Predicting Emergency Room Payer Mix

Emily Rinaldi

September 7, 2016

Introduction

The Problem

Public information about the patient population that visits a hospital's emergency room is generally very limited, so can demographic and economic factors for a particular geographic area predict the patient mix of emergency departments in that area?

The Client

The client is an emergency department physician staffing provider, which contracts with hospitals to staff physicians and performs all coding, billing, and collection functions related to the physicians' services. The client's main source of revenue is fee-for-service collections, and the revenue for a particular patient encounter is dependent on the health care coverage of the patient who was treated.

Most payers can be grouped into the following classes, in order from most expected revenue to least: Commercial Insurance, Medicare, Medicaid, and Self Pay. In the United States, emergency departments are subject to the Emergency Medical Treatment and Labor Act (EMTALA), which is a federal law that requires emergency department providers to stabalize and treat any patient that arrives, regardless of their ability to pay. Because treating each patient is costly for the client, operating at facilities where there are enough patients with insurance to cover provider staffing costs is paramount. Being able to accurately estimate the payer mix for potential client facilities would enable the company to focus business development efforts on those facilities located in geographic areas that indicate the most favorable payer mixes.

Data Sources

1. Emergency Department Data By Expected Payer Source 2010-2014: This dataset contains the distribution of emergency department encounters and admits by expected payer for California hospitals years 2010-2014.

2. 2010-2014 American Community Survey: The ACS collects information such as age, race, income, commute time to work, home value, veteran status, and other important data, and it is available by geographic area.

3. 2010 Census Demographic Profile: The Demographic Profile contains data on population characteristics including sex, age, race, household relationship, household type, group quarters population; and housing characteristics including occupancy and tenure.

Example Data

After cleaning the data and joining it into a single dataset, we have a dataframe of 66 variables with 346 observations. Each observation represents an emergency facility and the demographic and economic characteristics of the zip code in which it is located. A glimpse of the data is provided below.

```
glimpse(all_data)
```

```
## Observations: 346
## Variables: 66
## $ year                   <int> 2005, 2006, 2012, 2014, 2014, 2014, 201...
## $ id                     <int> 106551034, 106190230, 106301132, 106014...
## $ facility               <chr> "SONORA REGIONAL MEDICAL CENTER - FORES...
## $ MSSA_desig             <fctr> Rural, Urban, Urban, Urban, Rural, Urb...
## $ MSSA_name              <chr> "COLUMBIA/JAMESTOWN/SONORA", "DEL AIRE/...
## $ county                 <chr> "TUOLUMNE", "LOS ANGELES", "ORANGE", "A...
## $ address                <chr> "ONE SOUTH FOREST ROAD", "333 NORTH PRA...
## $ city                   <chr> "SONORA", "INGLEWOOD", "ANAHEIM", "CAST...
## $ zip                    <chr> "95370", "90301", "92807", "94546", "95...
## $ owner                  <fctr> Nonprofit, Investor, Nonprofit, Nonpro...
## $ owner_type             <chr> "Corporation", "Corporation", "Corporat...
## $ EMS_level              <chr> NA, NA, "Emergency - Basic", "Emergency...
## $ trauma_desig           <chr> NA, NA, NA, "Level II", NA, NA, "Level ...
## $ location               <chr> "ONE SOUTH FOREST ROAD\nSONORA, CA 9537...
## $ max_year               <int> 2005, 2006, 2012, 2014, 2014, 2014, 201...
## $ Medi-Cal               <int> 4364, 4889, 4731, 16639, 9780, 19540, 7...
## $ Medicare               <int> 4545, 2981, 14357, 6515, 5056, 8770, 81...
## $ Other                  <int> 1081, 417, 264, 886, 248, 1972, 1180, 3...
## $ Private Coverage       <int> 8015, 12582, 46996, 8392, 2282, 16127, ...
## $ Self Pay               <int> 1393, 6306, 2734, 4105, 1630, 7607, 341...
## $ tot_volume             <int> 19398, 27175, 69082, 36537, 18996, 5401...
## $ pct_comm               <dbl> 0.41318693, 0.46299908, 0.68029299, 0.2...
## $ pct_labor_force        <dbl> 54.7, 66.6, 68.3, 65.0, 50.0, 72.5, 66....
## $ pct_armed_forces       <dbl> 0.0, 0.0, 0.1, 0.0, 0.0, 0.0, 0.2, 0.2,...
## $ pct_unemployed         <dbl> 13.7, 15.3, 8.8, 9.2, 22.8, 8.5, 7.7, 7...
## $ pct_female_labforce    <dbl> 51.2, 61.8, 60.5, 60.8, 47.2, 67.1, 60....
## $ pct_pub_trans          <dbl> 0.4, 8.9, 0.3, 8.2, 1.2, 4.2, 1.9, 1.9,...
## $ pct_service_ind        <dbl> 24.2, 28.1, 11.4, 15.9, 30.2, 12.2, 16....
## $ pct_sales_office       <dbl> 21.2, 27.7, 28.6, 27.0, 22.9, 23.2, 27....
## $ pct_construction       <dbl> 11.0, 8.8, 4.9, 7.6, 19.2, 4.6, 6.2, 6....
## $ pct_transport_ind      <dbl> 9.9, 16.1, 8.8, 7.2, 10.6, 7.3, 6.8, 6....
## $ pct_under10K           <dbl> 5.9, 8.0, 2.0, 3.8, 13.0, 5.2, 3.0, 3.0...
## $ pct_10to15K            <dbl> 6.7, 8.9, 1.5, 4.1, 13.4, 3.3, 2.2, 2.2...
## $ pct_15to25K            <dbl> 11.5, 16.1, 4.9, 7.5, 21.8, 6.4, 5.0, 5...
## $ pct_25to35K            <dbl> 12.2, 13.5, 6.0, 5.8, 18.1, 7.5, 6.2, 6...
## $ pct_35to50K            <dbl> 15.1, 17.1, 8.3, 12.7, 14.0, 9.6, 7.4, ...
## $ pct_50to75K            <dbl> 19.2, 18.5, 14.3, 15.8, 10.0, 16.4, 14....
## $ pct_75to100K           <dbl> 11.6, 9.8, 15.9, 15.0, 5.4, 16.1, 15.2,...
## $ med_househ_income      <int> 48912, 37813, 94697, 75500, 25934, 7767...
```

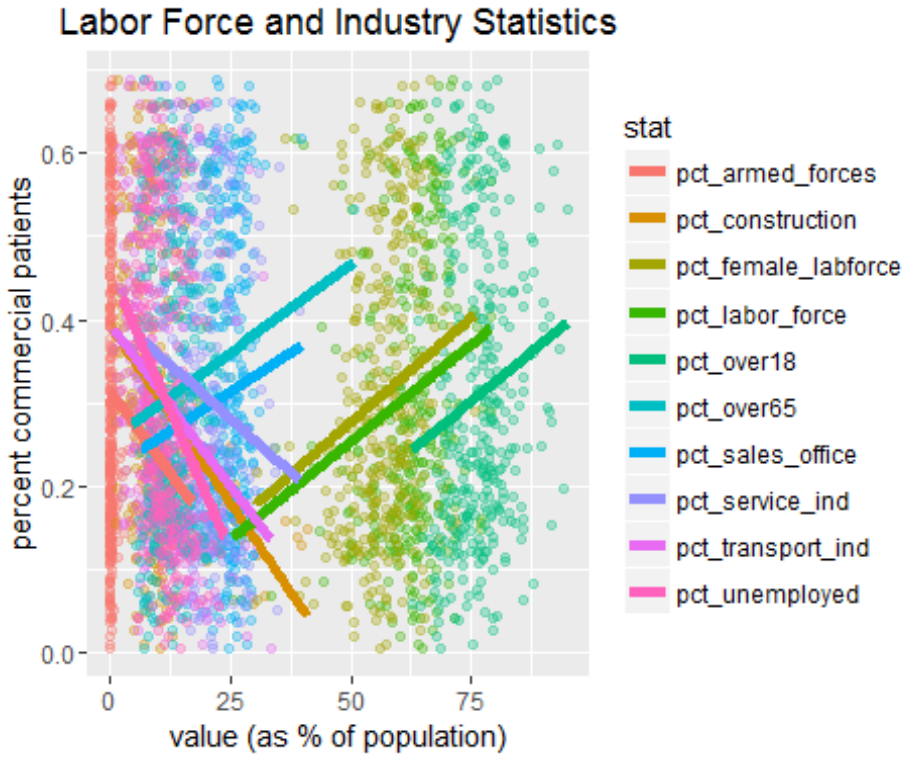
```
## $ mn_househ_income       <int> 64511, 47170, 115405, 91969, 36011, 904...
## $ pct_wSSI               <dbl> 8.2, 8.2, 3.5, 3.7, 14.6, 2.3, 3.6, 3.6...
## $ pct_wcash_assist       <dbl> 3.2, 6.5, 0.9, 2.7, 5.4, 2.2, 1.8, 1.8,...
## $ pct_SNAP               <dbl> 8.6, 13.7, 1.4, 5.3, 17.8, 1.3, 1.0, 1....
## $ per_cap_income         <int> 28416, 16762, 40578, 35537, 16588, 3869...
## $ med_worker_earnings    <dbl> 26946, 23195, 43827, 42637, 18586, 4418...
## $ med_male_earnings      <int> 47516, 30723, 75159, 62245, 34652, 5932...
## $ pct_private_ins        <dbl> 64.6, 42.4, 79.5, 76.1, 31.4, 77.4, 77....
## $ pct_public_ins         <dbl> 43.5, 34.2, 21.9, 26.1, 58.9, 17.4, 23....
## $ pct_no_ins             <dbl> 11.9, 28.0, 9.6, 9.9, 19.4, 12.7, 9.6, ...
## $ pct_poverty            <dbl> 13.8, 25.1, 5.2, 10.2, 34.4, 7.1, 6.4, ...
## $ tot_pop                <int> 26803, 36568, 36171, 42209, 15585, 3077...
## $ med_age                <dbl> 47.9, 32.6, 41.9, 41.2, 40.3, 38.9, 40....
## $ pct_over18             <dbl> 81.1, 73.1, 78.1, 77.5, 76.3, 80.8, 77....
## $ pct_over65             <dbl> 21.6, 8.6, 14.4, 14.7, 15.4, 12.0, 13.8...
## $ pct_black              <dbl> 0.4, 32.3, 2.0, 5.9, 4.0, 2.8, 1.3, 1.3...
## $ pct_asian              <dbl> 1.3, 2.0, 15.3, 16.5, 1.0, 10.6, 8.1, 8...
## $ pct_hisp               <dbl> 8.4, 61.7, 21.0, 18.8, 21.0, 26.2, 19.7...
## $ pct_nonhisp_wh         <dbl> 86.2, 3.2, 58.2, 54.0, 67.6, 57.8, 67.3...
## $ pct_house_wchildren    <dbl> 16.5, 21.9, 19.4, 20.6, 19.3, 17.4, 20....
## $ pct_extfamily_houses   <dbl> 4.5, 12.7, 7.7, 6.4, 7.8, 6.2, 6.5, 6.5...
## $ pct_nonrelative_houses <dbl> 6.8, 6.6, 4.5, 5.7, 10.6, 7.5, 5.8, 5.8...
## $ pct_group_qrts         <dbl> 2.5, 2.1, 0.2, 1.3, 3.0, 0.1, 1.9, 1.9,...
## $ pct_married_houses     <dbl> 49.2, 36.4, 63.7, 49.1, 33.0, 40.7, 62....
## $ pct_sing_mother_houses <dbl> 4.9, 12.6, 4.2, 6.8, 9.7, 4.6, 3.9, 3.9...
## $ avg_household_size     <dbl> 2.30, 3.01, 2.87, 2.58, 2.46, 2.33, 2.8...
## $ pct_vacant_houses      <dbl> 13.4, 5.3, 2.6, 4.8, 28.2, 4.8, 2.7, 2....
```

Data Exploration

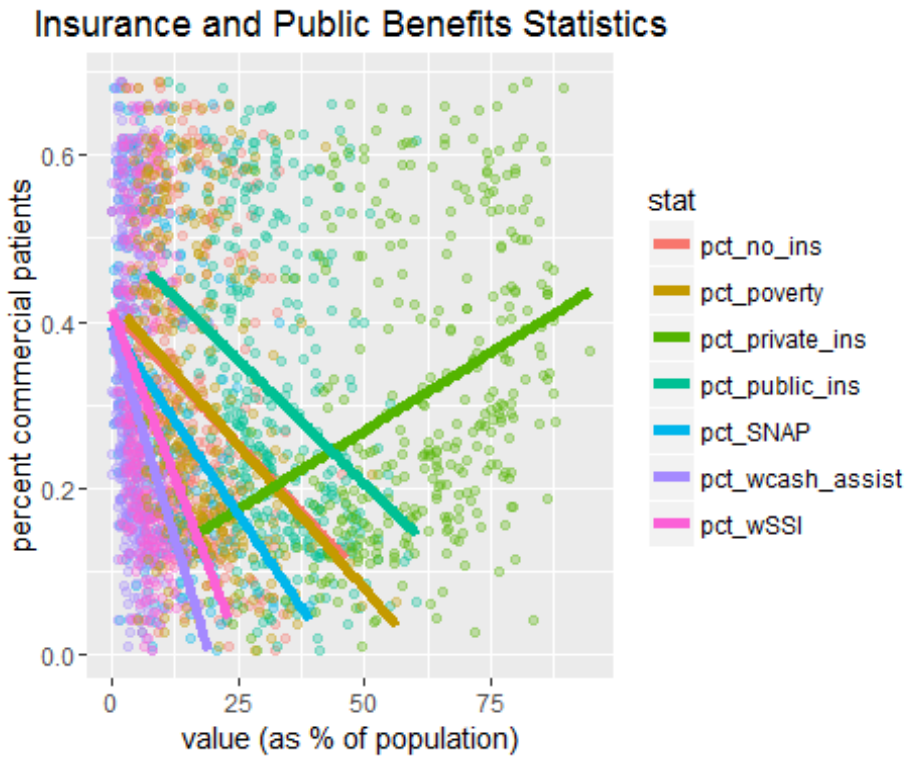
The dataset contains many potentially predictive variables, so we plotted the relationship between each variable and our dependent variable. The following graphs split the variables into related categories to observe correlation with percent of commercial patients as well as collinearity with other similar variables.

We see that the scatterplot data is very noisy, but in most cases, the best fit lines confirm our intuition about the variables' relationships to the percent of commercial patients.

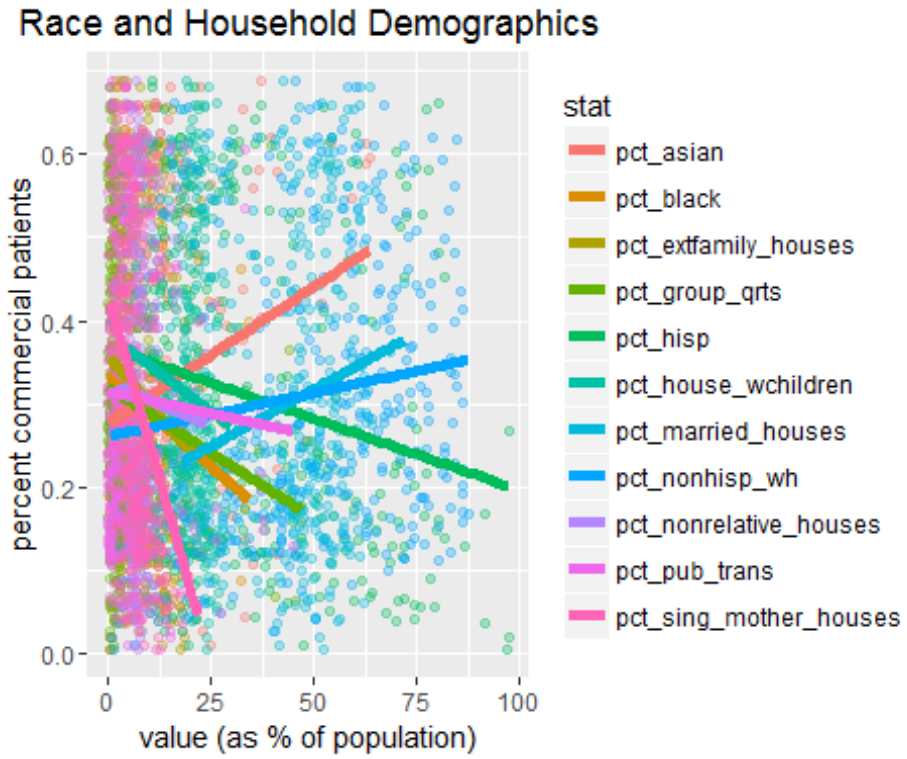
The graph below shows that a higher labor force participation rate indicates more commercial patients at emergency departments located in the same zip code. Interestingly, this also shows that the proportion of people over age 65 has a positive relationship with the percent of commercial patients, even though people over age 65 are eligible for Medicare coverage. Also, of the industries considered, only the sales and office jobs have a positive correlation with commercial patients.



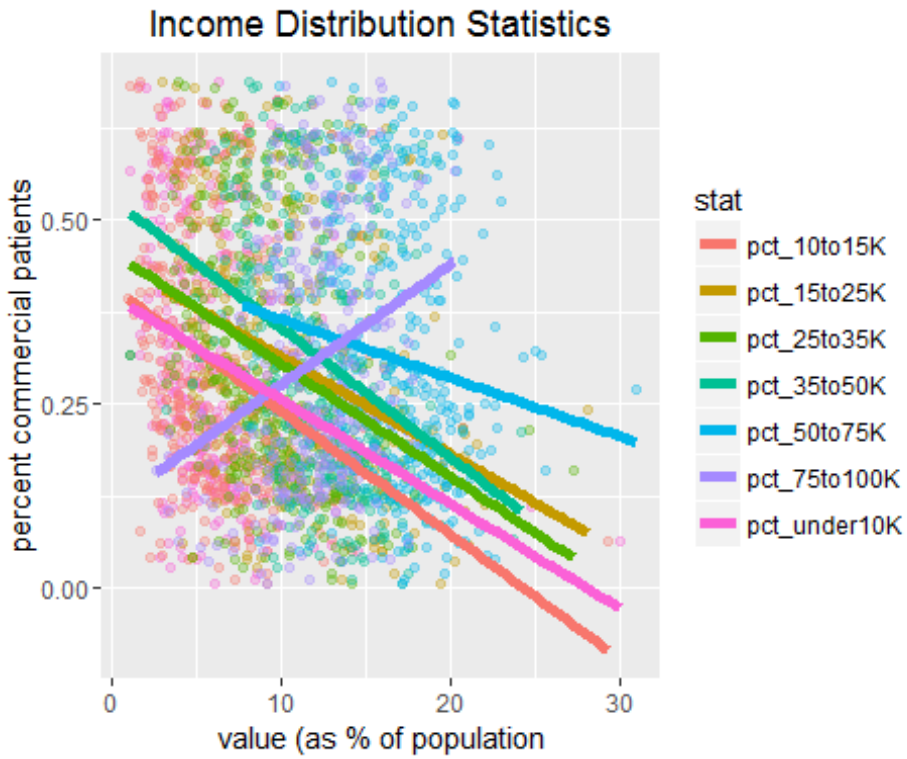
The following graph shows that more people on public assistance indicates fewer commercial patients at facilities in the same area. Also, the only line with a positive slope shows the percent of the population with private insurance. This may serve as confirmation that the Census and ACS data is truly representative of emergency department patients in the same zip code.



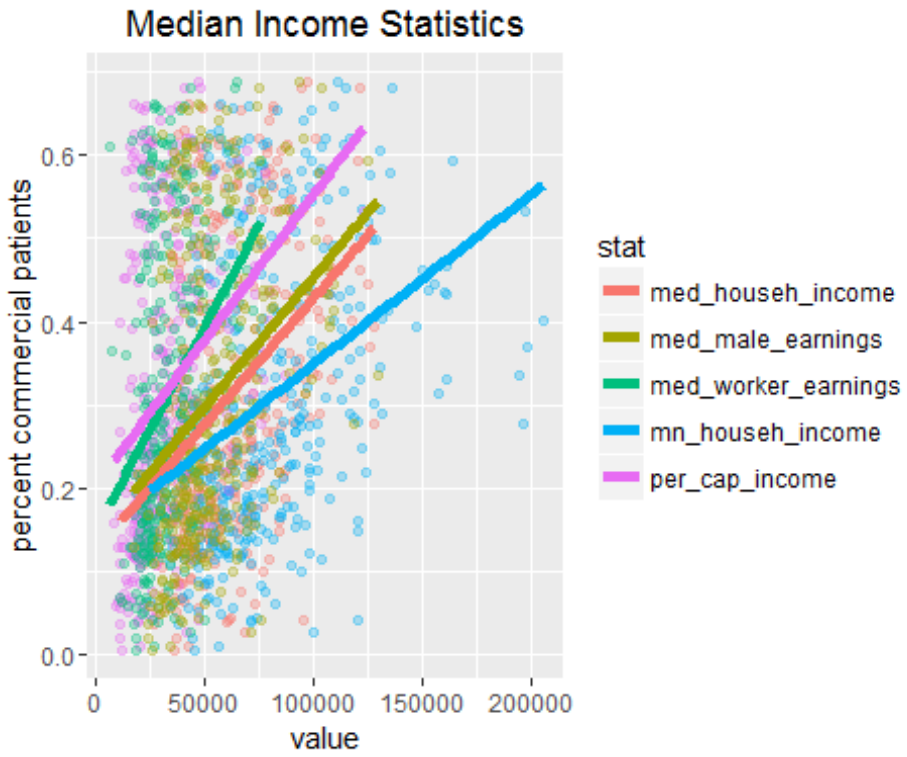
The Race and Household Demographics graph below shows that populations with more Asians, and to a lesser extent more whites, may have more commercially insured individuals. Also, it appears that the proportion of family households headed by a single mother may have a strong inverse relationship with commercial insurance patients.



When observing the income distrubtions of a population, a positive correlation to commercially insured patients is not reached until household incomes exceed $75,000.



As expected, higher median incomes for an area indicate more ER patients with commercial insurance.



Feature Selection and Preprocessing

regsubsets Function

To narrow down the dataset before modeling, we used the regsubsets function from the leaps package. This function performs an exhaustive search algorithm to find the best models of all sizes up to the specified nvmax, which we set at 16 variables. The results indicate which variables should be included in each model.

```
regfit <- regsubsets(pct_comm ~ ., model_data, nvmax = 16)
reg_sum <- summary(regfit)
```

We selected the model size based on which model minimizes Mallow's Cp, a metric that is less biased toward overfitting by adding more variables, and then filtered the dataset to only those variables which are included in the best model. We then added dummy variables for the factor values of owner, which was one of the predictors chosen by regsubsets.

The model that minimizes Mallow's Cp has 13 variables, and the dataset was filtered to include only these variables:

```
##  [1] "(Intercept)"       "ownerNonprofit"    "ownerPublic"
##  [4] "pct_service_ind"   "pct_sales_office"  "pct_75to100K"
##  [7] "med_male_earnings" "pct_public_ins"    "tot_pop"
## [10] "pct_black"         "pct_asian"         "pct_hisp"
## [13] "pct_nonhisp_wh"    "pct_vacant_houses"
```

Remove highly correlated variables

We used the caret package to identify and remove any highly correlated variables that still remain in our dataset, with a cutoff point of 0.75.

```
#remove highly correlated variables, excluding dummy variables
modelCor <- cor(model_filtered2[, -(1:4)])
summary(modelCor[upper.tri(modelCor)])

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.85040 -0.20820 -0.05462 -0.04797  0.16000  0.60550

highlyCorVar <- findCorrelation(modelCor, cutoff = 0.75) + 4
model_filtered3 <- model_filtered2[, -highlyCorVar]
```

Create training and testing datasets

We used the caret package's createDataPartition function to perform a .7/.3 split of our dataset, which is stratified based on the value of our dependent variable, pct_comm.

```
#Create training and testing data sets
set.seed(50)
inTraining <- createDataPartition(model_filtered3$pct_comm, p = .7, list = FA
LSE)
training <- model_filtered3[inTraining,]
testing <- model_filtered3[-inTraining,]
```

Imputation of NAs

Because our dataset is relatively small, we did not want to exclude any observations just because one predictor's value was missing. The owner variable contained 10 missing values which we filled using the bagged trees imputation method.

```
impute_NAs <- preProcess(training[,-1], method = "bagImpute")

set.seed(50)
trainingTransformed <- predict(impute_NAs, training)
testingTransformed <- predict(impute_NAs, testing)
```

Model Training and Analysis

We used the caret package to train two linear regression models and two stochastic gradient boosting models to our training set. For each model, we selected 10-fold cross validation for resampling. The results of our model training and analysis are below.

1. Linear Regression (lm)

Performing a linear regression of all remaining variables against pct_comm on the training set yields the following results:

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33840 -0.10489 -0.03571  0.09110  0.45568
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6.893e-02  1.603e-01  -0.430  0.66760
## ownerNonprofit    -4.786e-02  2.721e-02  -1.759  0.07999 .
## ownerPublic       -7.958e-02  3.633e-02  -2.191  0.02949 *
## pct_service_ind    5.627e-03  2.458e-03   2.290  0.02294 *
## pct_sales_office   2.355e-03  2.740e-03   0.860  0.39089
## pct_75to100K       1.051e-02  4.366e-03   2.406  0.01690 *
## med_male_earnings  3.603e-06  9.600e-07   3.754  0.00022 ***
## pct_public_ins    -1.923e-03  1.618e-03  -1.189  0.23567
## tot_pop           -1.245e-06  6.469e-07  -1.925  0.05546 .
## pct_black         -1.859e-03  1.837e-03  -1.012  0.31257
## pct_asian          2.063e-03  1.038e-03   1.987  0.04809 *
## pct_hisp           3.600e-04  7.148e-04   0.504  0.61498
## pct_vacant_houses  1.675e-03  1.312e-03   1.277  0.20290
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1544 on 231 degrees of freedom
## Multiple R-squared:  0.2603, Adjusted R-squared:  0.2219
## F-statistic: 6.775 on 12 and 231 DF,  p-value: 1.979e-10

##   intercept      RMSE  Rsquared     RMSESD RsquaredSD
## 1      TRUE 0.1575064 0.2259281 0.01722955  0.1455544
```

Below are the results of testing the model on the test dataset:

```
##      RMSE  Rsquared
## 0.1588232 0.2585875
```

The test result's Rsquared is similar to the model Rsquared, but there may be room for improvement. Removing all insignificant variables from Fit1 and training a new model yields the following results:

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33397 -0.11344 -0.03772  0.09895  0.39908
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.726e-01  9.215e-02  -1.873  0.06233 .
## ownerPublic       -2.596e-02  2.796e-02  -0.928  0.35424
## pct_service_ind    5.586e-03  2.292e-03   2.437  0.01554 *
## pct_75to100K       1.105e-02  3.778e-03   2.924  0.00379 **
## med_male_earnings  4.062e-06  6.919e-07   5.871 1.45e-08 ***
## pct_asian          1.991e-03  9.612e-04   2.071  0.03943 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1565 on 238 degrees of freedom
## Multiple R-squared:  0.2167, Adjusted R-squared:  0.2003
## F-statistic: 13.17 on 5 and 238 DF,  p-value: 2.488e-11

##   intercept      RMSE  Rsquared     RMSESD RsquaredSD
## 1      TRUE 0.1573267 0.2168303 0.01396007  0.1371444

##      RMSE  Rsquared
## 0.1694501 0.1592097
```

Interestingly, removing insignificant variables decreased both the Multiple R-squared and the Adjusted R-squared. It also resulted in a lower Rsquared when predicting the test set.

2. Stochastic Gradient Boosting (gbm)

In attempt to improve upon the linear regression models explained above, we tested training the model with interactions of every combination of two terms. None of these interactions were significant. To automate the testing of higher degree interactions, we used the caret package to train a stochastic gradient boosting model, with method set to "gbm".

The GBM model finds the model that maximizes Rsquared value across various tuning parameters. The results of the best tune and its prediction on the testing set are as follows:

```
##    shrinkage interaction.depth n.minobsinnode n.trees      RMSE  Rsquared
## 14      0.01                 1              5     750 0.1583647 0.2140304
```

```
##        RMSESD RsquaredSD
## 14 0.01493306  0.1366213
```

Test set prediction results:

```
##      RMSE  Rsquared
## 0.2042123 0.0171706
```

The best tune has a model Rsquared of 0.2140304 but a much lower Rsquared of 0.0171706 when predicting the test set. This is likely a result of overfitting the GBM model to the training set.

According to caret function varImp, the most important variables in the FitGBM1 model are as follows:

```
varImp(FitGBM1)

## gbm variable importance
##
##                   Overall
## med_male_earnings 100.000
## pct_public_ins     75.030
## pct_black          72.510
## pct_service_ind    64.176
## pct_hisp           44.244
## pct_vacant_houses  39.285
## pct_asian          36.401
## pct_75to100K       33.716
## tot_pop            30.376
## pct_sales_office   10.321
## ownerPublic         7.231
## ownerNonprofit      0.000
```

Fitting a GBM model with only the top 5 variables above produces the following results:

```
##    n.trees interaction.depth shrinkage n.minobsinnode
## 18     950                 1      0.01              5
```

Test set prediction results:

```
##      RMSE  Rsquared
## 0.1763890 0.1029211
```

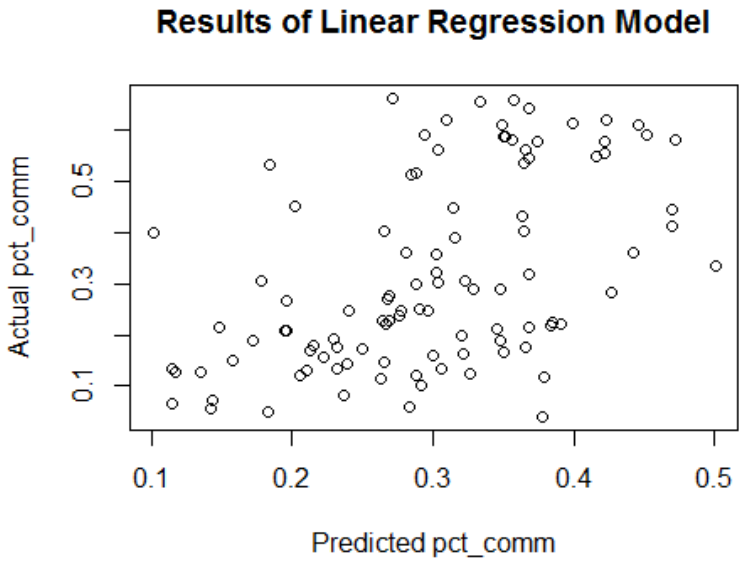
The Rsquared of the best tune increased in this model, as did the Rsquared from predicting the test set, but the testing Rsquared is still much lower than that obtained using the initial linear regression model.

Analysis of results

Our best result in predicting the percent of emergency department patients with commercial insurance was achieved with the first linear model we tested. This model

included all of the variables remaining after filtering our dataset for the predictors indicated by the regsubsets algorithm and then removing highly correlated predictors.

The following scatterplot shows the predicted and actual values of the dependent variable for each test set observation.



While the model only explains a fraction of the variance in the proportion of commercial patients at California emergency rooms, it can provide the client with a starting point of where to focus its business development and sales efforts.

Conclusion

In conclusion, we would recommend that the client take the following action as a result of this analysis:

1. Due to the limited scope of this publicly available dataset, which includes only emergency facilities in California, the client should test the model on its own facilities to see if the results are similar.

2. The client should consider whether the results pertain to all states or only those which also expanded Medicaid coverage, as California did in 2014. The client could run a similar analysis using Medicaid expansion as an additional predictor.

3. Further study should be done to determine whether the population residing in a facility's zip code is the best representation of its client base. A potential analysis could be done to match each zip code to its nearest emergency room to better capture the entire population which may present at the facility. This could improve results, given that many zip codes represent a relatively small geographic area, which may or may not include an emergency facility.