**Jeffrey Fujioka**
VEE Statistics: Regression Analysis
Winter 2016 Student Project

# Medicare Advantage Penetration Rate and Fee-for-Service Risk Scores

## Introduction:

The purpose of this study is to analyze the relationship between the Medicare Advantage penetration rate in a state and the average FFS risk score. In particular, the study is testing whether a higher Medicare Advantage penetration rate will be connected to higher Fee-for-Service risk scores. To account for the fact that different regions can have a different relationship, dummy variables are introduced to code for the region in which a state is located.

## Definitions and Descriptions of Data:

Fee-for-Service (FFS) Medicare is traditional Medicare offered through the federal government. Instead of FFS Medicare, individuals can choose to enroll in a Medicare Advantage (MA) plan which is a plan offered by an insurance company which covers the benefits covered by FFS Medicare and typically covers supplemental benefits as well. As part of the payment process for Medicare Advantage, the Centers for Medicare and Medicaid Services (CMS) introduced a risk score model which attributes a risk score to a members based on demographics such as age and gender, as well as certain diagnoses for chronic conditions. A higher average risk score for a group of individuals is meant to represent higher expected costs for these individuals. This risk score model can also be used for FFS Medicare individuals, and average risk scores for FFS individuals by county is publically available. This allows for a comparison of risk score among states. However, not all states have the same percentage of members in FFS vs MA. The MA penetration rate is the percentage of Medicare eligible individuals that have opted to enroll in an MA plan. Through running a regression analysis and examining p-values, we can determine if there is any relationship between Medicare Advantage penetration rate in a state and the average FFS risk score.

The data used for the FFS risk score came from the CMS website. For the purposes of the study, risk scores from 2011 are used, which are based on diagnoses collected from 2010 dates of service and demographic data such as age from 2011.  These risk scores are calculated under two CMS-HCC models and blended together to return one risk score. The risk scores provided are at the county level, and using the 2011 average enrollment between Part A and Part B, a weighted average for each state was calculated, as well as the District of Columbia. This provided a data set of 51 risk scores for 2011.

For the MA penetration rate, the CMS website was also used. To align with the risk scores, MA penetration by county for June 2011 was used. Similar to the risk score information, the overall penetration rate for each state and the District of Columbia was calculated. This was determined using the number of Medicare eligible individuals and those enrolled in an MA plan. This again provided a data set of 51 values, which were linked to the risk score data by state.

**Univariate Data Analysis and Transformations:**

The first analysis was performed on the FFS risk score data. This has a first quartile of 0.978, a median of 1.014, and a third quartile of 1.072. The median-hinge ratio of the data is 1.63, indicating a positive skew in the data. This positive skew can be corrected by the Box-Cox transformation, moving down the ladder of powers. Additionally, because the ratio of the smallest value to the largest value is relatively small, this indicates that subtracting a constant from all the values is helpful. This gives rise to a two variable Box-Cox Transformation. The following transformation is used:

$$r \to \frac{(r-0.7)^{-2}-1}{-2}$$

Using this new data, we now have a median-hinge ratio of 1.05, which indicates that the data is much more symmetric. This Box-Cox transformed FFS Risk Score will be represented by the variable $Y$.

A similar analysis can also be performed on the MA penetration data. In this case the median-hinge ratio for the original MA penetration data is a ratio of 1.51. This has a first quartile of 12.21%, median of 19.47%, and a third quartile of 30.46%.  Due to the positive skew, a transformation is needed. Because this data is bounded by 0 below and 1 above, the logit transformation is a reasonable consideration. This uses the following transformation:

$$p \to \ln\left(\frac{p}{1-p}\right)$$

After the logit transformation, the new data has a much more symmetric distribution with a ratio of 1.07. This logit transformed MA Penetration Rate will be represented by the variable $X_1$.

While examining the data for $X_1$, it is important to note that there is an outlier. This outlier is the transformed Alaska MA Penetration Rate. This transformed value lies between the inner and outer fences on the lower end. Calculating a hat value for this data point results in 0.35, which is significantly above the next highest hat value of 0.09. Using this observation as well as other factors that may cause Alaska to be different than the rest of the data set (i.e., low population density, physical distance from other states, etc.) in the following analysis this point will be ignored. This leaves the dataset with 50 data points.

**Analysis:**

Initial Regression:

The initial regression was run using the Box-Cox transformed FFS risk score data as the dependent variable and the logit transformed MA Penetration Rate as the explanatory variable. The results of the analysis are as follows:

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.203391 |
| R Square | 0.041368 |
| Adjusted R Square | 0.021396 |
| Standard Error | 2.536425 |
| Observations | 50 |

ANOVA

| | *Df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 13.32587 | 13.32587 | 2.07134 | 0.156579 |
| Residual | 48 | 308.8058 | 6.433454 | | |
| Total | 49 | 322.1316 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
|---|---|---|---|---|
| Intercept | -3.83114 | 0.800632 | -4.78514 | 0.000017 |
| X1 | 0.712174 | 0.494835 | 1.439215 | 0.156579 |

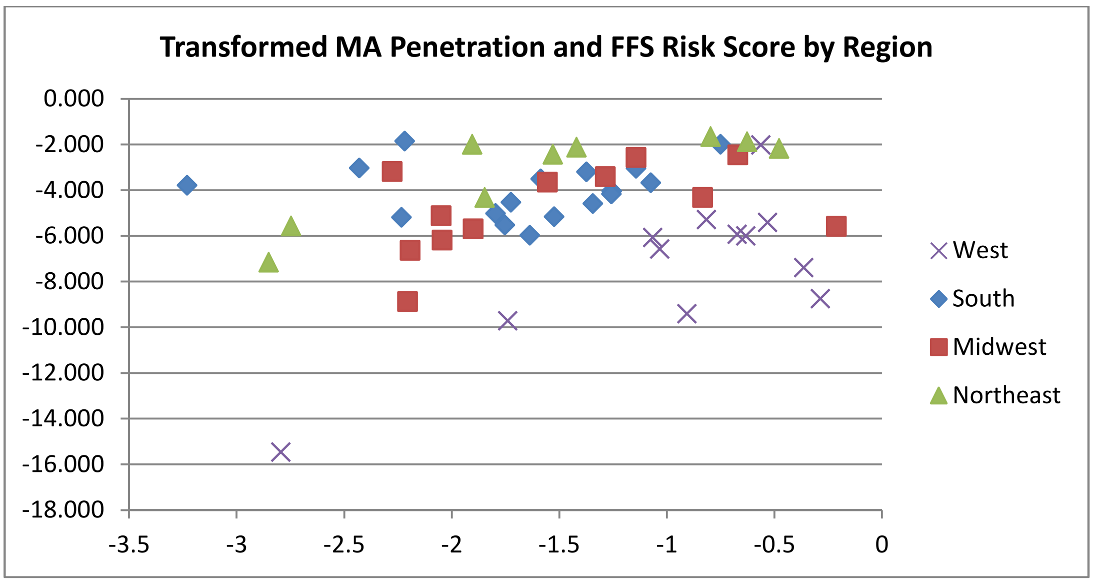
From this, the regression equation is:

$$\hat{Y} = -3.83114 + 0.712174X_1$$

where $\hat{Y}$ represents that Box-Cox transformed FFS risk score and $X_1$ represents the logit transformed MA penetration rate. Further, the analysis indicates that while there was some degree of correlation between the two variables, with $\rho = 0.203391$ and an $R$ Squared of 0.041368, there was a $p$-value of 0.1566. Using a null hypothesis of $\beta_1 = 0$, where $\beta_1$ represents the coefficient of $X_1$, this indicates that we cannot reject the null hypothesis with even $85\%$ confidence. Therefore, using a confidence level of $95\%$, the null hypothesis is not rejected and $\beta_1$ is assumed to be $0$.

Multivariate Scatterplot:

In order to determine whether or not regional variation would have an impact on the results, the states were separated into four regions. These four regions were established based on the Census Bureau's Regions of the US: West, Midwest, Northeast, and South. Prior to running a regression analysis, a scatterplot can be used to further analyze the data. A multivariable scatterplot was used, with the logit transformed MA Penetration Rate on the horizontal axis and the Box-Cox transformed FFS Risk Score on the vertical axis. The data points were colored to represent each region. Based on a visual inspection of the scatterplot, there is an indication that the regions could have an impact on the analysis.




Regression Analysis with Regions:

To run a regression analysis with a consideration for regions, three new dummy variables are introduced to represent the four regions. The first dummy variable, $D_1$ will be coded with a 1 when the state is in the Northeast and 0 otherwise. The second dummy variable, $D_2$ will be coded with a 1 when the state is in the Midwest and 0 otherwise. The third dummy variable, $D_3$ will be coded with a 1 when the state is in the South and 0 otherwise. Values of 0 for all three dummy variables will indicate that the state is in the West region.

After running the analysis, a new regression equation is obtained,

$$\hat{Y} = -5.6851 + 5.1701X_1 + 3.5363D_1 + 4.5602D_2 + 1.7325D_3$$

This regression equation is obtained based on the following results:

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.738647 |
| R Square | 0.5456 |
| Adjusted R Square | 0.505209 |
| Standard Error | 1.803556 |
| Observations | 50 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 4 | 175.755 | 43.93874 | 13.50791 | 2.6E-07 |
| Residual | 45 | 146.3767 | 3.252815 | | |
| Total | 49 | 322.1316 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
|---|---|---|---|---|
| Intercept | -5.6851 | 0.634968 | -8.95336 | 1.47E-11 |
| X1 | 5.170078 | 0.830647 | 6.22416 | 1.45E-07 |
| D1 – Northeast | 3.536328 | 0.76902 | 4.598488 | 3.46E-05 |
| D2 – Midwest | 4.560237 | 0.733062 | 6.220808 | 1.47E-07 |
| D3 – South | 1.732461 | 0.382296 | 4.531733 | 4.29E-05 |

Performing the new regression analysis using the four explanatory variables, the R squared improves to 54.56%. More importantly, the $p$-value for the transformed MA penetration rate has drastically reduced to $1.45 \times 10^{-7}$, indicating that the hypothesis that $\beta_1 = 0$ would be rejected at the 95% confidence level. In fact, the hypothesis could be rejected at a much higher confidence level, even above 99.99%. Additionally, the coefficients for the three dummy variables all have $p$-values less than 0.0001, indicating that there is above 99.99% confidence that the regional division will have an impact on the regression equation.

A comparison can be made between regions, based on the coefficients for $D_1$, $D_2$, and $D_3$. These coefficients can be added to the intercept to determine the intercept for each of the regions. For those in the West, the intercept will be -5.6851. Those in the Northeast will have an intercept of -2.14877, which is higher by 3.5363. The Midwest will have an intercept of -1.1249, higher than the West by 4.5602. In the South the intercept will be also higher, at -3.9526, which is 1.7325 above the West. This indicates that for a given level of MA penetration, the Midwest would have the highest risk score, followed by Northeast, then the South, and the West would have the lowest risk score.

**Conclusion:**

Based on this analysis, there is evidence that MA penetration rate can predict FFS Risk score in a state, once regional variation is considered. It is possible that this could be caused by the fact that members with lower risk scores (younger, less chronic conditions, etc.) choose to be in a MA plan. As a result, the higher percentage of members that are in an MA plan will leave fewer individuals in FFS and these fewer individuals will have higher risk scores. However, another possible explanation is that those states with higher risk scores in general may have a more intensive push for individuals to enter into MA plans, with perhaps the hopes that their health care will be monitored more closely. Further analysis could be performed to understand these possible causes, including an analysis on MA risk scores or an analysis on the FFS risk scores in one county/state as the MA Penetration Rate increases. Additionally, a further analysis can be done to see if any interaction exists between region and the MA penetration rate.