.72/1.72 = 1.00

t_1 = 1.81/1.08 = 1.68

t_2 = 1.08/1.08 = 1.00

To conclude, the base loss cost also needs to be adjusted to reflect the normalization:

$$\textit{Base loss cost} = \$100 \times 1.72 \times 1.08 = \$185.76$$

Our fitted versus actual loss costs are as follows:

Fitted	Urban	Rural	Total	Actual	Urban	Rural	Total
Male	615	366	529	**Male**	650	300	528
Female	312	186	248	**Female**	350	240	244
Total	499	267	402	**Total**	497	267	400

Table 5: Fitted vs Actual Loss Costs

It is important to note that the example above only considers one of the minimum bias methods (the multiplicative structure). Additionally, it only considers two rating variables with two levels each. Incorporating several rating variables requires some programming. Many minimum bias procedures are a subset of GLMs. GLMs consider all rating variables simultaneously and automatically adjust for exposure correlations between rating variables. Multivariate methods, such as GLMs, also remove unsystematic effects in the data as much as possible. The minimum bias method fails to do so.

2.3. Generalized Linear Models

GLMs comprise a wide range of models that include linear models as a case. However, the requirement for all components of Y to be normally distributed and have a common variance is removed. Another difference between GLMs and linear models is that the effect of the variables on Y is not assumed to be additive.

2.3.1. Components of Generalized Linear Models

The components of a general linear model as stated in *A Practitioner's Guide to Generalized Linear Models* are:

- *Random component:* Accounts for the probability distribution of Y (the response variable.) As previously stated, each of its components is independent and from one of the exponential family of distributions.
- *Systematic component:* Refers to the linear combination of explanatory variables that creates our predictor $\eta = X\beta$ (e.g., $\beta_0 + \beta_1 x_1 + \beta_2 x_2$.)
- *Link function:* Specifies the relationship (link) between the previous two components. The link function must be differentiable. It shows how the expected value of response variables relates to our predictor. E.g. $\eta = g(E(Y_i))$, $g(x)$ is the link function.

The exponential family of distributions includes several common distributions such as Normal, Poisson, Exponential, Gamma, Binomial and Inverse Gaussian. These are completely specified in terms of its mean and variance, while its variance is in turn a function of its mean.

The following table shows some of the models comprised by GLMs, according to Agresti (Ch. 4, 2013):

Model	Random	Link	Systematic
Linear Regression	Normal	Identity	Continuous
ANOVA	Normal	Identity	Categorical
ANCOVA	Normal	Identity	Mixed
Logistic Regression	Binomial	Logit	Mixed
Loglinear	Poisson	Log	Categorical
Poisson Regression	Poisson	Log	Mixed
Multinomial response	Multinomial	Generalized Logit	Mixed

Table 6: GLM Models

2.3.2. Further Assumptions about Generalized Linear Models

- Errors need to be independent but do not need to be normally distributed.
- GLMs rely on sufficiently large samples (detailed further on.)
- GLMS estimate the parameters using maximum likelihood estimators instead of ordinary least squares.

2.3.3. Typical GLMs

Different GLM models are used in accordance with the assumptions we need to make about distribution of data. For example, typical model for insurance claim frequencies is Poisson because of its time memoryless character (i.e. measuring the frequencies per month and frequencies per year will yield the same results). On the other hand, to model insurance severities, Multiplicative Gamma distribution is typically used due to it being invariant to measures of currency. This distribution makes it possible to get the same result whether the measurements are made in cents or in dollars.

In general, typical model forms are as in the table below (Anderson et al., 2007):

Y	**Claim frequencies**	**Claim numbers or counts**	**Average claim amounts**	**Probability (e.g. of renewing)**
Link function *g(x)*	ln(x)	ln(x)	ln(x)	ln(x/(1-x))
Error	Poisson	Poisson	Gamma	Binomial
Scale parameter *φ*	1	1	Estimated	1
Variance function *V(x)*	x	x	x^2	x(1-x)
Prior weights *ω*	Exposure	1	# claims	1
Offset *ξ*	0	ln(exposure)	0	0

Table 7: Typical Model Forms

2.4. Maximum Likelihood Estimator

After deciding the distribution of the response variable and the link function, we used the Maximum Likelihood function to determine the values of the covariates. To compute these values, we need to maximize the likelihood function which is the same as maximizing its logarithm. The core of this method is attempting to find parameters which will result in fitted values as close as possible to the original ones.

The likelihood function is defined as the product of probabilities of observing each value of the y-variate. Typically, we consider the log of the likelihood function since being a summation across observations rather than a product makes the calculations more manageable.

In the tables below the observed average claim severity for the following cases is presented:

MALE	LUXURY	REGULAR
OLD CAR	1400	1080
NEW CAR	1550	1230

Table 8: Example 3 Average Claim Severity Male

FEMALE	LUXURY	REGULAR
OLD CAR	1420	1100
NEW CAR	1570	1250

Table 9: Example 3 Average Claim Severity Female

The first step to applying the GLM procedure to analyze the following case is to identify the factors that account for the variations in observed average claim in the given cases. There are three such factors, namely: gender, classification of the car, age of the car. Each of these factors

has two levels: male (X_1) and female (X_2) for gender, luxury (X_3) or regular (X_4) for the classification of the car and old car (X_5) and new car (X_6) for the age of the car. These indicator variables take the value 1 or 0. For example, the male covariate, (X_1), is equal to 1 if the gender is male, and 0 otherwise.

The purpose of the linear model is to express the observed item Y (average claim severity) as a linear combination of a specific selection of the six variables, plus a normal random variable ε with mean zero and variance σ^2, often written $\varepsilon \sim N(0, \sigma^2)$. One such model might be:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon$$

However, this model has as many parameters as it does combinations of rating factor levels being considered, and there is a linear dependency between the six covariates. This means that the model in the above form is not uniquely defined. To make this model uniquely defined we consider selecting a base case, and reducing the number of variables to three instead of six. We will do this by assigning variables only to one level for each of the given factors. We chose the average claim severity for males who have a regular, old car as our base case (1080 severities). This leads to our updated model:

$$Y = 1080 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 + \varepsilon$$

Our new parameters are: $\beta_2 \rightarrow$ effect of being a female, $\beta_3 \rightarrow$ effect of having a luxury car and $\beta_6 \rightarrow$ effect of having a new car.

The next step of applying the GLM procedure is to specify the design matrix and the vector of parameters β.

Since the parameters are β_2, β_3, and β_6 the vector of parameters is:

$$\beta = \begin{bmatrix} \beta_2 \\ \beta_3 \\ \beta_6 \end{bmatrix}$$

Based on this the design matrix will be:

$$X = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

And the response matrix will be:

$$Y = \begin{bmatrix} 1400 - 1080 \\ 1080 - 1080 \\ 1550 - 1080 \\ 1230 - 1080 \\ 1420 - 1080 \\ 1100 - 1080 \\ 1570 - 1080 \\ 1250 - 1080 \end{bmatrix} = \begin{bmatrix} 320 \\ 0 \\ 470 \\ 150 \\ 340 \\ 20 \\ 490 \\ 170 \end{bmatrix}$$

The classical linear model case assumes a normal error structure and an identity link function.

The predicted values in the example take the form:

$$E[Y] = g^{-1}(X * \beta) = \begin{bmatrix} \beta_3 \\ 0 \\ \beta_3 + \beta_6 \\ \beta_6 \\ \beta_2 + \beta_3 \\ \beta_2 \\ \beta_2 + \beta_3 + \beta_6 \\ \beta_2 + \beta_6 \end{bmatrix}$$

Since error will have Normal distribution, then the response variable "*Y*" will have Normal distribution as well. Therefore, we will have to consider probability density function of Normal distribution as below:

$$f(Y;\ \mu,\sigma^2) = exp\left\{\frac{-(Y-\mu)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\}$$

To get the best estimation, we use the strategy of maximizing the likelihood function. Likelihood function is described with the following expression:

$$L(Y;\ \mu,\sigma^2) = \prod_{i=1}^{n} exp\left\{\frac{-(Y_i-\mu_i)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\}$$

Since maximizing likelihood function is the same as maximizing log-likelihood function, we use the log-likelihood function as below:

$$l(Y;\ \mu,\sigma^2) = \sum_{i=1}^{n} \frac{-(Y_i-\mu_i)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$$

After applying identity link function:

$$l(Y;\ \mu,\sigma^2) = \sum_{i=1}^{n} \frac{-\left(Y_i-\sum_{j=1}^{p} X_{ij} * \beta_j\right)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$$

Now, by setting to 0 the partial derivatives of log-likelihood function we get the following:

$$\frac{\partial(l(Y;\mu,\sigma^2))}{\partial\beta_2} = 0 \rightarrow 2\beta_2 + \beta_3 + \beta_6 = 510$$

$$\frac{\partial(l(Y;\mu,\sigma^2))}{\partial\beta_3} = 0 \rightarrow 2\beta_3 + \beta_2 + \beta_6 = 810$$

$$\frac{\partial(l(Y;\mu,\sigma^2))}{\partial\beta_6} = 0 \rightarrow 2\beta_6 + \beta_2 + \beta_3 = 640$$

Solving the created system of equations, we get:

$$\beta_2 = 20, \beta_3 = 320, \beta_6 = 150.$$

Using these values we get the following tables of fitted results:

Fitted		
MALE	LUXURY	REGULAR
OLD CAR	1400	1080
NEW CAR	1550	1230

Actual		
MALE	LUXURY	REGULAR
OLD CAR	1400	1080
NEW CAR	1550	1230

Table 10: Example 3 Fitted vs Actual Results Male

Fitted		
FEMALE	LUXURY	REGULAR
OLD CAR	1420	1100
NEW CAR	1570	1250

Actual		
FEMALE	LUXURY	REGULAR
OLD CAR	1420	1100
NEW CAR	1570	1250

Table 11: Example 3 Fitted vs Actual Results Female

We see that these tables are exactly the same as the original ones. This happens because the data we have has a strong linear correlation that can be noticed even with a careful observation. For example, every respective entry in the table of females is exactly 20 more that the entry in the table of males (i.e. 1420 = 1400 + 20; 1100= 1080 + 20 etc.). The same happens with other variables as well. Since the linear correlation is perfect, then by using Normal distribution to model the errors and the identity link function in the GLM procedure, we replicate the simple linear model. And, since the linear correlation of covariates is perfect, this model will conclude in perfectly predicting the original results.

Now, let us try doing the GLM procedure using Poisson distribution for errors and a different link function. In this case, the analysis of covariates is the same as in the first part of the example. We only need to describe the second part of the procedure where logarithm link function and Poisson distribution for errors are involved.

The predicted values will take the form:

$$E[Y] = g^{-1}(X * \beta) = \begin{bmatrix} e^{\beta_3} \\ 1 \\ e^{\beta_3+\beta_6} \\ e^{\beta_6} \\ e^{\beta_2+\beta_3} \\ e^{\beta_2} \\ e^{\beta_2+\beta_3+\beta_6} \\ e^{\beta_2+\beta_6} \end{bmatrix}$$

Since error will have Poisson distribution, then the response variable "*Y*" will have Poisson distribution as well. Therefore, we will have to consider probability density function of Poisson distribution as below:

$$f(Y;\ \mu) = \frac{e^{-\mu}\ \mu^{Y}}{Y!}$$

To get the best estimation, we use the strategy of maximizing the likelihood function. Likelihood function is described with the following expression:

$$L(Y;\ \mu) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\ \mu_i^{Y_i}}{Y_i!}$$

Since maximizing likelihood function is the same as maximizing log-likelihood function, we use the log-likelihood function as below:

$$l(Y;\ \mu) = \sum_{i=1}^{n} -\mu_i + Y_i \ln(\mu_i) - \ln(Y_i!)$$

After applying logarithm link function:

$$l\left(Y;\ e^{X\beta}\right) = \sum_{i=1}^{n} -exp\left(\sum_{j=1}^{p} X_{ij} * \beta_j\right) + Y_i * \left(\sum_{j=1}^{p} X_{ij} * \beta_j\right) - \ln(Y!)$$

Now, by setting to 0 the partial derivatives of log-likelihood function we get the following:

$$\frac{\partial\left(l(Y;\ \mu,\sigma^2)\right)}{\partial\beta_2} = 0 \ \rightarrow e^{\beta_2}\left(e^{\beta_3} + 1 + e^{\beta_3+\beta_6} + e^{\beta_6}\right) = 1020$$

$$\frac{\partial(l(Y;\ \mu,\sigma^2))}{\partial\beta_3} = 0 \ \rightarrow e^{\beta_3}\left(1 + e^{\beta_6} + e^{\beta_2} + e^{\beta_2+\beta_6}\right) = 1620$$

$$\frac{\partial(l(Y;\ \mu,\sigma^2))}{\partial\beta_6} = 0 \ \rightarrow e^{\beta_6}\left(e^{\beta_3} + 1 + e^{\beta_2+\beta_3} + e^{\beta_2}\right) = 1280$$

Solving the created system of equations, we get:

$$\beta_2 = 0.51, \beta_3 = 4.88, \beta_6 = 1.29.$$

Using these values we get the following tables of fitted results:

Fitted		
MALE	LUXURY	REGULAR
OLD CAR	1211.63	1080
NEW CAR	1558.19	1083.63

Actual		
MALE	LUXURY	REGULAR
OLD CAR	1400	1080
NEW CAR	1550	1230

Table 12: Example 3 Poisson Fitted vs Actual Results Male

Fitted		
FEMALE	LUXURY	REGULAR
OLD CAR	1299.20	1081.67
NEW CAR	1876.32	1086.05

Actual		
FEMALE	LUXURY	REGULAR
OLD CAR	1400	1080
NEW CAR	1550	1230

Table 13: Example 3 Poisson Fitted vs Actual Results Female

We notice that even though the fitted values are relatively close to the original values, the fit is far from perfect. This happens because the data is perfectly linear and every other fit except identity link function and Normal distribution of response variable will give less accurate results.

Chapter 3: Methodology

3.1. Introduction

The goal of this project was to work with physician Medicare data to identify key factors that would affect the ratio of the actual payment that providers received to the amount that they charged, which we will define as the Reimbursement Ratio (RR). For instance, if a provider has submitted a charge of $100 and Medicare paid them $75, then their RR would be $75/$100 = 75%. In general, RR can be calculated by the following formula:

$$\text{Reimbursement Ratio (RR)} = \frac{\text{Total Amount Paid}}{\text{Total Submitted Charge}}$$

We wanted to identify variables that would have a strong effect on this ratio. To accomplish the overarching project goal, we executed the following set of objectives. The first one was to gather, review, and clean up the physician Medicare data. The second objective was to analyze the relationship between the RR and the factors that we selected after the data cleaning process. These factors were: state of the provider, provider type, number of services performed by the provider, and the "greediness factor." We define this factor as the amount (per service) charged by the provider compared to the average of their specialty. The greediness factor for a given provider can be calculated using the following formula:

$$\text{Greediness factor} = \frac{\text{Submitted Charge of the Provider (per service)}}{\text{Average Submitted Charge per service of their specialty}}$$

Finally, our third objective was to utilize SAS to analyze variables that had more influence on the RR. The overall methodology for this project is represented by the flow chart in Figure 1.

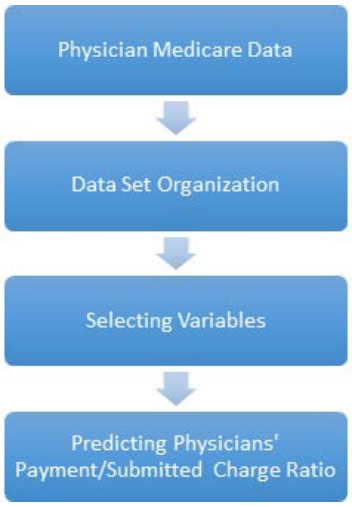


Figure 1: Methodology Flow Chart

3.2. Data Set Organization

3.2.1. Removing Factors

Our first step was to obtain Medicare information on providers through the government website, data.cms.gov. We were able to find a spreadsheet titled, "Medicare Physicians and Other Supplier National Provider Identifier Aggregate Report Calendar Year 2014", which consisted of information on utilization and payment data of doctors and medical organizations throughout the year of 2014. We had access to 986,677 providers and 70 categories of information, which included personal data such as their name, credentials, state and address, as well as more relevant data such as "total submitted charge amount" and "total Medicare payment". However, we limited our analysis to the payment data of individual medical providers by removing Alaska, Hawaii, and other territories outside of the continental USA to eliminate changes in price due to a possible rise in shipping costs. Whereas, since individuals and

organizations greatly differ in the number of services that they provide, we chose to work only with individuals (they represented 94% of the providers). Additionally, we removed possible outliers such as providers that reported more than 2000 services (as they were above 300% of the mean) and providers that did not report their specialty or whose specialty had very little representation, such as providers who worked in the sleep medicine industry. This procedure would allow our GLM to fit the RR more efficiently. Figure 2 represents our cleaning process along with the number of providers that remained after each step.

Figure 2: Data cleaning process

3.2.2. Summarizing Data

Next, the data was summarized graphically, allowing us to see the trends in the data and the relationship between different variables. Graphs helped the team to understand the data provided and revealed trends that needed to be further investigated. Additionally, they allowed us to focus on specific factors in our data and helped us eliminate factors that did not have any important effects on the RR. An example of a graph that would help describe the data is the distribution of the RR based on the number of services performed by the physicians as seen in Figure 3 (bigger version in Appendix A)

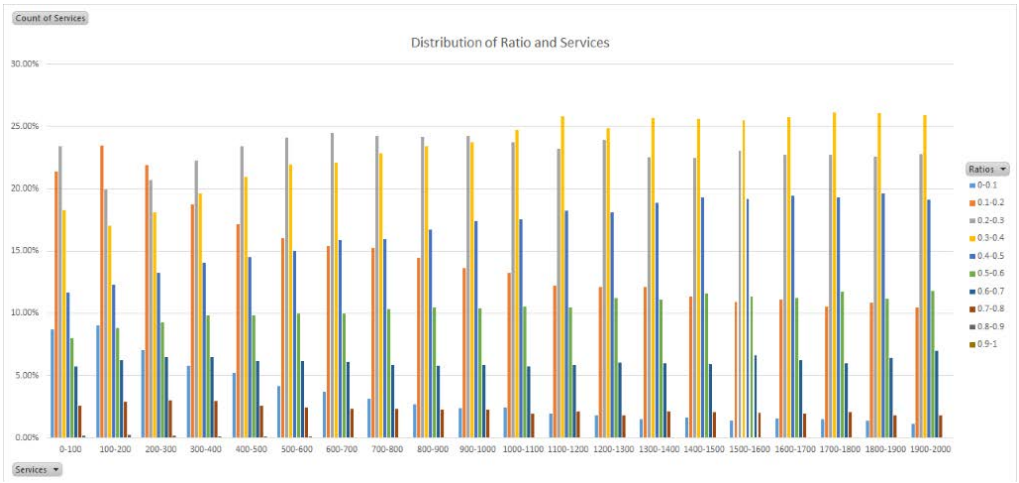


Figure 3: Distribution of Reimbursement Ratio and Number of Services

The graph above shows the RR as a percent based on the number of services the providers performed. The number of services is broken up into intervals of 100 up till 2000 services. We can see in the graph that in most categories, the common ratio range encountered is 30%-40% (yellow bar), however, 20%-30% (gray bar) is predominant in the some of the ranges. This graph was consistent with the average of the ratios which is about 27%. What this graph tells us is that generally, physicians who perform more services have a larger RR. In other words, they get more of what they ask for. Once this data set organization was complete, the team worked on analyzing and selecting our independent variables.

3.3. Selecting Variables

The next step was to determine which factors were meaningful. We performed one-way analysis on all the different categories of data by comparing them to the RR to check for a correlation. We then selected the four factors that had the greatest correlation to the RR, which were: state of the provider, provider type, number of services and the "greediness factor". Identifying these variables was an important part of the generalized linear modeling result because the results of the model are interpreted based on the impact of these predictors on our

response variable. Additionally, we analyzed these factors graphically for better visualization. Figure 4 below shows the RR depending on the state of the provider (bigger version in Appendix A).

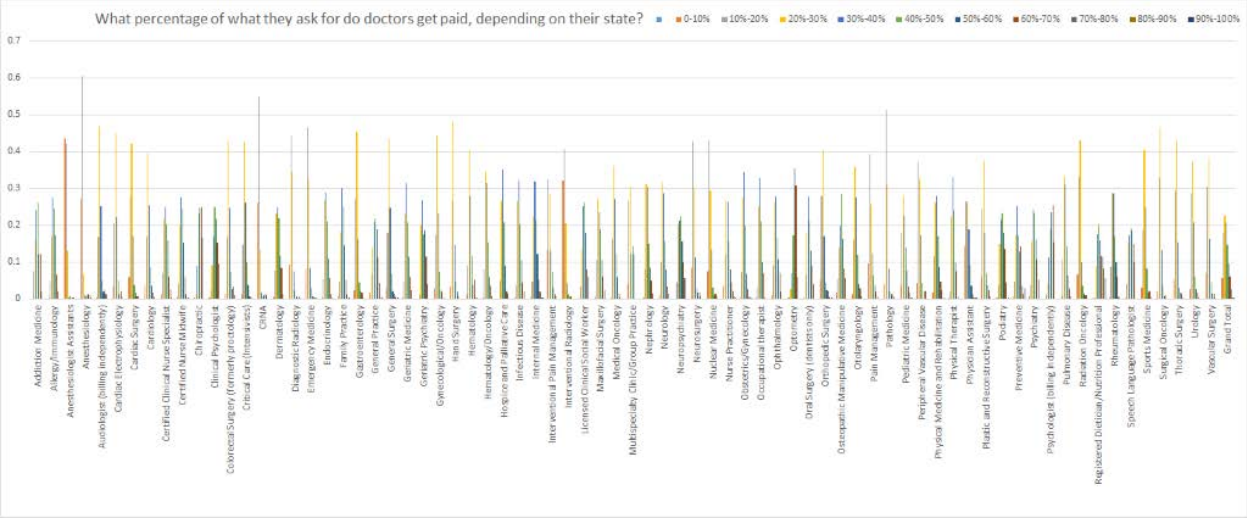


Figure 4: Payment ratio depending on State

We can see in the graph above that most states in the yellow bars (30%-40%) are highest followed by the gray bars (20%-30%). However there are a few states for which the blue and orange bars are highest. This tells us that the state the physician is from is a relevant factor for our analysis.

We illustrate an example of a distribution we found to not be relevant in Figure 5 below.

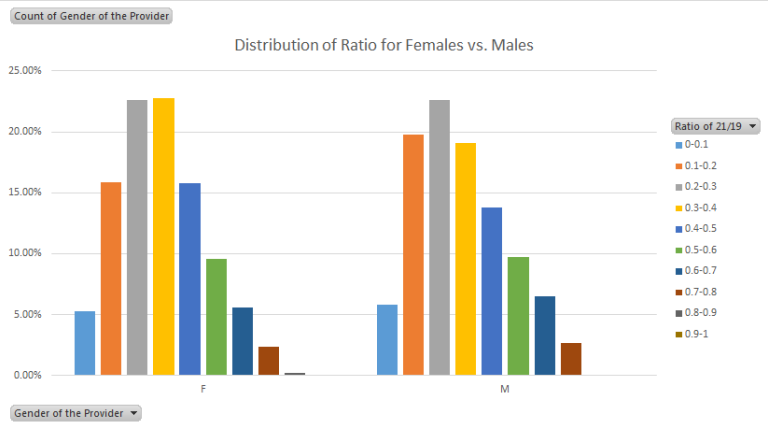


Figure 5: Distribution of Payment Ratio for Females and Males

The graph above shows that both male and female providers share a similar distribution, therefore, we did not take this factor into consideration for our GLM analysis. Several other factors showed similar distributions as in this graph so we did not consider those factors either. After selecting our input variables we moved onto the modeling process.

3.4. Generalized Linear Model Analysis

3.4.1 Setting up the Data

First, we split the data in half in order to use one half for calibration and the other half for testing. We separated the data using the "rand" function in Excel to assign each row of data a random number. We then sorted the data in numerical order based on these random values and then split the data in half. We used the first half of the data to build our GLM and the second half to test that model. Since the number of services and the greediness factor had over a thousand different values, we grouped them so that we had about the same amount of entries as we did for state and provider types. To do this, we grouped the services into intervals of 50, starting at 1 and ending at 2000, and the greediness in intervals of 10 percent, starting from 0 to 300 percent.

However, since several providers had a greediness factor much higher than 300 percent, we grouped them all in a "+300%" interval so that we did not have groups with a single provider.

3.4.2 Running the data in SAS

The next step was to run the first half of the data using the GENMOD procedure in SAS (Appendix B). This procedure allowed us to acquire estimates for each level of our four factors. We ran the GENMOD procedure with three different combinations of distributions and link functions (see Table 4 in Chapter 2 for the full list). The distributions with their respective link functions are listed below:

1. Normal distribution & Identity link function
2. Poisson distribution & Logarithmic link function
3. Gamma distribution & Logarithmic link function

3.4.3 Calculating Results

We then compared the fitted values of RR produced by each of these combinations to the actual RR of each provider. The Normal distribution with the identity link function, which is the linear model, yielded the best result in terms of correlation between the actual and fitted values. Therefore we used this model to estimate the fitted values of the remaining half of the data. We then calculated the correlation between the model's estimates for the second half of the data and their respective RR's. To conclude our project, we compared the correlations for each half of the data and calculated several statistical measurements such as the R-squared, the root mean squared error (RMSE), and the mean absolute error (MAE).

Chapter 4: Results

After conducting the GLM procedure on our four variables, we were able to analyze our results. This section presents the results of our analysis of the three distributions, the fitted model, the estimates for the four significant variables, and the comparison of the fitted model to the estimates.

4.1. Fitted Model

The fitted values of the aforementioned combinations of distributions and link functions were compared to the actual RRs and yielded the following statistical measurements:

Distributions/Link functions	Correlation	R-squared	RMSE	MAE
Normal/Identity	0.8289	0.6871	0.0947	0.0700
Poisson/Logarithmic	0.8276	0.6849	0.0950	0.0672
Gamma/Logarithmic	0.8165	0.6667	0.0986	0.0679

Table 14: Statistical measurements of the distributions/link functions

As shown in Table 14 above, the most optimal results were obtained when using the Normal distribution with the identity link function. The correlation and R-squared for this combination are greater, while the root mean squared error (RMSE) is smaller than the rest. Therefore, we selected the results provided by this distribution and link function to create our fitted model. Our model can be described by the following equation:

$$\hat{Y} = \beta_0 + \beta_{STi} + \beta_{PRj} + \beta_{SRk} + \beta_{GRh}$$

where i = 0, 1, …, 49, j = 0, 1, …, 73, k = 0, 1, …, 40, and h = 0, 1, …, 31.

As seen in the formula, β_0 is the y-intercept in our model, β_{STi} is the estimate of the effect of the i-th level of the state, β_{PRj} is the estimate of the effect of the j-th level of the provider type, β_{SRk} is the estimate of the effect of the k-th level of the number of services, and β_{GRh} is the estimate of the effect of the h-th level of the "greediness". Table 15 also shows some of the estimates for each factor.

Coefficients	i, j, k, h = 1	i, j, k, h = 2	…	i, j, k, h = n-1	i, j, k, h = n	
Intercept (β_0)	$\beta_0 = 0.2068$	$\beta_0 = 0.2068$	…	$\beta_0 = 0.2068$	$\beta_0 = 0.2068$	
State (β_{STi})	$\beta_{ST1} = -0.0023$	$\beta_{ST2} = -0.022$	…	$\beta_{ST(n-1)} = -0.0207$	$\beta_{STn} = 0$	n = 49
Provider (β_{PRj})	$\beta_{PR1} = 0.045$	$\beta_{PR2} = 0.1478$	…	$\beta_{PR(n-1)} = 0.0144$	$\beta_{PRn} = 0$	n = 73
Services (β_{SRk})	$\beta_{SR1} = 0.0016$	$\beta_{SR2} = 0.0077$	…	$\beta_{SR(n-1)} = -0.0045$	$\beta_{SRn} = -0.0046$	n = 40
Greediness (β_{GRh})	$\beta_{GR1} = 0.1792$	$\beta_{GR2} = 0.1993$	…	$\beta_{GR(n-1)} = -0.143$	$\beta_{GRn} = -0.1715$	n = 31

Table 15: Examples of Estimate for each Factor

Each subscripted variable displayed in the table refers to the estimate for the specific level of each factor. The whole list of these estimates is shown in Appendix C. Note that each factor has a different number of levels, indicated in the last column of the table.

4.2. Estimate Analysis

By analyzing each factor and its estimates, we were able to understand the weight that each level has on the resulting RR. The maximum and minimum for each of the four factors are listed below:

	State	Greediness	Provider Type	Number of Services
Maximum	New Jersey (0.0432)	10%-20% (0.1993)	Chiropractor (0.3467)	100-150 (0.0087)
Minimum	Wisconsin (-0.0467)	+300% (-0.1715)	Interventional Radiology (-0.1065)	1800-1850 (-0.0075)

Table 16: Maximum and Minimum for Estimate of Significant Variables

The middle row of the table represents the level of each factor that has the greatest positive impact on the RR of a physician. The bottom row represents the levels with the greatest negative impact. For example, our model estimates that living in New Jersey will add 0.0432 to a physician's RR, while working on Interventional Radiology would subtract 0.1065 from such RR.

It becomes clear then that the best possible scenario (upper bound) for physicians that wish to maximize their RR is working in New Jersey, as Chiropractors, asking for 10 to 20 percent of the average of their industry, and performing from 100 to 150 services per year. On the other hand, the lower bound for a physician's RR would be achieved by working in Wisconsin, as an Interventional Radiologist, asking for more than 300 percent of the industry average (greediness factor), and performing from 1800 to 1850 services per year.

As an important note, we must clarify that our model indicates the effect of factors and levels on the RR, not on the overall income of the physicians. It is crucial to consider this in order to correctly understand the impact of the greediness factor. For instance, even though our model considers 10 to 20 percent as the ideal greediness level for doctors, it is clear that asking for about 15% of the industry average is most likely not the strategy to follow when it comes to generating greater earning. Our model analyzes the RR of a physician, and does not help physicians achieve the greatest possible income. Information on the greediness level that would produce the most substantial payment will be provided later.

4.3. Testing the Model

To test the validity of our model, we assigned the estimated value for each level of each factor to the second half of our data, and then compared the fitted RRs to the actual ones. Our results were satisfactory and are displayed below:

Correlation	R-squared	RMSE	MAE
0.8291	0.6874	0.0946	0.0699

Table 17: Statistical measurements of the fit of our model to the second part of the data

As shown in Table 17 above, the statistical measurements between the fitted and actual RRs were virtually identical (even slightly better) than those obtained by applying our GLM to the first half of the data. This shows that our model is not over fit to our dataset and efficiently estimates the RR based on our four factors. We were surprised to see that the correlation was roughly 83%, since we expected it to be much lower because we did not include provider qualifications. However, we realized that it is not the best correlation because the R-squared is

only 69%. Therefore, the model explains only 69% of the variability of the RR, leaving more room for improvement.

4.4. Real World Applications

As mentioned previously, our model is concerned with estimating the RR of providers and not their overall pay. However, the greediness factor can help providers to come up with an optimal strategy to maximize earnings.

The table in Appendix D: Greediness Ranking Levels is intended to explain the outcome of charging for each of the 30 possible greediness categories (0%-10%, 10%-20% etc.). These intervals are expressed as their midpoint for calculation purposes. The first column simply lists the 30 categories. The second column shows our fitted RR for each level. For example, our estimation is that providers charging 10 to 20 percent per service of the industry average will get 40.61% of what they ask for (which is their RR). The third column is the product of the first two, this accounts for the percentage of the industry average submitted charge that providers receive depending on their greediness level. For example, the first row of data tells us that providers that ask for 5% of the average industry charge per service will receive 38.6% of this amount. That would mean 1.93% (of the average industry charge per service).

Finally, the last two columns of our table represent a reordering of the greediness levels. These are ranked from the largest to smallest when it comes to generating returns. The optimal strategy would be to charge from 230 to 240 percent (per service) of the industry average. However, it is also important to note that asking for 230 to 240 percent (per service) of the industry average will yield less than 1% more than asking for 90 to 100 percent (per service). In other words, being "greedy" is just slightly better than asking for the industry average.

Additionally, this table shows that the worst possible strategy for medical providers is to ask for less than 70 percent of their industry average as this would lead to the lowest income.

It is important to note that these calculations take into account the base levels of State, Provider Type, and Number of Services (Wyoming, Vascular Surgery and “950-100”), and these have an estimated value of zero. It is clear that the returns generated by the previously discussed greediness levels are also subject to these other three factors.

Chapter 5: Conclusion and Recommendations

5.1. Conclusion

Once we finished studying our results, we came up with conclusions that summarized the highlights of our analysis. This section presents what we interpreted from our results and useful information for providers.

5.1.1. Factor ranking

One of the most important conclusions inferred from our model, was the importance of each factor in estimating the RR. We compared our estimates of the levels of each factors and observed that the higher estimates (in absolute value) belonged to type of the provider. That indicates that the industry of each provider plays the most significant role in their rate of reimbursement (RR).

A crucial achievement of our project was discovering that the "greediness level", which we defined as the amount (per service) charged by the provider compared to the average of their industry, plays a very significant role in determining the RR. For instance, being more "greedy" will result in having a lower rate of reimbursement, while being less "greedy" will do the exact opposite. However, as we explained under Section 4.4, this holds only for the rate of reimbursement and not the real income.

Moreover, we discovered that the state where the provider operates is the third factor with most influence to the RR, leaving the number of services as the least important factor. We also observed that the number of services had little weight in our model and could also be omitted with very little loss of estimating accuracy.

Perhaps the most meaningful part of our results is the greediness levels' ranking. Out of the four factor that we took into consideration (state, greed, number of services and provider type), greed is the only one that is completely controlled by the providers. Therefore it was interesting for us to see how the greediness level affected providers' income in ways that are not obvious (Appendix D).

5.1.2. Applications

As we reveal in our results section, asking for 230 to 240 percent (per service) of the industry average yields the largest earnings for medical providers. However, this income does not differ much from that of the medical providers that ask for 90 to 100 percent (per service) of the industry average. In other words, being very "greedy" does not help much more than asking the usual average of your industry. Additionally, this table shows that the strategy that medical providers should definitely avoid is to ask for less than 70 percent of the industry average, since that would lead to significantly lower payments.

5.2. Recommendations

This section discusses the challenges we faced while carrying out the project. It also includes information on how to make future projects more precise.

5.2.1. What went well

In order to get the results, we had to analyze a large amount of data efficiently so that we did not make computational errors. We needed to work with Excel and SAS effectively to avoid these errors. By combining both of these software, we were able to work faster and more accurately. For example, to run SAS correctly, we had to rename provider types and group services. Additionally, we sorted greediness into intervals so that SAS could read them and not

cut off our entries. Since we had over 700,000 data entries, it was difficult to ensure that all of the changes were made to each cell but by using more advanced formulas in Excel, we were able to make precise changes. Otherwise, our fitted model could have interpreted a worse correlation, making our results useless.

Another aspect that went well with our project was the timely feedback. Whenever we came across a problem or question that could hold our project back, we contacted people in the Medicare industry to get clearer definitions on variables. For example, when we first looked at our dataset, we did not know what it meant by "services" so we contacted the website where we found our data and they explained the exact meaning of a service to us. By getting the feedback quickly, we were able to continue our analysis and move onto testing without losing valuable data.

Creating the "greediness" factor also had a significant role in getting meaningful results. The amount (per service) charged by the provider compared to the average of their industry was very significant in fitting the RR. This factor helped us get a more profound understanding of our results, since it allowed us to see the impact that the submitted charge had on the reimbursement ratio of the providers.

5.2.2. Difficulties

The most difficult part of our project was finding data that would allow us to build a general linear model. We only had access to data published publically by the government. Initially, we thought finding car insurance claims data would be ideal for our project but that information is private. Another problem was that most of the data we had access too had very little predictive power. In other words, these datasets contained mostly one field that we would

be interested in modeling and all the other fields consisted only of personal information such as first and last name, address, limited demographics (age was often not included), which had no predictive power on the response. After careful research, we selected the "Medicare Physician and Other Supplier National Provider Identifier (NPI) Aggregate Report, Calendar Year 2014", which was a much more complete information in that it consisted of more than 900.000 data points and 70 different columns that included information such as provider, state, number of services, etc., that we would expect to play a role on our selected response variable (the RR).

Another difficulty in our study was understanding and interpreting this data correctly. Our background on medical field was not very advanced and many concepts were initially vague. To improve our understanding, we used the information and explanations that were given in the website where we found the data and did some online research. However, we were still missing information about some concepts which were specific to our dataset such as what constitutes of a service. This information was crucial, since we believed that the number of services played an important role in estimating the RR in our data. Therefore, we contacted the data providers and asked for an explanation. Fortunately, they sent us a detailed explanation, where they defined service as an appointment of 11-15 minutes with a medical provider. For instance, if a patient had a 30 minute appointment with a medical provider, then the medical provider would have performed two full services with that patient.

Unfortunately, the recording of provider qualifications in our data also posed as a complication. We believed that this factor could have an impact on the RR and wanted to include it in our analysis. However, the formatting made that impossible for us. Same qualifications were declared in more than 20 different manners and even with all of our efforts, we could not format this field so that it would be usable to our project and therefore did not include it in our analysis.

Deciding which software to use to analyze the data was also a challenge. Through research, we discovered that fitting a general linear model was already a built-in procedure in two statistical software, SAS and R. Since SAS often gives a more detailed output, we selected it to analyze our data. After that, we had to find the correct procedure (GENMOD) to fit a general linear model to our data. We watched various tutorials to understand how to correctly use the procedure to get the desired output.

This was not the biggest problem though. Once we were able to use SAS, we had to learn how SAS operated so that we did not lose data due to incorrect input. We spent a lot of time re-entering data into SAS to correctly learn why some of our data was rejected. We learned a few things from this. One, SAS can only take so many entries with different inputs. Since two out of four factors we had are categorical, we made the mistake of putting all of the services and greediness into SAS separately, without groups. This error resulted in SAS taking over an hour to run, which we knew should not take that long. To fix this, we grouped each continuous factor to make them categorical factors. The second lesson we learned is that SAS outputs only eight characters. We had a few provider names that were similar in name, yet were completely different titles. However, SAS did not notice this and labeled them under the same name. This created incorrect models, making us have to go back and rename each provider type with the eight character limit.

Due to all of these difficulties, we spent most of our time going back and correcting mistakes in SAS and renaming provider types. But, from these difficulties, we learned more on how GLMs operate in SAS and how to improve our project so that we can improve the correlation for each distribution.

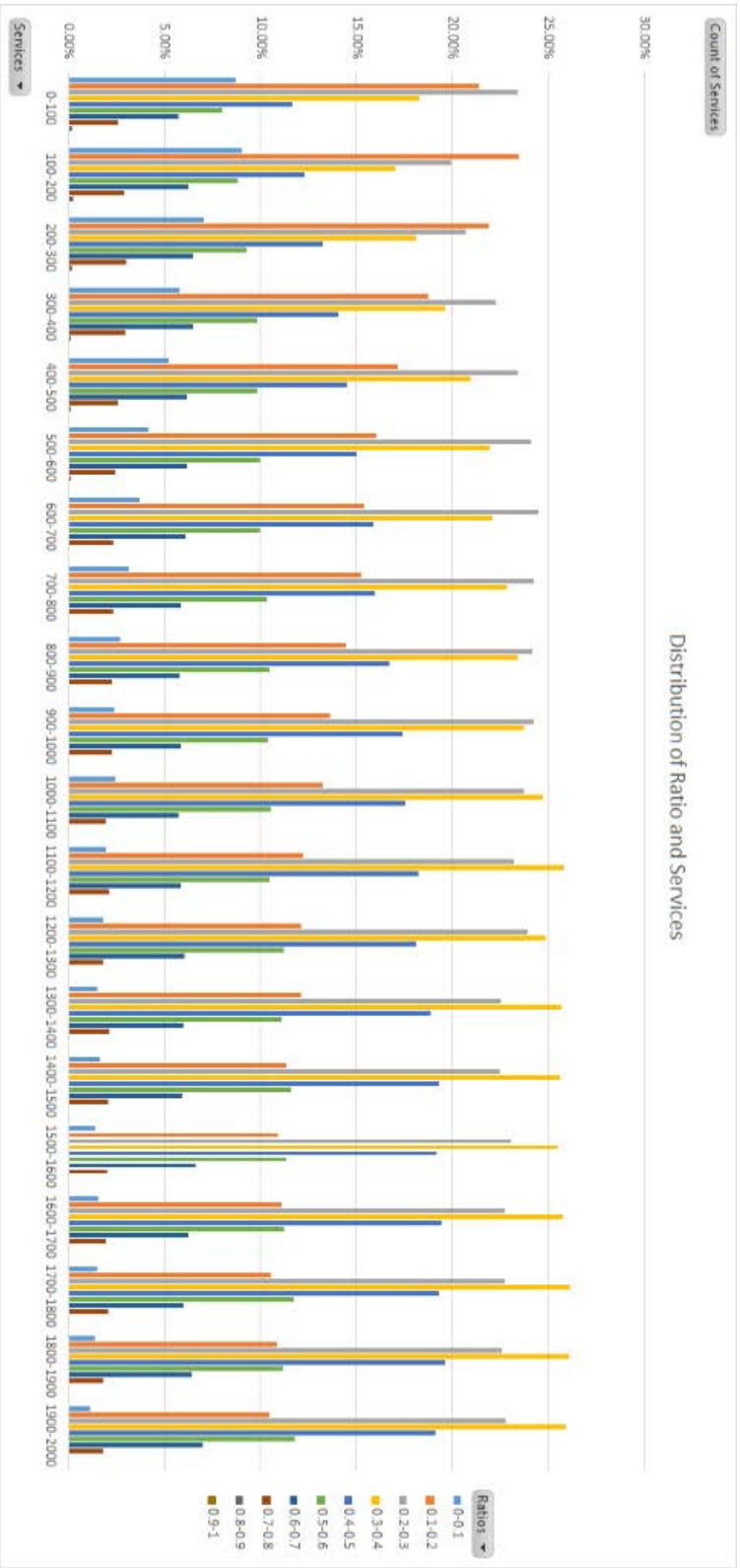
5.2.3. Feasible Improvements

We believe that being able to use provider qualifications would increase the predictive power of our model. Medical providers that belong to different industries usually require different compensation amounts. Therefore, it seems logical that providers with a higher level of qualification would usually require a higher amount of compensation than those with a lower level of qualification. Since working in a specific industry had the most impact on the RR, we believe that providing provider qualifications would further increase the correlation between fitted and actual values of the RR.

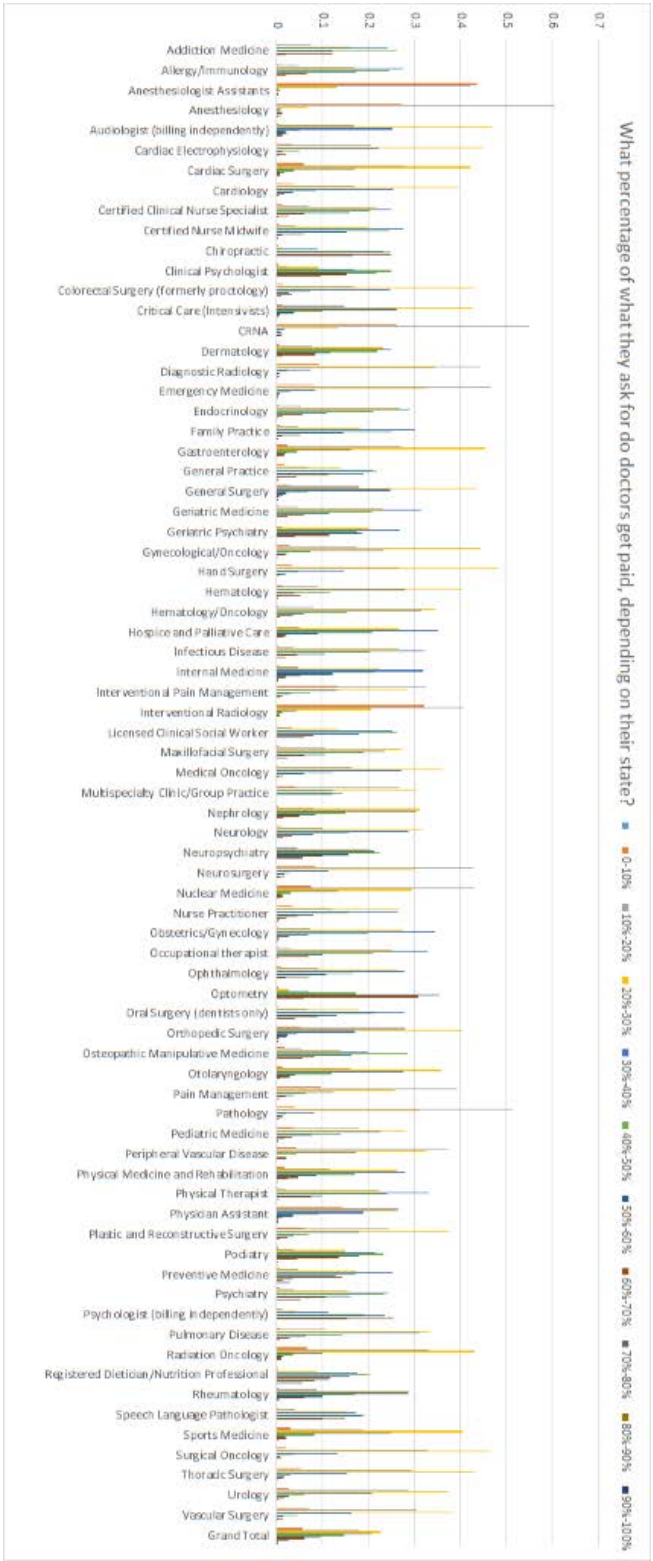
We learned from this project that so far only provider type and greediness were significant enough to impact the correlation of the linear model. Another improvement would be to find more factors, including provider qualifications, which could increase the fitted RR. By having at least four or five significant factors, the model could represent the data better.

Appendix A: Methodology Graphs

Distribution of Reimbursement Ratio and Number of Services



Payment Ratio Depending on State



Appendix B: SAS Code for GLM Procedures

```
***CODE FOR BACKGROUND***

data claims;
input GENDER $ TYPE $  AGE $ YIELD;
datalines;
male      luxury     old   1400
male      luxury     new   1550
male      regular    old   1080
male      regular    new   1230
female    luxury     old   1420
female    luxury     new   1570
female    regular    old   1100
female    regular    new   1250
;

proc genmod data=claims;
class GENDER TYPE AGE;
model YIELD = GENDER TYPE AGE / dist=Normal
                                link=identity
                                lrci;
output out=d predicted=Fitted;
run;

proc print data=d;
var Fitted;
run;
```

```
***CODE FOR GLM PROCEDURE***
data ratio;
input STATE $ PROVIDER $ SERVICES $ GREED $ RATIO;
datalines;


***insert large amount of data***


proc genmod data=ratio;
class STATE PROVIDER SERVICES GREED;
model RATIO = STATE PROVIDER SERVICES GREED / dist=Normal
                                              link=identity
                                              lrci;
output out=d predicted=Fitted;
run;

proc print data=d;
var Fitted;
run;
```

Appendix C: Level Estimates for Four Factors

States

State	Estimate	Std. Error	Confidence Interval Lower Bound	Confidence Interval Upper Bound	State	Estimate	Std. Error	Confidence Interval Lower Bound	Confidence Interval Upper Bound
AL	-0.0023	0.0039	-0.0098	0.0053	NC	-0.0021	0.0037	-0.0093	0.005
AR	-0.022	0.004	-0.0299	-0.0141	ND	-0.0289	0.0044	-0.0374	-0.0203
AZ	0.0265	0.0037	0.0192	0.0338	NE	-0.0244	0.004	-0.0323	-0.0165
CA	0.0227	0.0036	0.0157	0.0298	NH	0.0025	0.0041	-0.0054	0.0105
CO	0.0162	0.0038	0.0088	0.0236	NJ	0.0432	0.0037	0.036	0.0504
CT	0.0102	0.0038	0.0028	0.0176	NM	-0.0044	0.0041	-0.0124	0.0035
DC	0.0249	0.0044	0.0162	0.0335	NV	0.0088	0.0041	0.0008	0.0169
DE	0.0219	0.0045	0.0131	0.0307	NY	0.0245	0.0036	0.0175	0.0316
FL	0.0274	0.0036	0.0203	0.0345	OH	0.0061	0.0036	-0.001	0.0133
GA	-0.006	0.0037	-0.0133	0.0012	OK	-0.0089	0.0039	-0.0165	-0.0013
IA	-0.0275	0.0039	-0.0351	-0.0199	OR	-0.0294	0.0038	-0.0368	-0.0219
ID	0.0098	0.0042	0.0016	0.0179	PA	0.0178	0.0036	0.0107	0.0249
IL	0.0095	0.0036	0.0024	0.0167	RI	0.0245	0.0042	0.0163	0.0326
IN	-0.0027	0.0037	-0.01	0.0046	SC	-0.014	0.0038	-0.0214	-0.0065
KS	-0.0005	0.0039	-0.0081	0.0072	SD	0.0233	0.0044	0.0148	0.0318
KY	-0.0155	0.0038	-0.0229	-0.008	TN	-0.0153	0.0037	-0.0226	-0.008
LA	-0.0097	0.0038	-0.0172	-0.0022	TX	0.0001	0.0036	-0.007	0.0071
MA	-0.0071	0.0037	-0.0142	0.0001	UT	0.0009	0.004	-0.0069	0.0087
MD	0.0364	0.0037	0.0291	0.0437	VA	0.0078	0.0037	0.0005	0.015
ME	0.0103	0.004	0.0025	0.0182	VT	-0.0006	0.0045	-0.0094	0.0082
MI	0.0223	0.0036	0.0152	0.0295	WA	-0.0132	0.0037	-0.0204	-0.006
MN	-0.0109	0.0037	-0.0181	-0.0037	WI	-0.0467	0.0037	-0.054	-0.0395
MO	-0.0036	0.0037	-0.0109	0.0037	WV	-0.0207	0.0041	-0.0287	-0.0127
MS	-0.0283	0.004	-0.0362	-0.0203	WY	0	0	0	0
MT	0.022	0.0043	0.0136	0.0304					

Providers

Provider	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Addiction Medicine	0.045	0.014	0.0175	0.0725
Allergy Immunology	0.1478	0.0048	0.1384	0.1572
Anesthesiologist	-0.0779	0.0031	-0.0841	-0.0718
Audiologist	0.0689	0.0035	0.0619	0.0758
Certified Nurse Clinical Specialist	0.1519	0.0043	0.1434	0.1604
Nurse Midwife	0.1577	0.006	0.146	0.1694
CRNA	-0.0709	0.0032	-0.0771	-0.0647
Cardiac Electrophysiology	0.0353	0.0067	0.0221	0.0485
Cardiac Surgery	0.0001	0.0047	-0.009	0.0092
Cardiology	0.0345	0.0034	0.0278	0.0412
Chiropractic	0.3467	0.0032	0.3406	0.3529
Clinical Psychologist	0.2646	0.0033	0.2581	0.271
Colorectal Surgery	0.0578	0.0049	0.0482	0.0673
Critical Care	0.0614	0.0042	0.0532	0.0695
Dermatology	0.129	0.0037	0.1217	0.1363
Diagnostic Radiology	-0.065	0.0034	-0.0716	-0.0584
Emergency Medicine	-0.0249	0.0031	-0.0311	-0.0188
Endocrinology	0.1524	0.0038	0.145	0.1597
Family Practice	0.1578	0.0031	0.1517	0.1639
Gastroenterology	0.0291	0.0033	0.0226	0.0356
General Practice	0.1516	0.0037	0.1443	0.1589
General Surgery	0.0322	0.0032	0.0259	0.0385
Geriatric Medicine	0.1644	0.0048	0.155	0.1738
Geriatric Psychiatry	0.1898	0.0116	0.1671	0.2125
Gynecological Oncology	0.0371	0.0058	0.0257	0.0485
Hand Surgery	0.011	0.0054	0.0004	0.0216
Hematology Oncology	0.1174	0.0038	0.1099	0.1249
Hematology	0.1053	0.0066	0.0923	0.1183
Hospice Palliative Care	0.1469	0.0065	0.1341	0.1597
Interventional Pain Management	-0.0135	0.006	-0.0253	-0.0016
Internal Medicine	0.1527	0.0031	0.1466	0.1587
Interventional Radiology	-0.1065	0.0057	-0.1178	-0.0953
Infectious Disease	0.1547	0.0038	0.1472	0.1621
Licensed Clinical Social Worker	0.2167	0.0032	0.2104	0.2231
Maxillofacial Surgery	0.1213	0.0052	0.1111	0.1315
Medical Oncology	0.09	0.0045	0.0812	0.0989
Multispecialty Clinic	0.0342	0.02	-0.005	0.0734

			95% Confidence Interval	
Provider	**Estimate**	**Std. Error**	**Lower Bound**	**Upper Bound**
Nephrology	0.1269	0.0038	0.1194	0.1344
Neuropsychiatry	0.1789	0.0137	0.152	0.2058
Neurology	0.106	0.0034	0.0994	0.1126
Neurosurgery	-0.0262	0.0037	-0.0334	-0.0189
Nuclear Medicine	-0.042	0.0075	-0.0567	-0.0272
Nurse Practitioner	0.0892	0.0031	0.0831	0.0953
Obstetrics Gynecology	0.1081	0.0032	0.1019	0.1144
Occupational Therapist	0.1596	0.0039	0.1521	0.1672
Ophthalmology	0.1259	0.0034	0.1192	0.1325
Optometry	0.3321	0.0032	0.3258	0.3383
Oral Surgery	0.1462	0.0047	0.1371	0.1553
Orthopedic Surgery	-0.0106	0.0033	-0.017	-0.0042
Osteopathic Manipulative Medicine	0.1656	0.0067	0.1526	0.1787
Otolaryngology	0.0564	0.0035	0.0495	0.0632
Physical Medicine and Rehabilitation	0.1171	0.0037	0.1099	0.1243
Physical Therapist	0.1757	0.0032	0.1694	0.1819
Pain Management	-0.0089	0.0054	-0.0196	0.0018
Pathology	-0.008	0.0034	-0.0147	-0.0012
Pediatrician	0.0669	0.0044	0.0583	0.0755
Peripheral Vascular Disease	-0.0157	0.0196	-0.0541	0.0227
Physician Assistant	-0.0399	0.0031	-0.0461	-0.0338
Plastic Surgery	0.012	0.0037	0.0048	0.0193
Podiatry	0.2105	0.0034	0.2038	0.2172
Preventive Medicine	0.1993	0.0084	0.1829	0.2158
Psychiatry	0.2028	0.0032	0.1965	0.2091
Psychologist	0.3252	0.0089	0.3078	0.3427
Pulmonary Disease	0.1005	0.0036	0.0935	0.1075
Radiation Oncology	0.0028	0.0042	-0.0054	0.0111
Registered Dietician	0.3305	0.0044	0.3218	0.3393
Rheumatology	0.1316	0.0043	0.1231	0.1401
Speech Language Pathologist	0.2274	0.0067	0.2142	0.2406
Sports Medicine	0.0293	0.0068	0.016	0.0426
Surgical Oncology	-0.0059	0.0057	-0.0171	0.0054
Thoracic Surgery	-0.0049	0.0043	-0.0133	0.0035
Urology	0.0144	0.0038	0.0069	0.0219
Vascular Surgery	0	0	0	0

Services

Services	Estimate	Std. Error	95% Confidence Interval		Services	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound				Lower Bound	Upper Bound
"1-50"	0.0016	0.0014	-0.0011	0.0043	"1000-1050"	-0.0013	0.0018	-0.0048	0.0022
"51-100"	0.0077	0.0014	0.005	0.0104	"1050-1100"	0.0034	0.0018	-0.0002	0.007
"100-150"	0.0087	0.0014	0.006	0.0114	"1100-1150"	0.0016	0.0019	-0.0021	0.0052
"150-200"	0.0085	0.0014	0.0058	0.0113	"1150-1200"	-0.0008	0.0019	-0.0045	0.0029
"200-250"	0.0074	0.0014	0.0046	0.0102	"1200-1250"	0	0.0019	-0.0037	0.0037
"250-300"	0.0072	0.0014	0.0044	0.0101	"1250-1300"	0.0019	0.0019	-0.0019	0.0057
"300-350"	0.0062	0.0015	0.0034	0.0091	"1300-1350"	0.0011	0.002	-0.0027	0.005
"350-400"	0.0056	0.0015	0.0027	0.0085	"1350-1400"	-0.0008	0.002	-0.0047	0.0032
"400-450"	0.0046	0.0015	0.0016	0.0076	"1400-1450"	-0.0006	0.002	-0.0046	0.0034
"450-500"	0.0029	0.0015	-0.0002	0.0059	"1450-1500"	0.0009	0.002	-0.0031	0.0049
"500-550"	0.0032	0.0016	0.0001	0.0062	"1500-1550"	0.0011	0.0021	-0.003	0.0052
"550-600"	0.0023	0.0016	-0.0008	0.0054	"1550-1600"	-0.0011	0.0021	-0.0053	0.003
"600-650"	0.0021	0.0016	-0.001	0.0053	"1600-1650"	-0.0013	0.0021	-0.0055	0.003
"650-700"	0.0025	0.0016	-0.0007	0.0057	"1650-1700"	-0.0012	0.0022	-0.0055	0.0031
"700-750"	0.0005	0.0017	-0.0028	0.0037	"1700-1750"	-0.0012	0.0022	-0.0056	0.0031
"750-800"	0.0029	0.0017	-0.0004	0.0062	"1750-1800"	-0.0015	0.0023	-0.0059	0.0029
"800-850"	0.0012	0.0017	-0.0022	0.0045	"1800-1850"	-0.0075	0.0023	-0.012	-0.003
"850-900"	0.0021	0.0017	-0.0013	0.0055	"1850-1900"	-0.0017	0.0023	-0.0062	0.0027
"900-950"	0.0014	0.0017	-0.002	0.0049	"1900-1950"	-0.0045	0.0023	-0.009	0.0001
"950-1000"	0	0	0	0	"1950-2000"	-0.0046	0.0024	-0.0093	0

Greediness

			95% Confidence Interval	
Greediness	Estimate	Std. Error	Lower Bound	Upper Bound
"0-10"	0.1792	0.0021	0.1751	0.1834
"10-20"	0.1993	0.0012	0.197	0.2017
"20-30"	0.1829	0.001	0.181	0.1849
"30-40"	0.1359	0.0009	0.1341	0.1377
"40-50"	0.1136	0.0008	0.1119	0.1152
"50-60"	0.0942	0.0008	0.0926	0.0957
"60-70"	0.0695	0.0008	0.068	0.0709
"70-80"	0.046	0.0007	0.0446	0.0475
"80-90"	0.0219	0.0008	0.0204	0.0234
"90-100"	0	0	0	0
"100-110"	-0.0223	0.0008	-0.0239	-0.0207
"110-120"	-0.0406	0.0009	-0.0423	-0.0389
"120-130"	-0.0577	0.0009	-0.0596	-0.0559
"130-140"	-0.0714	0.001	-0.0734	-0.0694
"140-150"	-0.0811	0.0011	-0.0832	-0.0789
"150-160"	-0.088	0.0012	-0.0904	-0.0856
"160-170"	-0.096	0.0014	-0.0986	-0.0933
"170-180"	-0.0969	0.0015	-0.0998	-0.0939
"180-190"	-0.1053	0.0016	-0.1085	-0.1021
"190-200"	-0.1072	0.0018	-0.1107	-0.1037
"200-210"	-0.1128	0.0019	-0.1166	-0.109
"210-220"	-0.1141	0.0021	-0.1183	-0.11
"220-230"	-0.1191	0.0023	-0.1236	-0.1146
"230-240"	-0.1192	0.0025	-0.1241	-0.1142
"240-250"	-0.1234	0.0026	-0.1286	-0.1182
"250-260"	-0.1329	0.0029	-0.1386	-0.1272
"260-270"	-0.1312	0.0031	-0.1374	-0.1251
"270-280"	-0.1335	0.0032	-0.1398	-0.1271
"280-290"	-0.1356	0.0036	-0.1425	-0.1286
"290-300"	-0.143	0.0039	-0.1506	-0.1354
"300+"	-0.1715	0.0012	-0.1739	-0.1691

Appendix D: Greediness Ranking Levels

They ask for ___ of the industry average (greed)	Our SAS fitted RR + Intercept	They get ___ of the industry average submitted charge
5%	0.386	1.93%
15%	0.4061	6.09%
25%	0.3897	9.74%
35%	0.3427	11.99%
45%	0.3204	14.42%
55%	0.301	16.56%
65%	0.2763	17.96%
75%	0.2528	18.96%
85%	0.2287	19.44%
95%	0.2068	19.65%
105%	0.1845	19.37%
115%	0.1662	19.11%
125%	0.1491	18.64%
135%	0.1354	18.28%
145%	0.1257	18.23%
155%	0.1188	18.41%
165%	0.1108	18.28%
175%	0.1099	19.23%
185%	0.1015	18.78%
195%	0.0996	19.42%
205%	0.094	19.27%
215%	0.0927	19.93%
225%	0.0877	19.73%
235%	0.0876	20.59%
245%	0.0834	20.43%
255%	0.0739	18.84%
265%	0.0756	20.03%
275%	0.0733	20.16%
285%	0.0712	20.29%
295%	0.0638	18.82%

Ranking of the best percentages to ask for	
235%	20.59%
245%	20.43%
285%	20.29%
275%	20.16%
265%	20.03%
215%	19.93%
225%	19.73%
95%	19.65%
85%	19.44%
195%	19.42%
105%	19.37%
205%	19.27%
175%	19.23%
115%	19.11%
75%	18.96%
255%	18.84%
295%	18.82%
185%	18.78%
125%	18.64%
155%	18.41%
165%	18.28%
135%	18.28%
145%	18.23%
65%	17.96%
55%	16.56%
45%	14.42%
35%	11.99%
25%	9.74%
15%	6.09%
5%	1.93%

References

Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., & Thandi, N. (2007, February). *A Practitioner's Guide to Generalized Linear Models* (3rd ed.).

Agresti, A. (2013, January). *Categorical Data Analysis* (3rd ed.)

Feldblum, S., & Brosius, J., Dr. (2002, January). *The Minimum Bias Procedure*.

Mosley, R. (2005). The Use of Predictive Modeling in the insurance Industry. PINNACLE.

Werner, G., & Modlin, C. (2010, October). Basic ratemaking. In *Casualty Actuarial Society*.

"Medicare Physician and Other Supplier National Provider Identifier (NPI) Aggregate Report,Calendar Year 2014." *Data.CMS.gov*. N.p., 2015. Web. 25 Apr. 2017. https://data.cms.gov/Medicare-Physician-Supplier/Medicare-Physician-and-Other Supplier-National-Pro/4a3h-46r6