**SNF Deficiency Mitigation**

**The Problem**

Skilled Nursing Providers (SNFs) are highly regulated by the Centers for Medicare and Medicaid Services (CMS). Each SNF is surveyed by the Officer of Inspector General (OIG) each year during an annual survey and also following complaints made against the provider. If the OIG agents determine that the provider is out of compliance on any CMS regulation, then the provider can be tagged with deficiencies on the final report. The provider can be charged civil money penalties (CMP) for each deficiency ranging from $50 to $10,000 per day out of compliance. Large CMPs can be very damaging to a SNF company if they are not prepared for the financial impact.

**Data**

The first step taken was to extract the publicly reported data from the Medicare website at <https://data.medicare.gov/data/nursing-home-compare> and compile the data. The CSV flat files download multiple CSV formatted files in a zipped file with the most recently published data from Medicare. I chose to use only a few of these files.

The first file is labeled “HealthDeficiencies\_Download.csv” and contains all of the health deficiencies that each provider received for the past three year with the provider number, provider address, date of the survey, deficiency identification number, deficiency description, scope or severity of the deficiency, date that the deficiency was corrected, the cycle period with 1 being the most recent 12 month lookback, whether the deficiency was tagged on a standard or complaint survey and the date that the data was published to the Medicare website. The second file is labeled “FireSafetyDeficiencies\_Download.csv” and contains all of the fire safety deficiencies that each provider received for the past three years with all of the same categories as the health deficiencies file. The third file is labeled “Penalties\_Download.csv” and contains all of the CMPs for the past three years with the provider number, the provider name, provider address, date the penalty was received and the CMP fine amount. The last file from the CSV flat files is labeled “ProviderInfo\_Download.csv” has the provider information for each provider at the time the report is published including provider number, provider name, address, ownership type, number of certified beds, current overall star rating, current survey star rating, current quality star rating and current staffing star rating. Since the provider info file only contains the information for the most recent year and the deficiencies files include the past three years, I downloaded the provider information files for 2018 and 2017 from the archive of the Medicare website. I also created a CSV file that has each US state and the CMS regional location that it’s associated to for grouping.

**Data Wrangling**

The next step was to review all the data for missing data and combine all the data frames. I first searched the data for any blanks using excel filters and pivot tables to summarize the data. Then I uploaded each file into R using the read.csv function and reviewed each data frame using the summary function. There wasn’t any missing data. I combined the data frames using the rename function to rename columns to match other data frames, the rbind function to combine data frames by rows, the data.frame function to create new data frames and left\_join to join tables by vectors. Below is the code that I wrote to combine the data frames.

Health\_Deficiencies <- rbind(Health\_Deficiencies, Fire\_Safety\_Deficiencies)

Penalties <- Penalties %>% rename(PROVNUM = provnum)

Health\_Providers\_All <- rbind(Health\_Providers\_2019, Health\_Providers\_2018, Health\_Providers\_2017)

In addition to combining the data, I also grouped together two categories. I combine the deficiencies scope category to Substandard Care or Not Substandard Care. I combine the ownership type category to For Profit, Government or Non Profit using the code below.

Health\_Providers\_All$Own\_Type <- ifelse(Health\_Providers\_All$OWNERSHIP == "For profit - Corporation"| Health\_Providers\_All$OWNERSHIP == "For profit - Individual" | Health\_Providers\_All$OWNERSHIP == "For profit - Limited Liability company" | Health\_Providers\_All$OWNERSHIP == "For profit - Partnership", "For Profit", ifelse(Health\_Providers\_All$OWNERSHIP == "Government - City"| Health\_Providers\_All$OWNERSHIP == "Government - City/county" | Health\_Providers\_All$OWNERSHIP == "Government - County" | Health\_Providers\_All$OWNERSHIP == "Government - Federal" | Health\_Providers\_All$OWNERSHIP == "Government - Hospital district" | Health\_Providers\_All$OWNERSHIP == "Government - State", "Government", ifelse(Health\_Providers\_All$OWNERSHIP == "Non profit - Church related"| Health\_Providers\_All$OWNERSHIP == "Non profit - Corporation" | Health\_Providers\_All$OWNERSHIP == "Non profit - Other", "Non Profit", "N/A")))

**Method**

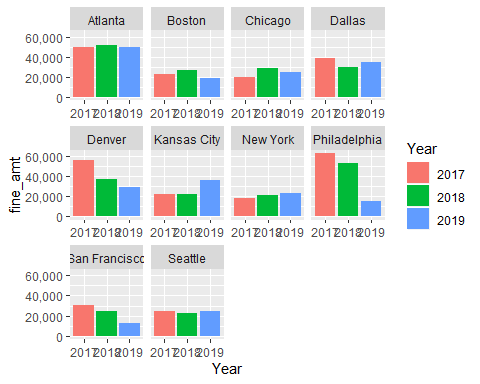
Once all of the data was combined, I started creating linear regression models with the CMP fine amount as the dependent variable. The first model I created used quality star rating, staffing star rating, overall star rating, number of not substandard care deficiencies received, number of substandard care deficiencies received and the number of licensed beds as the coefficients. The summary of this model showed that each of the coefficients were very significant with each Pr(>|t|) level close to zero, but the adjusted r-squared value was only .02521, which indicated that the model was not a good fit.

After creating the first model, I realized that there are a lot of providers that didn’t receive any deficiencies or CMP fines that could be skewing the model. I created a new data frame without providers that didn’t have any deficiencies or fines. I also wanted to normalize the data to minimize the impact of outliers, so I added a logarithmic transformation to the intercept in the second model. I also added year, ownership type, quality star rating and staffing star rating to the coefficients. The summary of the second model showed that the adjusted r-squared value increased to .06175. However, the ownership type, quality star rating and staffing star rating were not very significant.

For the third model, I replaced the ownership type with the grouped ownership type categories to narrow it down from thirteen coefficients to three. I also removed the quality star rating and the staffing star rating. I added the CMS regional location as a coefficient, since the regional offices could have a big influence on the fine amount decision. I also reviewed the number of licensed beds and found that there are a few outliers. There are a few providers that have over 750 beds and one with 1,389, while most providers have less than 200 beds. Since there some outliers, I added a logarithmic transformation to the number of licensed beds coefficient. The diagnostics for this model are located in appendix A.

**Findings**

After establishing that the final linear model is sound, I created a few graphs to help explain the data. I first created a bar graph using of fine amount against year and faceted by CMS regional location. I then realized that each region has a different number of providers, so I created the same graph, except calculating the average fine amount per provider. Below is the code and graph.

ggplot(data=subset(DF\_With\_Fines, !is.na(STATE))) +  
 stat\_summary(mapping = aes(x = Year, y = fine\_amt, fill = Year), fun.y = "mean", geom = "bar", na.rm = TRUE, inherit.aes = FALSE) +  
 facet\_wrap(. ~ Region, scales = "free\_x") +  
 scale\_y\_continuous(labels = comma)

The regions of Atlanta, Denver and Philadelphia show that they give out higher fines per provider, which is also indicated in the linear regression model.

**Further Research**

If I were to analyze the data further, there are a few items I would look into. Since I excluded all of the providers that did not receive a deficiency or a CMP fine, I would like to create a more comprehensive model that includes all providers. I would see if there’s a correlation between deficiency type and scope. If there’s a pattern between deficiency type and scope, then a SNF could focus on correcting behavior at the SNF to reduce the chance of receiving deficiencies that are associated with a more severe scope and mitigate higher CMPs. I would also look at if there’s a correlation between receiving the same deficiency year over year. If there is, then a SNF could focus on deficiencies that it received on its last survey. I would also look at the acuity of the patients to see if there’s correlation between a specific diagnosis or treatment type that may increase the chance of receiving deficiencies.

**Conclusion**

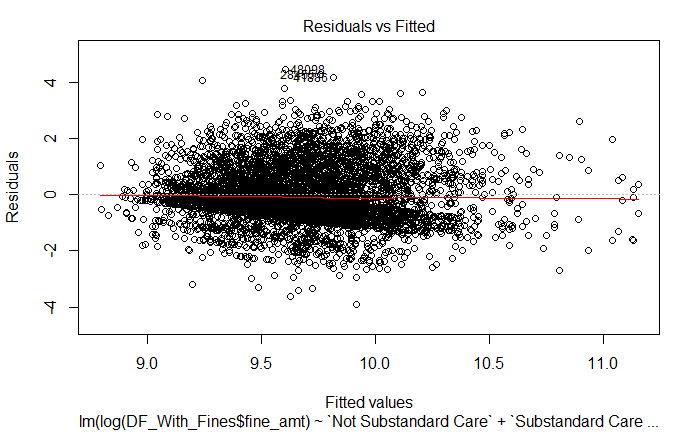
Based off the data analysis, I would make a few recommendations. Since the CMS region locations of Atlanta, Denver and Philadelphia assess higher penalties per provider, then I would recommend to either sell or not purchase SNFs that fall into the Atlanta, Denver or Philadelphia CMS region locations. Since the fine amount increases along with the number of licensed beds, then I would also recommend to purchase SNFs with a smaller number of licensed beds. Finally, I would recommend that each provider put an audit process in place, such as a mock survey, in an attempt to reduce or eliminate future deficiencies, since the fine amount increases with each deficiency received.

**Appendix A**

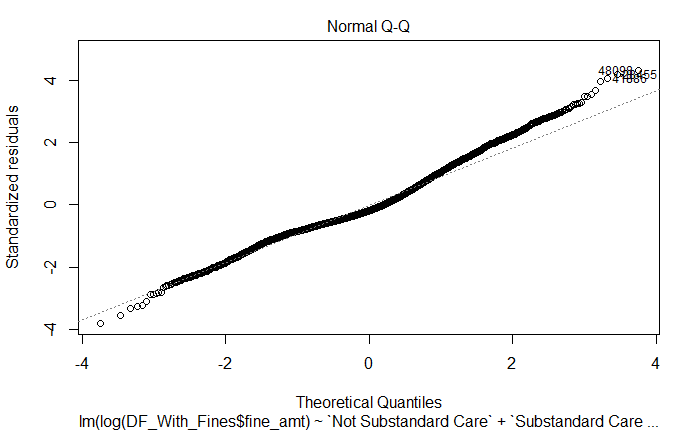
**Diagnostic of Linear Regression Model 3**

The summary of the third model showed an adjusted r-squared value of .08707 and that each of the coefficients were significant. Each coefficient’s estimate made sense, such as fine amount increasing along with the number of not substandard care deficiencies received, number of substandard care deficiencies received or the number of licensed beds.

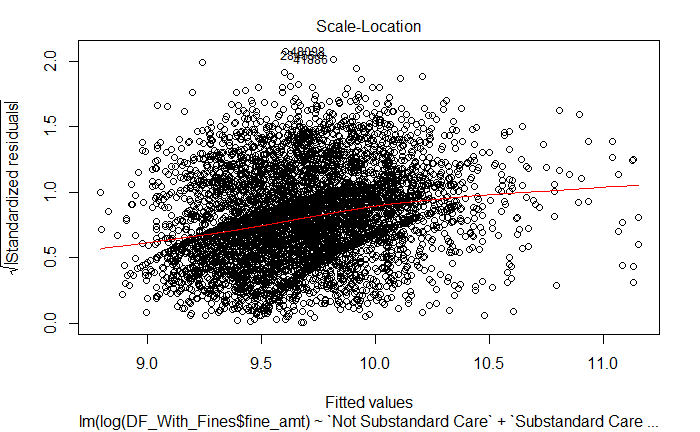
I analyzed the third model further by plotting it out. The first graph below shows the residuals vs the fitted values. As you can see in the graph, the majority of the plots are fairly equally spread along the horizontal line with few outliers, which shows that the model doesn’t have any non-linear relationships.



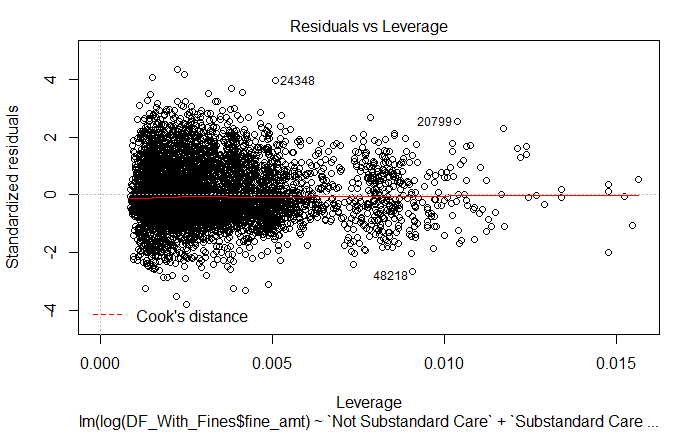
The Normal Q-Q graph shows the standardized residuals against the theoretical quantiles. The majority of the points follow the dotted line. This indicates that the residuals are normally distributed.



The Scale-Location graph shows if residuals are spread equally. The fairly horizontal line and evenly spread points indicates that that they are randomly spread.



The Residuals vs Leverage graph shows how influential an outlier is in the linear regression model. Most of the points below are close to the red line and none are outside of the Cook’s distance, which indicates that there aren’t any outliers influential enough to skew the regression model.



I also analyzed the correlation between the coefficients by calculating the variance inflation factor using the vif function in R. Each of the coefficients had a gvif that were close to 1.0, which indicates each coefficient doesn’t have a correlation with the other coefficients that could cause an issue.