Capstone Project

Tim Harvey

5/26/2017

The goal of this project is to determine how the local uninsured rate influences the readmission rate of hospitals and whether good patient counseling, as judged by the patient, has any effect.

The audience for this analysis is both hospital administrators and those that have an interest in insurance policy. If an administrator finds the local population to have a high uninsured rate, the ability to mitigate this with additional efforts targeted toward discharge counseling could be a viable technique to lower readmission rates. The readmission rate is a key indicator of quality of care and healthcare facilities are penalized by Medicare for high readmission rates. These penalties result in reimbursements being withheld and in October of 2016 these penalties were increased by 20 percent. To increase the quality of patient outcomes, advocates for increased insurance coverage may be interested in the analysis as evidence that increases in insurance coverage lower readmission rates. Lower readmission rates have the effect of lowering overall healthcare costs.

The data used for this analysis will be the Small Area Health Insurance Estimates (SAHIE) by the US Census Bureau (https://www.census.gov/did/www/sahie/data/20082014/index.html), and the hospital-level HCAHPS and readmission data available from Medicare (https://data.medicare.gov/data/hospital-compare).

The first step is using the SAHIE data to determine the mean uninsured rate for each county in the United States. Once the mean uninsured rate of the population served in each county is determined it can then be correlated to the mean readmission rate for hospitals in that county. If there is a correlation, then the results of the HCAHPS survey can be used to determine if there is a correlation between discharge counseling and readmission rate among hospitals that have similar rates of uninsured patients.

The key pieces of data in the Medicare data sets are the state and county, the readmission rate for each facility, and the HCAHPS scores for the metric associated with patient education. In the SAHIE data the key data fields are the state and county as well as the estimated uninsured rate in each county.

One of the limits of the data SAHIE data is that is the uninsured rate is an estimate. Ideally, this data would be a more exact metric, but SAHIE estimates are an accepted measure of the uninsured rate. Another limitation of the data is the three data sets do not measure the same period. The SAHIE data follows calendar years, January – December, whereas the period measured by the Medicare readmission data is July –

June, and the HCAHPS date is April - March. This is addressed by using the HCAHPS data time period, and weighting the time periods of the other two data sets.

Another issue with the data is the granularity of the uninsured rate. The uninsured rate is aggregated at the county level while the data for readmission and HCAHPS scores are collected at the hospital level. Any correlation would be more meaningful if all data was at the hospital level, avoiding the need to aggregate hospital-level metrics at the county level and yielding less accurate results. Additionally, potential users of the analysis would be handicapped by the aggregation, as many of the changes that could affect the readmission rate would take place at the hospital level. A missing element of the data set that would be useful would be bins for age ranges. As it is, there is no way to take age into consideration and conclusions can't be drawn regarding the impacts on various age groups.

The first step in munging the SAHIE data is removing the explanatory information at the top of the file, stripping the trailing white space, filtering out the data that covers specific populations, rather than the population as a whole, and selecting only the columns that are necessary to either further cleanse the data or for the final analysis. In this case, the year will be used to determine the final metric weighting, the state and county will be used to join the other data sets, and the PCTELIG column is the uninsured rate for the county population.

```r
sahie_2014_clean <- read.csv(file.path("Data/", "sahie_2014.csv"), stringsAsFactors = FALSE,
    sep = ",", skip = 79, strip.white = TRUE, na.strings = c("")) %>% filter(agecat ==
    0, racecat == 0, sexcat == 0, iprcat == 0, county_name != "NA") %>% select(year,
    state_name, county_name, PCTELIG)
```

```r
sahie_2015_clean <- read.csv(file.path("Data/", "sahie_2015.csv"), stringsAsFactors = FALSE,
    sep = ",", skip = 79, strip.white = TRUE, na.strings = c("")) %>% filter(agecat ==
    0, racecat == 0, sexcat == 0, iprcat == 0, county_name != "NA") %>% select(year,
    state_name, county_name, as.numeric(PCTELIG))
```

The next step is formatting the data in each column. The uninsured rate needs to be converted to a numeric data type, and the state and county names do not match those in the other two sets. The county name needs to be capitalized and the word "County" and the word "Parish" needed to be removed from the name. The state name is given as the full name, for example, "Louisiana" and needs to be converted to the abbreviation "LA".

```
sahie_2014_clean$PCTELIG <- as.numeric(sahie_2014_clean$PCTELIG)
sahie_2014_clean$state_name <- as.factor(state.abb[match(sahie_2014_clean$state_name,
    state.name)])
sahie_2014_clean$county_name <- as.factor(toupper(sahie_2014_clean$county_name))
sahie_2014_clean$county_name <- as.factor(gsub(" COUNTY", "", sahie_2014_clean$county_name))
sahie_2014_clean$county_name <- as.factor(gsub(" PARISH", "", sahie_2014_clean$county_name))
```

```
sahie_2015_clean$PCTELIG <- as.numeric(sahie_2015_clean$PCTELIG)
sahie_2015_clean$state_name <- as.factor(state.abb[match(sahie_2015_clean$state_name,
    state.name)])
sahie_2015_clean$county_name <- as.factor(toupper(sahie_2015_clean$county_name))
sahie_2015_clean$county_name <- as.factor(gsub(" COUNTY", "", sahie_2015_clean$county_name))
sahie_2015_clean$county_name <- as.factor(gsub(" PARISH", "", sahie_2015_clean$county_name))
```

The final step in cleansing the SAHIE data is combining the two years of data, with appropriate weighting, and averaging the score for all records within a county and state. The final version of the data set is completed by changing the columns headers for the county and state name to match those in the other two data sets, as well as changing the "PCTELIG" column name to the more descriptive "combined_sahie_metric".

```
sahie_final <- inner_join(sahie_2014_clean, sahie_2015_clean, by = c(state_name = "state_name",
    county_name = "county_name")) %>% mutate(combined_sahie_metric = (PCTELIG.x *
    0.75) + (PCTELIG.y * 0.25)) %>% select(state_name, county_name, PCTELIG.x, PCTELIG.y,
    combined_sahie_metric) %>% group_by(State = state_name, County.Name = as.factor(county_name)) %>%
    summarise(county_unins_average = mean(combined_sahie_metric))
```

The HCAHPS data is the most straightforward to clean as it is the base timeline, so there is only one file. The file is read, eliminating white space and formatting NA values. Any hospital that did not fall in a county are eliminated, all measures other than the chosen metric are filtered out, and any hospital for which the data was not collected are filtered out.

```
hcahps_clean <- hcahps_201404_201503 <- read.csv(file.path("Data", "HCAHPS_Hospital_201404-201503.csv")
    stringsAsFactors = FALSE, sep = ",", strip.white = TRUE, na.strings = c("", "Not Available")) %>%
    filter(HCAHPS.Measure.ID == "H_COMP_6_Y_P", County.Name != "NA", HCAHPS.Answer.Percent !=
        "NA") %>% select(Measure.Start.Date, Measure.End.Date, State, County.Name,
    HCAHPS.Measure.ID, HCAHPS.Answer.Percent)
```

The date format does not include a leading zero for single-digit months and days. The leading zero is required for conversion to a date data type. Since there are only two dates they can be replaced with a simple gsub. The state and county names are converted to factors.

```r
hcahps_clean$Measure.End.Date <- as.Date(hcahps_clean$Measure.End.Date, "%m/%d/%Y")
hcahps_clean$Measure.Start.Date <- as.Date(hcahps_clean$Measure.Start.Date, "%m/%d/%Y")
hcahps_clean$State <- as.factor(hcahps_clean$State)
hcahps_clean$County.Name <- as.factor(hcahps_clean$County.Name)
```

The final HCAHPS data is generated by selecting the state name, county name, and the necessary metric. The data is then converted to a numeric data type and averaged across county and state yielding a mean score for each county. The name of the measure is changed to "county__discharge__counsel__avg" for clarity.

```r
hcahps_final <- select(hcahps_clean, State, County.Name, HCAHPS.Answer.Percent) %>%
    group_by(State, County.Name) %>% summarise(county_discharge_counsel_avg = mean(as.numeric(HCAHPS.Ans
```

The munging of the readmission data is very similar to the HCAHPS data, as they are from the same source and the same formatting conventions are used. The notable exception is there are two files that must be weighted to derive the final metric. The file is read eliminating white space and formatting NA values and the unnecessary scores and missing data are filtered out.

```r
readmit_201307_201406_clean <- read.csv(file.path("Data", "Readmissions_and_Deaths 201307-201406_Hospita
    stringsAsFactors = FALSE, sep = ",", strip.white = TRUE, na.strings = c("", "Not Available")) %>%
    filter(Measure.ID == "READM_30_HOSP_WIDE", County.Name != "NA", Score != "NA") %>%
    select(Measure.Start.Date, Measure.End.Date, State, County.Name, Measure.ID,
        Score)
```

```r
readmit_201407_201506_clean <- read.csv(file.path("Data", "Readmissions_and_Deaths 201407-201506_Hospita
    stringsAsFactors = FALSE, sep = ",", strip.white = TRUE, na.strings = c("", "Not Available")) %>%
    filter(Measure.ID == "READM_30_HOSP_WIDE", County.Name != "NA", Score != "NA") %>%
    select(Measure.Start.Date, Measure.End.Date, State, County.Name, Measure.ID,
        Score)
```

The dates are formatted in the same way as the HCAHPS file, so the same gsub is used. The state and county names are then converted to factors.

```
readmit_201307_201406_clean$Measure.Start.Date <- as.Date(gsub("7/1/13", "07/01/2013",
    readmit_201307_201406_clean$Measure.Start.Date), "%m/%d/%Y")
readmit_201307_201406_clean$Measure.End.Date <- as.Date(gsub("6/30/14", "06/30/2014",
    readmit_201307_201406_clean$Measure.End.Date), "%m/%d/%Y")
readmit_201307_201406_clean$State <- as.factor(readmit_201307_201406_clean$State)
readmit_201307_201406_clean$County.Name <- as.factor(readmit_201307_201406_clean$County.Name)
```

```
readmit_201407_201506_clean$Measure.Start.Date <- as.Date(gsub("7/1/14", "07/01/2014",
    readmit_201407_201506_clean$Measure.Start.Date), "%m/%d/%Y")
readmit_201407_201506_clean$Measure.End.Date <- as.Date(gsub("6/30/15", "06/30/2015",
    readmit_201407_201506_clean$Measure.End.Date), "%m/%d/%Y")
readmit_201407_201506_clean$State <- as.factor(readmit_201407_201506_clean$State)
readmit_201407_201506_clean$County.Name <- as.factor(readmit_201407_201506_clean$County.Name)
```

The final readmission data is derived by joining the two yearly files by county and state. The readmission rates are then weighted and combined into a final metric. The final columns are State, County.Name, the individual yearly scores, and the final combined metric. The mean of the combined metric per county is derived, yielding the final readmission file.

```
readmit_final <- inner_join(readmit_201307_201406_clean, readmit_201407_201506_clean,
    by = c(State = "State", County.Name = "County.Name")) %>% mutate(combined_readmit_metric = (Score.x
    0.25) + (Score.y * 0.75)) %>% select(State, County.Name, Score.x, Score.y, combined_readmit_metric)
    group_by(State = as.factor(State), County.Name = as.factor(County.Name)) %>%
    summarise(county_readmit_average = mean(combined_readmit_metric))
```

The final data set on which the analysis will be preformed is derived by using nested joins and the state and county names.

```
data_final <- inner_join(readmit_final, hcahps_final, by = c(State = "State", County.Name = "County.Nam
    inner_join(., sahie_final, by = c(State = "State", County.Name = "County.Name"))
```

This is some basic analysis to start.

```
lm1 <- lm(county_readmit_average ~ county_discharge_counsel_avg + county_unins_average,
    data = data_final)
summary(lm1)
```

```
##
## Call:
## lm(formula = county_readmit_average ~ county_discharge_counsel_avg +
##     county_unins_average, data = data_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9854 -0.3622 -0.0527  0.2989  3.4847
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  18.067582   0.274593  65.798   <2e-16 ***
## county_discharge_counsel_avg -0.030541   0.003036 -10.059   <2e-16 ***
## county_unins_average          0.001472   0.002761   0.533    0.594
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5892 on 2108 degrees of freedom
## Multiple R-squared:  0.04986,    Adjusted R-squared:  0.04896
## F-statistic: 55.31 on 2 and 2108 DF,  p-value: < 2.2e-16
```

```r
anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: county_readmit_average
##                                Df Sum Sq Mean Sq  F value Pr(>F)
## county_discharge_counsel_avg    1  38.30  38.302 110.3424 <2e-16 ***
## county_unins_average            1   0.10   0.099   0.2844 0.5939
## Residuals                    2108 731.72   0.347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```