Final Project - Medicare PUF

Arjun Srinivasan

March 5, 2019

Medicare PUF Analysis

Research Question

The cost of social security programs is a persistent political hot topic in the United States; as national debt rises, politicians and citizens continue to question whether programs like Medicare and Medicaid are affordable. The affordability of these programs is an undoubtedly complex issue, including a vast body of research on administrative costs, drug prices, and payments to medical providers. This investigation looks at Medicare payments to providers using the Basic Stand Alone (BSA) Medicare Claims Public Use Files (PUFs), a 2010 dataset that contains information from Medicare Carrier claims, to infer the relationship between a variety of claim-related factors-from patient sex and age group to diagnoses to type of provider-and the cost of any given claim in the dataset.

Prior research on Medicare PUFs has found beneficiaries diagnosed with cancer-even without any other chronic conditions-cost five times more than beneficiaries with no chronic conditions (Prada 2014). Another study used Medicare PUFs to look at factors related to Alzheimer's disease (Prada 2013). This project is one similarly focused on inference and description, as I am interested in understanding a variety of potential relationships between a variety of factors and the costs of services.

Research Design

To investigate this relationship, I will use a variety of different regressions-linear, ridge, and lasso-in addition to decision trees (pruned and unpruned). Before fitting the models, the dataset must undergo extensive modification in transforming all relevant categorical and multi-level factors into quantitative binary factors. Then, I will take two small samples (n = 20,000) from the large dataset (n = 70,052,393; for df1 loaded from .csv, n = 2,801,660) as training and test datasets.

Project Setup

```
## I. Load packages.
library(Matrix)
library(glmnet)
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-16
```

```
library(ggplot2)
library(rpart)
library(rpart.plot)
library(tree)

## II. Import data into data frame.
df1 <- read.csv("2010_BSA_Carrier_PUF.csv", header=TRUE, stringsAsFactors = FALSE)

## Rename headers.
names(df1) <- c("Sex", "Age_Group", "ICD9", "HCPCS_Service_Code", "BETOS_Service_Code", "Service_Count"
## Testing what dataset looks like.
## summary(df1)
## head(df1)

## III. In creating data frame 2 (df2), the data frame used for model fitting, I will begin with a Cost

## For ICD9, create 20 binary variables representing the 20 different categories of diagnoses as shown i

##Note: The first 45 observations have no ICD9.

Infectious_dz <- c(rep(0,45), rep(1,5681-45), rep(0,2801660-5681))
Neoplasms <- c(rep(0,5681), rep(1,22867-5681), rep(0,2801660-22867))
Immunity_dz <- c(rep(0,22867), rep(1,35186-22867), rep(0,2801660-35186))
Blood_dz <- c(rep(0,35186), rep(1,39893-35186), rep(0,2801660-39893))
Mental_disorder <- c(rep(0,39893), rep(1,53448-39893), rep(0,2801660-53448))
Nervous_dz <- c(rep(0,53448), rep(1,63342-53448), rep(0,2801660-63342))
Sense_dz <- c(rep(0,63342), rep(1,71939-63342), rep(0,2801660-71939))
Circulatory_dz <- c(rep(0,71939), rep(1,97895-71939), rep(0,2801660-97895))
Respiratory_dz <-c(rep(0,97895), rep(1,111513-97895), rep(0,2801660-111513))
Digestive_dz <-c(rep(0,111513), rep(1,127463-111513), rep(0,2801660-127463))
Genitourinary_dz <-c(rep(0,127463), rep(1,140839-127463), rep(0,2801660-140839))
Pregnancy_complic <-c(rep(0,140839), rep(1,140840-140839), rep(0,2801660-140840))
Skin_dz <-c(rep(0,140840), rep(1,150355-140840), rep(0,2801660-150355))
Musculoskeletal_dz <-c(rep(0,150355), rep(1,180065-150355), rep(0,2801660-180065))
Congenital_anom <-c(rep(0,180065), rep(1,181323-180065), rep(0,2801660-181323))
Perinatal_condit <-c(rep(0,181323), rep(1,181336-181323), rep(0,2801660-181336))
Ill_defined_condit <-c(rep(0,181336), rep(1,208063-181336), rep(0,2801660-208063))
Injury_Poison <-c(rep(0,208063), rep(1,228985-208063), rep(0,2801660-228985))
External_Cz_of_Inj <-c(rep(0,228985), rep(1,229008-228985), rep(0,2801660-229008))
Fact_Inf_Hlth_Srvc <-c(rep(0,229008), rep(1,2801660-229008))

df.ICD9 <- cbind.data.frame(Infectious_dz, Neoplasms, Immunity_dz, Blood_dz, Mental_disorder, Nervous_dz

df2 <- cbind.data.frame(df1$Cost, df1$Sex, df.ICD9)
colnames(df2)[1] <- "Cost"
colnames(df2)[2] <- "Sex"

## For HCPCS_Service_Code, create 7 binary variables for the 10 most frequent HCPCS as shown in the code

## 99213, 99214 = Established Patient Office or Other Outpatient Services
EPOOS <- rep(0,2801660)
## 36415 = Venous Procedures
Venous <- rep(0,2801660)
## 99232 = Subsequent Hospital Care
Hosp_Care <- rep(0,2801660)
## 85025, 85610 = Hematology and Coagulation Procedures
Hematology <- rep(0,2801660)
## 80053, 80061 = Organ or Disease Oriented Panels
Organ_Dz_Pnl <- rep(0,2801660)
## 97110 = Physical Medicine and Rehabilitation Therapeutic Procedures
Phys_Med_Rhb <- rep(0,2801660)
## The last category in the codebook is 'all other values'.
Other_Service <- rep(0,2801660)
```

```r
for(i in 1:2801660){
  if(df1[i,4] == '99213' || df1[i,4] == '99214'){
    EPOOS[i] <- 1
  }
  else if(df1[i,4] == '36415'){
    Venous[i] <- 1
  }
  else if(df1[i,4] == '99232'){
    Hosp_Care[i] <- 1
  }
  else if(df1[i,4] == '85025' || df1[i,4] == '85610'){
    Hematology[i] <- 1
  }
  else if(df1[i,4] == '80053' || df1[i,4] == '80061'){
    Organ_Dz_Pnl[i] <- 1
  }
  else if(df1[i,4] == '97110'){
    Phys_Med_Rhb[i] <- 1
  }
  else{
    Other_Service[i] <- 1
  }
}

df.HCPCS <- cbind.data.frame(EPOOS, Venous, Hosp_Care, Hematology, Organ_Dz_Pnl, Phys_Med_Rhb, Other_Se

df2 <- cbind.data.frame(df2, df.HCPCS)

## For BETOS_Service_Code, create 10 binary variables for the 10 most frequent BETOS codes.

## M1B = Office Visit
Ofc_Vst <- rep(0,2801660)
## T1H = Lab Test - Other/Non-Medicare Fee Schedule
LbTst_Othr <- rep(0,2801660)
## M2B = Hospital Visit - Subsequent
Hosp_Vst <- rep(0,2801660)
## P6C = Minor Procedures
Mnr_Pcdr <- rep(0,2801660)
## T1A = Lab Test - Venipuncture
LbTst_Vnpctr <- rep(0,2801660)
## T1B = Lab Test - Automated General Profiles
LbTst_AGP <- rep(0,2801660)
## I1A = Standard Imaging - Chest
Chest_Img <- rep(0,2801660)
## M5C = Specialist - Opthamology
Opthmlgy <- rep(0,2801660)
## T1D = Lab Test - Blood Count
LbTst_BldCt <- rep(0,2801660)
## T2A = Other Tests - Electrocardiograms
ECG <- rep(0,2801660)
## All other values
Other_BETOS <- rep(0,2801660)
```

```r
for(i in 1:2801660){
  if(df1[i,5] == 'M1B'){
    Ofc_Vst[i] <- 1
  }
  else if(df1[i,5] == 'T1H'){
    LbTst_Othr[i] <- 1
  }
  else if(df1[i,5] == 'M2B'){
    Hosp_Vst[i] <- 1
  }
  else if(df1[i,5] == 'P6C'){
    Mnr_Pcdr[i] <- 1
  }
  else if(df1[i,5] == 'T1A'){
    LbTst_Vnpctr[i] <- 1
  }
  else if(df1[i,5] == 'T1B'){
    LbTst_AGP[i] <- 1
  }
  else if(df1[i,5] == 'I1A'){
    Chest_Img[i] <- 1
  }
  else if(df1[i,5] == 'M5C'){
    Opthmlgy[i] <- 1
  }
  else if(df1[i,5] == 'T1D'){
    LbTst_BldCt[i] <- 1
  }
  else if(df1[i,5] == 'T2A'){
    ECG[i] <- 1
  }
  else{
    Other_BETOS[i] <- 1
  }
}

df.BETOS <- cbind.data.frame(Ofc_Vst, LbTst_Othr, Hosp_Vst, Mnr_Pcdr, LbTst_Vnpctr, LbTst_AGP, Chest_Img

df2 <- cbind.data.frame(df2, df.BETOS)

## For Provider type, create 5 binary variables for the 5 distinct types of providers in the PUF.

Clinic <- rep(0,2801660)
Solo <- rep(0,2801660)
Institutional <- rep(0,2801660)
Clinic_Mult_Specialties <- rep(0,2801660)
Other_Provider <- rep(0,2801660)

for(i in 1:2801660){
  if(df1[i,7] == 0){
    Clinic[i] <- 1
  }
  else if(df1[i,7] == 1){
    Solo[i] <- 1
  }
  else if(df1[i,7] == 3){
    Institutional[i] <- 1
  }
  else if(df1[i,7] == 5){
    Clinic_Mult_Specialties[i] <- 1
  }
  else{
    Other_Provider[i] <- 1
  }
}

df.Provider <- cbind.data.frame(Clinic, Solo, Institutional, Clinic_Mult_Specialties, Other_Provider)
df2 <- cbind.data.frame(df2, df.Provider)

## Add service count.

df2 <- cbind.data.frame(df2, df1$Service_Count)
colnames(df2)[46] <- "Service_Count"

## For Service code, create 20 binary variables for the 20 types of services in the PUF codebook.

Med_care <- rep(0,2801660)
Diag_lab <- rep(0,2801660)
Diag_radiol <- rep(0,2801660)
Surgery <- rep(0,2801660)
Flu_vacc <- rep(0,2801660)
Ambulance <- rep(0,2801660)
Outpatient_MH <- rep(0,2801660)
Vision <- rep(0,2801660)
Anesthesia <- rep(0,2801660)
Thrp_radiol <- rep(0,2801660)
Ambul_surg_cntr <- rep(0,2801660)
Hearing <- rep(0,2801660)
Asst_at_surg <- rep(0,2801660)
Other_med_itm <- rep(0,2801660)
Consultation <- rep(0,2801660)
Prosthtc_Orthtc <- rep(0,2801660)
Med_supply <- rep(0,2801660)
Imnsprsv_drg <- rep(0,2801660)
Kidney_dnr <- rep(0,2801660)
Whole_bld <- rep(0,2801660)

for(i in 1:2801660){
  if(df1[i,8] == '1'){
    Med_care[i] <- 1
  }
  else if(df1[i,8] == '5'){
    Diag_lab[i] <- 1
  }
  else if(df1[i,8] == '4'){
    Diag_radiol[i] <- 1
  }
  else if(df1[i,8] == '2'){
    Surgery[i] <- 1
  }
  else if(df1[i,8] == 'V'){
    Flu_vacc[i] <- 1
  }
  else if(df1[i,8] == 'D'){
    Ambulance[i] <- 1
  }
  else if(df1[i,8] == 'T'){
    Outpatient_MH[i] <- 1
  }
  else if(df1[i,8] == 'Q'){
    Vision[i] <- 1
  }
  else if(df1[i,8] == '7'){
    Anesthesia[i] <- 1
  }
  else if(df1[i,8] == '6'){
    Thrp_radiol[i] <- 1
  }
  else if(df1[i,8] == 'F'){
    Ambul_surg_cntr[i] <- 1
  }
  else if(df1[i,8] == 'K'){
    Hearing[i] <- 1
  }
  else if(df1[i,8] == '8'){
    Asst_at_surg[i] <- 1
  }
  else if(df1[i,8] == '9'){
    Other_med_itm[i] <- 1
  }
  else if(df1[i,8] == '3'){
    Consultation[i] <- 1
  }
  else if(df1[i,8] == 'P'){
    Prosthtc_Orthtc[i] <- 1
  }
  else if(df1[i,8] == 'S'){
    Med_supply[i] <- 1
  }
  else if(df1[i,8] == 'G'){
    Imnsprsv_drg[i] <- 1
  }
  else if(df1[i,8] == 'N'){
    Kidney_dnr[i] <- 1
  }
  else if(df1[i,8] == 'O'){
    Whole_bld[i] <- 1
  }
}
```

```
df.ServiceCd <- cbind.data.frame(Med_care, Diag_lab, Diag_radiol, Surgery, Flu_vacc, Ambulance, Outpati

df2 <- cbind.data.frame(df2, df.ServiceCd)

## Finally add age.

df2 <- cbind.data.frame(df2, df1$Age_Group)
colnames(df2)[67] <- "Age_Group"


## Get 2 random samples of 20,000 observations.

set.seed(1)
ind <- seq(from = 1, to = 2801660)
ind2 <- sample(ind, size = 20000, replace = FALSE)

df2.sample <- df2[ind2,]
df2.nonsample <- df2[-ind2,]

ind3 <- seq(from = 1, to = 2781660)
ind4 <- sample(ind3, size = 20000, replace = FALSE)
df2.validset <- df2[ind4,]

summary(df2)
```

```
##       Cost               Sex        Infectious_dz        Neoplasms
##  Min.   :    0.00   Min.   :1.000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:   15.00   1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :   45.00   Median :2.000   Median :0.000000   Median :0.000000
##  Mean   :   82.01   Mean   :1.549   Mean   :0.002012   Mean   :0.006134
##  3rd Qu.:   85.00   3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :44000.00   Max.   :2.000   Max.   :1.000000   Max.   :1.000000
##   Immunity_dz          Blood_dz       Mental_disorder
##  Min.   :0.000000   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.000000
##  Median :0.000000   Median :0.00000   Median :0.000000
##  Mean   :0.004397   Mean   :0.00168   Mean   :0.004838
##  3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.00000   Max.   :1.000000
##    Nervous_dz          Sense_dz        Circulatory_dz
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.003531   Mean   :0.003069   Mean   :0.009265
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##  Respiratory_dz      Digestive_dz      Genitourinary_dz
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.004861   Mean   :0.005693   Mean   :0.004774
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##  Pregnancy_complic    Skin_dz         Musculoskeletal_dz
##  Min.   :0e+00     Min.   :0.000000   Min.   :0.0000    
##  1st Qu.:0e+00     1st Qu.:0.000000   1st Qu.:0.0000    
##  Median :0e+00     Median :0.000000   Median :0.0000    
##  Mean   :4e-07     Mean   :0.003396   Mean   :0.0106    
##  3rd Qu.:0e+00     3rd Qu.:0.000000   3rd Qu.:0.0000    
##  Max.   :1e+00     Max.   :1.000000   Max.   :1.0000    
##  Congenital_anom    Perinatal_condit  Ill_defined_condit
##  Min.   :0.000000   Min.   :0.0e+00   Min.   :0.00000   
##  1st Qu.:0.000000   1st Qu.:0.0e+00   1st Qu.:0.00000   
##  Median :0.000000   Median :0.0e+00   Median :0.00000   
##  Mean   :0.000449   Mean   :4.6e-06   Mean   :0.00954   
##  3rd Qu.:0.000000   3rd Qu.:0.0e+00   3rd Qu.:0.00000   
##  Max.   :1.000000   Max.   :1.0e+00   Max.   :1.00000   
##  Injury_Poison      External_Cz_of_Inj Fact_Inf_Hlth_Srvc
##  Min.   :0.000000   Min.   :0.0e+00    Min.   :0.0000    
##  1st Qu.:0.000000   1st Qu.:0.0e+00    1st Qu.:1.0000    
##  Median :0.000000   Median :0.0e+00    Median :1.0000    
##  Mean   :0.007468   Mean   :8.2e-06    Mean   :0.9183    
##  3rd Qu.:0.000000   3rd Qu.:0.0e+00    3rd Qu.:1.0000    
##  Max.   :1.000000   Max.   :1.0e+00    Max.   :1.0000    
##      EPOOS             Venous           Hosp_Care         Hematology     
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.00000   Min.   :0.000000  
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.000000  
##  Median :0.00000   Median :0.000000   Median :0.00000   Median :0.000000  
##  Mean   :0.07491   Mean   :0.008123   Mean   :0.01399   Mean   :0.007108  
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.000000  
##  Max.   :1.00000   Max.   :1.000000   Max.   :1.00000   Max.   :1.000000  
##   Organ_Dz_Pnl       Phys_Med_Rhb      Other_Service       Ofc_Vst      
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:1.0000   1st Qu.:0.0000  
##  Median :0.000000   Median :0.000000   Median :1.0000   Median :0.0000  
##  Mean   :0.009519   Mean   :0.004626   Mean   :0.8817   Mean   :0.1051  
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:1.0000   3rd Qu.:0.0000  
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.0000   Max.   :1.0000  
##    LbTst_Othr         Hosp_Vst          Mnr_Pcdr        LbTst_Vnpctr     
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000000  
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000  
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.000000  
##  Mean   :0.09279   Mean   :0.04293   Mean   :0.04519   Mean   :0.008123  
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000000  
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.000000  
##    LbTst_AGP         Chest_Img          Opthmlgy        LbTst_BldCt    
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000  
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000  
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.0000  
##  Mean   :0.01389   Mean   :0.01581   Mean   :0.01929   Mean   :0.0107  
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000  
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000  
##       ECG           Other_BETOS         Clinic            Solo       
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.00000   Median :1.0000   Median :0.0000   Median :1.0000  
##  Mean   :0.01277   Mean   :0.6334   Mean   :0.1077   Mean   :0.7214  
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000  
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##  Institutional     Clinic_Mult_Specialties Other_Provider   
##  Min.   :0.00000   Min.   :0.0000          Min.   :0.00000  
##  1st Qu.:0.00000   1st Qu.:0.0000          1st Qu.:0.00000  
##  Median :0.00000   Median :0.0000          Median :0.00000  
##  Mean   :0.02741   Mean   :0.1048          Mean   :0.03867  
##  3rd Qu.:0.00000   3rd Qu.:0.0000          3rd Qu.:0.00000  
##  Max.   :1.00000   Max.   :1.0000          Max.   :1.00000  
##  Service_Count        Med_care         Diag_lab       Diag_radiol    
##  Min.   :  0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:  1.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :  1.000   Median :0.0000   Median :0.0000   Median :0.0000  
##  Mean   :  2.068   Mean   :0.3694   Mean   :0.2364   Mean   :0.1611  
##  3rd Qu.:  1.000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000  
##  Max.   :999.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##     Surgery          Flu_vacc          Ambulance       Outpatient_MH    
##  Min.   :0.0000   Min.   :0.000000   Min.   :0.00000   Min.   :0.00000  
##  1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.00000  
##  Median :0.0000   Median :0.000000   Median :0.00000   Median :0.00000  
##  Mean   :0.1153   Mean   :0.003678   Mean   :0.02712   Mean   :0.01043  
##  3rd Qu.:0.0000   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.00000  
##  Max.   :1.0000   Max.   :1.000000   Max.   :1.00000   Max.   :1.00000  
##      Vision          Anesthesia       Thrp_radiol      Ambul_surg_cntr   
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000000  
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000  
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.000000  
##  Mean   :0.00829   Mean   :0.03914   Mean   :0.01084   Mean   :0.007788  
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000000  
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.000000  
##     Hearing          Asst_at_surg      Other_med_itm     
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.0000000  
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.0000000  
##  Median :0.000000   Median :0.000000   Median :0.0000000  
##  Mean   :0.001841   Mean   :0.004709   Mean   :0.0005932  
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.0000000  
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.0000000  
##   Consultation      Prosthtc_Orthtc       Med_supply       
##  Min.   :0.000000   Min.   :0.0000000   Min.   :0.0000000  
##  1st Qu.:0.000000   1st Qu.:0.0000000   1st Qu.:0.0000000  
##  Median :0.000000   Median :0.0000000   Median :0.0000000  
##  Mean   :0.002171   Mean   :0.0006475   Mean   :0.0004697  
##  3rd Qu.:0.000000   3rd Qu.:0.0000000   3rd Qu.:0.0000000  
##  Max.   :1.000000   Max.   :1.0000000   Max.   :1.0000000  
##   Imnsprsv_drg        Kidney_dnr         Whole_bld         Age_Group    
##  Min.   :0.00e+00   Min.   :0.00e+00   Min.   :0.0e+00   Min.   :1.000  
##  1st Qu.:0.00e+00   1st Qu.:0.00e+00   1st Qu.:0.0e+00   1st Qu.:2.000  
##  Median :0.00e+00   Median :0.00e+00   Median :0.0e+00   Median :3.000  
##  Mean   :1.07e-05   Mean   :1.14e-05   Mean   :2.9e-06   Mean   :3.365  
##  3rd Qu.:0.00e+00   3rd Qu.:0.00e+00   3rd Qu.:0.0e+00   3rd Qu.:5.000  
##  Max.   :1.00e+00   Max.   :1.00e+00   Max.   :1.0e+00   Max.   :6.000  
```

Describing the Data

Given the complexity and quantity of the variables in the dataset, it is difficult to immediately recognize patterns in the dataset. A summary of the data frame (df2) shows the distributions of relevant variables, most of which are between 0 and 1 since they represent binary factors. Interesting associations between variables can be seen later in the regression stage of this investigation.

There are 66 non-cost variables in the dataset, which fall into 10 broad groups of variables: sex of the beneficiary, beneficiary age group at the year 2010, the beneficiary's International Classification of Diseases, Ninth Revision, Clinical Modification (ICD 9) diagnosis, the provider type, the number of services processed per line item on the carrier claim, the type of service, the place of service, the payment made for the line item, the Healthcare Common Procedure Coding System (HCPCS) codes which identify items and services, and the Berenson-Eggers Type of Service (BETOS) code for the line item based on generally agreed upon clinically meaningful groupings of procedures and services.

While clustering likely will not reveal much of interest due to variable complexity (no clear binary variable is likely strongly related to cost, and if so, finding which one requires doing the modeling portion of this investigation first), a cluster of 2 excluding sex to see whether claims of different sexes are significantly distinct may be of interest. Interestingly enough, the clusters produce very similar sex ratios, suggesting that sex is a relatively uninportant parameter of interest in the first split of the dataset.

Clustering

```r
## cluster using dataset exlucding sex
df3 <- df2.sample[,-2]
clst <- kmeans(df3, centers = 2)

## counters for number of obs in each cluster
ct1 <- 0
ct2 <- 0

## counters for number of spam in each cluster
ct3 <- 0
ct4 <- 0

## check which cluster and if spam for each obs. count up.
for(i in 1:20000){
  if(clst$cluster[i] == 1){
    ct1 <- ct1 + 1
    if(df2.sample$Sex[i] == 1){
      ct3 <- ct3 + 1
    }
  }
  else{
    ct2 <- ct2 + 1
    if(df2.sample$Sex[i] == 1){
      ct4 <- ct4 + 1
    }
  }
}

## calculate and output percentages
pct1 <- ct3/ct1
pct2 <- ct4/ct2
pct1
```

```
## [1] 0.4475138
```

```
pct2
```

```
## [1] 0.4517262
```

Cost Models

```
## I. Linear Regression
## First, regress on all factors.
mod1 <- glm(Cost ~ ., family = gaussian, data = df2.sample)
summary(mod1)
```

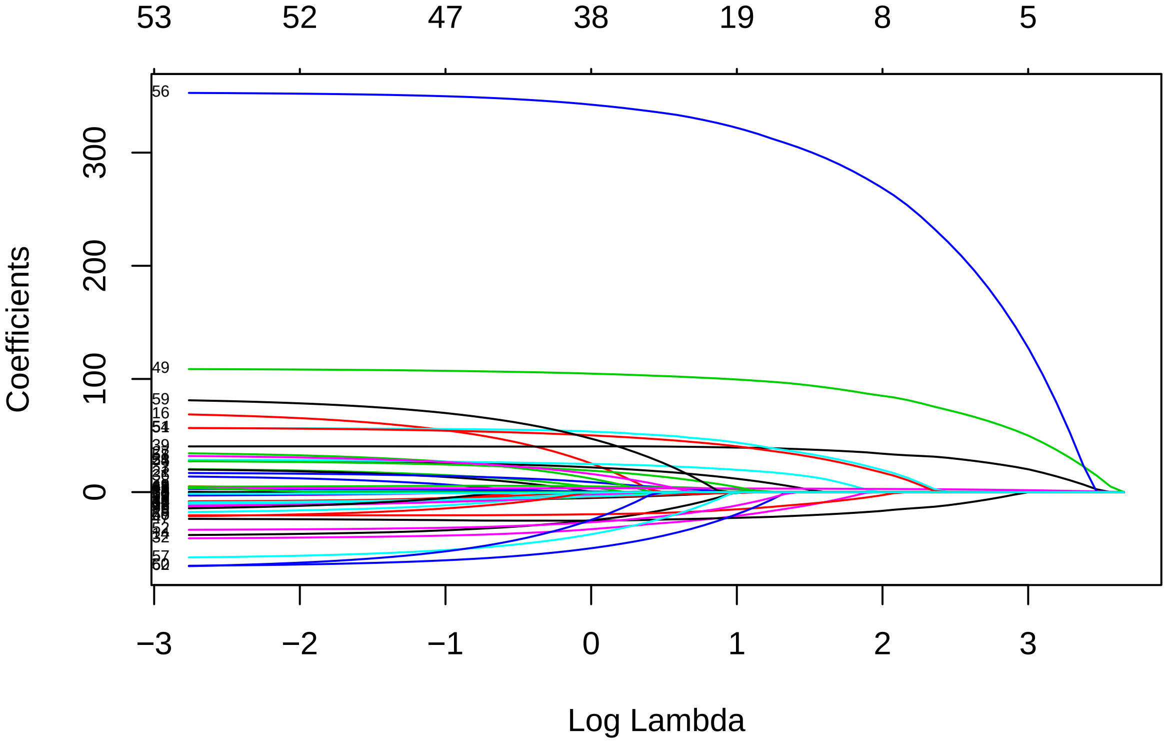
```
##
## Call:
## glm(formula = Cost ~ ., family = gaussian, data = df2.sample)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1104.9    -48.2    -12.3     13.3   9556.6
##
## Coefficients: (12 not defined because of singularities)
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              20.5650    54.3774   0.378 0.705293
## Sex                      -8.5665     2.5825  -3.317 0.000911 ***
## Infectious_dz           -28.7340    25.8003  -1.114 0.265419
## Neoplasms                 3.9776    16.2402   0.245 0.806517
## Immunity_dz             -15.8837    17.3598  -0.915 0.360219
## Blood_dz                -36.2272    27.7574  -1.305 0.191861
## Mental_disorder         -20.8997    17.5656  -1.190 0.234136
## Nervous_dz                3.8649    22.5119   0.172 0.863689
## Sense_dz                -11.1131    22.1436  -0.502 0.615768
## Circulatory_dz          -11.7444    12.3865  -0.948 0.343058
## Respiratory_dz          -20.6081    17.0153  -1.211 0.225854
## Digestive_dz            -19.6301    16.4755  -1.191 0.233482
## Genitourinary_dz        -30.0677    18.0527  -1.666 0.095819 .
## Pregnancy_complic             NA         NA      NA       NA
## Skin_dz                 -56.1830    19.9888  -2.811 0.004948 **
## Musculoskeletal_dz      -17.7696    12.4542  -1.427 0.153653
## Congenital_anom          54.0827    51.3022   1.054 0.291805
## Perinatal_condit              NA         NA      NA       NA
## Ill_defined_condit      -19.9543    12.5050  -1.596 0.110568
## Injury_Poison           -17.7839    13.2671  -1.340 0.180115
## External_Cz_of_Inj            NA         NA      NA       NA
## Fact_Inf_Hlth_Srvc            NA         NA      NA       NA
## EPOOS                     2.6976     8.2872   0.326 0.744795
## Venous                  -54.4091    14.0224  -3.880 0.000105 ***
## Hosp_Care                27.8181    12.4726   2.230 0.025736 *
## Hematology               -1.0256    19.4557  -0.053 0.957958
## Organ_Dz_Pnl              0.2649    15.6355   0.017 0.986484
## Phys_Med_Rhb             35.8829    19.1800   1.871 0.061381 .
## Other_Service                 NA         NA      NA       NA
## Ofc_Vst                 -40.1635     7.5269  -5.336 9.61e-08 ***
## LbTst_Othr              -43.8632     6.2630  -7.004 2.57e-12 ***
## Hosp_Vst                -11.5132     7.5831  -1.518 0.128965
## Mnr_Pcdr                -81.7203     6.3968 -12.775  < 2e-16 ***
## LbTst_Vnpctr                  NA         NA      NA       NA    
## LbTst_AGP               -50.4428    13.5472  -3.723 0.000197 ***
## Chest_Img               -50.6033    10.4356  -4.849 1.25e-06 ***
## Opthmlgy                -25.1311    11.7920  -2.131 0.033085 *  
## LbTst_BldCt             -52.0050    16.1778  -3.215 0.001308 ** 
## ECG                     -53.0860    11.4996  -4.616 3.93e-06 ***
## Other_BETOS                   NA         NA      NA       NA    
## Clinic                    1.0292     7.2254   0.142 0.886735    
## Solo                     14.1606     6.2829   2.254 0.024218 *  
## Institutional            14.2226    87.6878   0.162 0.871153    
## Clinic_Mult_Specialties   9.6095     8.3593   1.150 0.250339    
## Other_Provider                NA         NA      NA       NA    
## Service_Count             3.4027     0.1114  30.533  < 2e-16 ***
## Med_care                 68.0322    53.7724   1.265 0.205819    
## Diag_lab                 44.7243    53.8512   0.831 0.406257    
## Diag_radiol              47.5604    53.7946   0.884 0.376647    
## Surgery                 177.1005    53.8229   3.290 0.001002 ** 
## Flu_vacc                  1.8720    57.7475   0.032 0.974140    
## Ambulance               122.9227   102.9222   1.194 0.232364    
## Outpatient_MH            34.4240    55.1928   0.624 0.532829    
## Vision                   51.8404    56.2331   0.922 0.356601    
## Anesthesia              125.2597    54.0670   2.317 0.020527 *  
## Thrp_radiol              95.9720    54.9755   1.746 0.080875 .  
## Ambul_surg_cntr         421.4021    55.6747   7.569 3.92e-14 ***
## Hearing                   9.0450    60.0825   0.151 0.880337    
## Asst_at_surg            101.1097    56.8098   1.780 0.075125 .  
## Other_med_itm           151.5484    69.4542   2.182 0.029122 *  
## Consultation             45.3672    59.6564   0.760 0.446980    
## Prosthtc_Orthtc          74.4998    74.1897   1.004 0.315304    
## Med_supply                    NA         NA      NA       NA    
## Imnsprsv_drg                  NA         NA      NA       NA    
## Kidney_dnr                    NA         NA      NA       NA    
## Whole_bld                     NA         NA      NA       NA    
## Age_Group                -1.4944     0.7823  -1.910 0.056109 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 28821.1)
## 
##     Null deviance: 679538623  on 19999  degrees of freedom
## Residual deviance: 574836762  on 19945  degrees of freedom
## AIC: 262192
## 
## Number of Fisher Scoring iterations: 2
```

```r
## Second, regress on factors that were significantly related (at the highest level of significance) to
mod2 <- glm(Cost ~ Venous+Ofc_Vst+LbTst_Othr+Mnr_Pcdr+LbTst_AGP+Chest_Img+ECG+
              Service_Count+Surgery+Ambul_surg_cntr, family = gaussian, data = df2.sample)
summary(mod2)
```

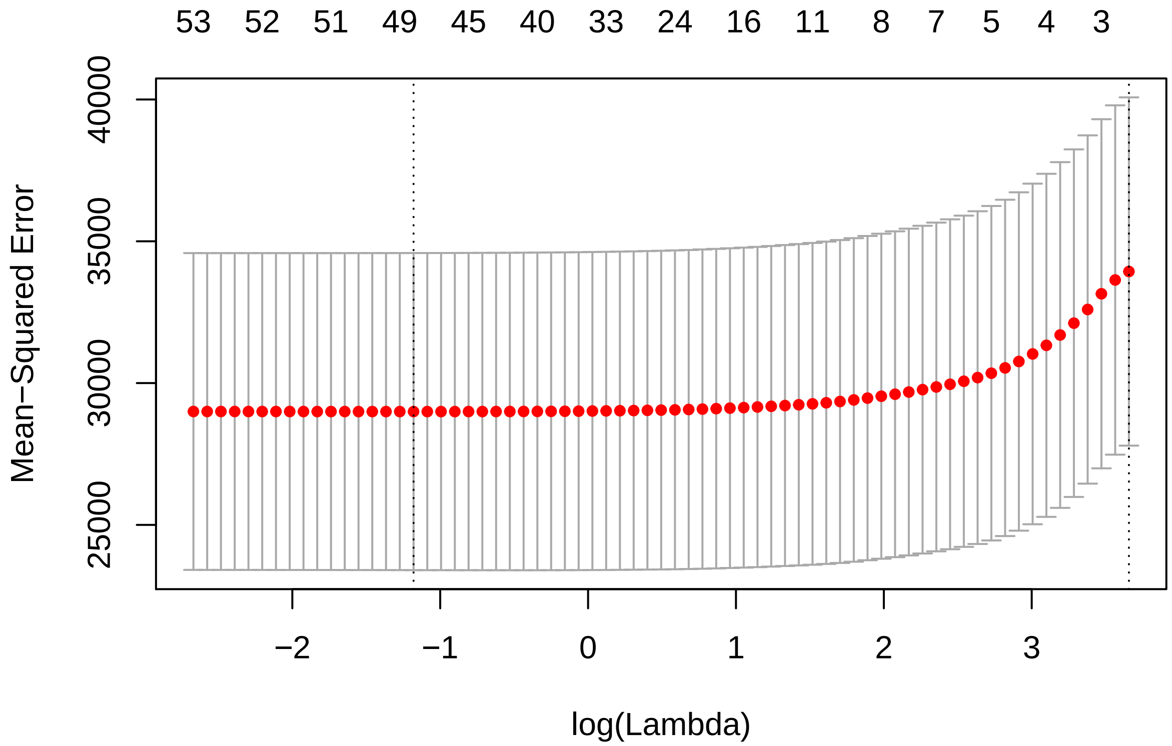
```
## 
## Call:
## glm(formula = Cost ~ Venous + Ofc_Vst + LbTst_Othr + Mnr_Pcdr + 
##     LbTst_AGP + Chest_Img + ECG + Service_Count + Surgery + Ambul_surg_cntr, 
##     family = gaussian, data = df2.sample)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1136.8    -53.4    -13.4     11.6   9564.6
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      74.8305     1.5798  47.366  < 2e-16 ***
## Venous          -73.6811    13.5743  -5.428 5.77e-08 ***
## Ofc_Vst         -33.7292     4.0543  -8.319  < 2e-16 ***
## LbTst_Othr      -62.4998     4.2897 -14.570  < 2e-16 ***
## Mnr_Pcdr        -71.7200     6.0073 -11.939  < 2e-16 ***
## LbTst_AGP       -68.8022    10.3797  -6.629 3.48e-11 ***
## Chest_Img       -65.5131    10.1336  -6.465 1.04e-10 ***
## ECG             -70.4504    10.9760  -6.419 1.41e-10 ***
## Service_Count     3.5400     0.1108  31.951  < 2e-16 ***
## Surgery         114.8864     3.8996  29.461  < 2e-16 ***
## Ambul_surg_cntr 357.0211    13.6987  26.062  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 29273.82)
##
##     Null deviance: 679538623  on 19999  degrees of freedom
## Residual deviance: 585154349  on 19989  degrees of freedom
## AIC: 262460
##
## Number of Fisher Scoring iterations: 2
```

```
## II. Ridge Regression
x1 <- as.matrix(df2.sample[,2:67])
y1 <- df2.sample$Cost
cv.ridge1 <- cv.glmnet(x1, y1, alpha=0)
fit.ridge1 <- glmnet(x1, y1, alpha=0)
## plot(fit.ridge1, xvar="lambda", label=T)
## plot(cv.ridge1)

## III. Lasso Regression
fit.lasso1 <- glmnet(x1, y1, alpha=1)
plot(fit.lasso1, xvar="lambda", label=T)
```



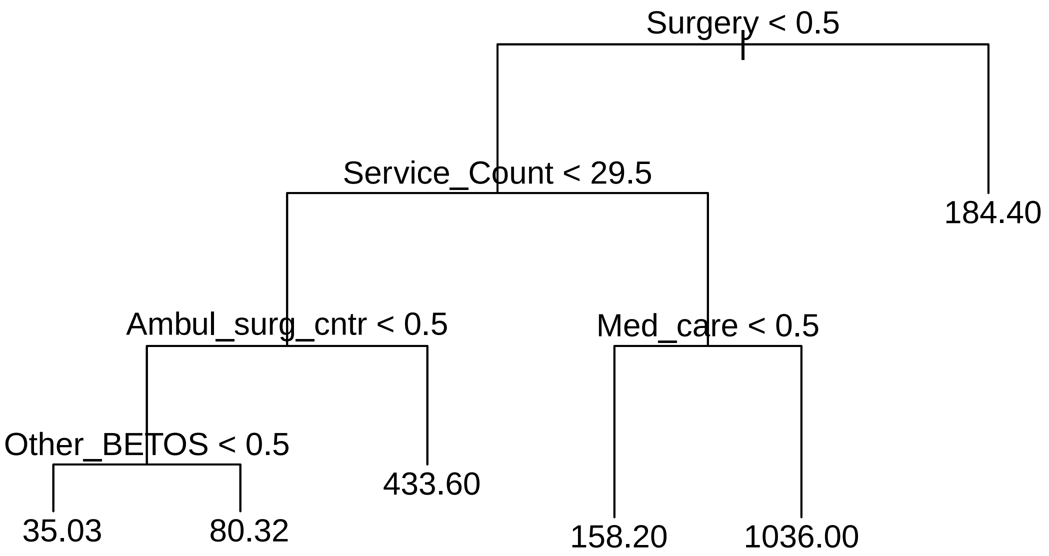
```
cv.lasso1 <- cv.glmnet(x1, y1, alpha=1)
plot(cv.lasso1)
```



```
coef(cv.lasso1)
```

```
## 67 x 1 sparse Matrix of class "dgCMatrix"
##                                 1
## (Intercept)              82.97575
## Sex                       .      
## Infectious_dz             .      
## Neoplasms                 .      
## Immunity_dz               .      
## Blood_dz                  .      
## Mental_disorder           .      
## Nervous_dz                .      
## Sense_dz                  .      
## Circulatory_dz            .      
## Respiratory_dz            .      
## Digestive_dz              .      
## Genitourinary_dz          .      
## Pregnancy_complic         .      
## Skin_dz                   .      
## Musculoskeletal_dz        .      
## Congenital_anom           .      
## Perinatal_condit          .      
## Ill_defined_condit        .      
## Injury_Poison             .      
## External_Cz_of_Inj        .      
## Fact_Inf_Hlth_Srvc        .      
## EPOOS                    .
## Venous                   .
## Hosp_Care                .
## Hematology               .
## Organ_Dz_Pnl             .
## Phys_Med_Rhb             .
## Other_Service            .
## Ofc_Vst                  .
## LbTst_Othr               .
## Hosp_Vst                 .
## Mnr_Pcdr                 .
## LbTst_Vnpctr             .
## LbTst_AGP                .
## Chest_Img                .
## Opthmlgy                 .
## LbTst_BldCt              .
## ECG                      .
## Other_BETOS              .
## Clinic                   .
## Solo                     .
## Institutional            .
## Clinic_Mult_Specialties  .
## Other_Provider           .
## Service_Count            .
## Med_care                 .
## Diag_lab                 .
## Diag_radiol              .
## Surgery                  .
## Flu_vacc                 .
## Ambulance                .
## Outpatient_MH            .
## Vision                   .
## Anesthesia               .
## Thrp_radiol              .
## Ambul_surg_cntr          .
## Hearing                  .
## Asst_at_surg             .
## Other_med_itm            .
## Consultation             .
## Prosthtc_Orthtc          .
## Med_supply               .
## Imnsprsv_drg             .
## Kidney_dnr               .
## Whole_bld                .
## Age_Group                .
```

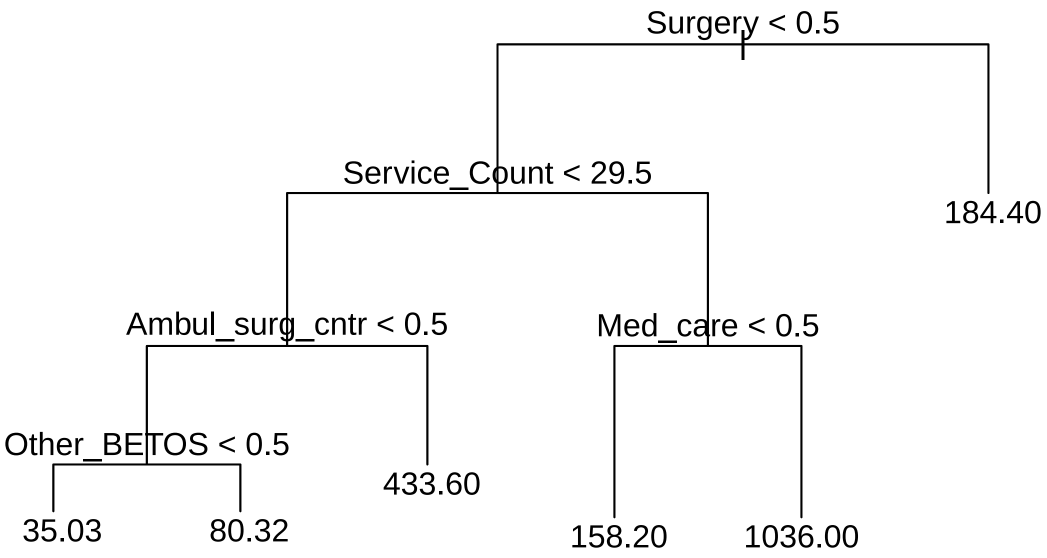
```
## IV. Decision Trees
cost_tree <- tree(Cost ~ ., data=df2.sample)
plot(cost_tree);text(cost_tree, pretty = 1)
```



```
cv.cost_tree <- cv.tree(cost_tree)
cv.cost_tree
```

```
## $size
## [1] 6 5 4 1
##
## $dev
## [1] 571869454 580447046 599320611 679621733
##
## $k
## [1]      -Inf  8524408 21587768 28718925
##
## $method
## [1] "deviance"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

```
prune.cost_tree <- prune.tree(cost_tree, best = 6)
plot(prune.cost_tree);text(prune.cost_tree, pretty = 1)
```



```
## Below, each model makes predictions based on the validation set, and then mean squared errors for ea
mod1pred <- predict(mod1, df2.validset)
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading
```

```
mse1 <- mean((df2.validset$Cost-mod1pred)[1:20000]^2)

mod2pred <- predict(mod2, df2.validset)
mse2 <- mean((df2.validset$Cost-mod2pred)[1:20000]^2)

ridge_pred <- predict(fit.ridge1, s=cv.ridge1$lambda.min, newx=as.matrix(df2.validset[,2:67]))
mse3 <- mean((df2.validset$Cost-ridge_pred)[1:20000]^2)

lasso_pred <- predict(fit.lasso1, cv.lasso1$lambda.min, newx=as.matrix(df2.validset[,2:67]))
mse4 <- mean((df2.validset$Cost-lasso_pred)[1:20000]^2)

tree_pred <- predict(cost_tree, df2.validset)
mse5 <- mean((df2.validset$Cost-tree_pred)[1:20000]^2)

prune_tree_pred <- predict(prune.cost_tree, df2.validset)
mse6 <- mean((df2.validset$Cost-prune_tree_pred)[1:20000]^2)

mean_vec <- rep(mean(df2.sample$Cost), 20000)
mse7 <- mean((df2.validset$Cost-mean_vec)[1:20000]^2)

df.mse <- data.frame(mse1, mse2, mse3, mse4, mse5, mse6, mse7)
```

```
names(df.mse) <- c("Linear Regression - All Variables","Linear Regression - Significant Variables", "Ri
df.mse
```

```
##   Linear Regression - All Variables
## 1                          35779.96
##   Linear Regression - Significant Variables Ridge Regression
## 1                                   36305.8          35740.1
##   Lasso Regression     Tree Pruned Tree Guess Average Cost
## 1         35764.28 34912.97    34912.97           40114.89
```

Results and Conclusion

I created six models to infer the relationship between a given Medicare claim's cost and a variety of factors related to said claim. The first model linearly regressed Cost on all 66 factors in the data frame, while the second linearly regression Cost on the 10 factors which were significantly related to Cost in the first regression (at the most stringent level of significance - ***). The third model used cross-validation to fit a ridge regression that minimized mean squared error, while the fourth model used cross-validation to fit a lasso regression that minimized mean squared error. Finally, the fifth model created a basic tree and the sixth pruned said tree using cross-validation.

The first model included all features possible, while the second selected features that I knew were likely to be related to Cost given the output of the first model. The third model also used all features possible, whil the fourth performed an automatic form of feature selection through the lasso penalty. The tree models also selected relevant features for me by choosing what features created optimal splits. Feature selection was therefore either a choice made by the machine learning tool or a choice to include all features as to best predict Cost; both options attempt to best understand the relationship between the factors and Cost, and are consistent with my research goal.

The above dataframe shows that both trees - which were equivalent since the pruned tree did not remove any nodes from the basic tree - had the lowest mean squared error. Those trees show that the factors of whether the claim's type of service was surgery, whether the service count was greater than 29.5, whether type of service was medical care (if not surgery), whether the type of service was facility usage of an ambulatory surgical center, and whether the Berenson-Eggers Type of Service code was not one of the 9 most frequent codes (i.e. was label Other_BETOS) were the most determinative factors in inferring the cost of a claim. Comparing the last mean squared error (mse7) to the mse of the trees shows that trees moderately improve modeling of the relationship between a claim's cost and other factors compared to simply guessing the cost to be the mean of the claims' costs in the sample set.

One can conclude from this analysis that important and interpretable focal points for reducing the incidence of especially high cost Medicare claims include the costs and usage of ambulatory surgical centers and costs and incidence of surgery. These two factors have a significant relationship with Medicare claim cost that merits public policy attention for understand one facet of what may make Medicare costly.