# Predicting Emergency Room Payer Mix

Emily Rinaldi

September 13, 2016

## Abstract

In this analysis, we tested the hypothesis that the demographic and economic characteristics of a particular zip code can predict the commercial patient proportion of total volume at emergency department facilities located in that zip code. We found that approximately 22.6% of the variation in commercial patient volume for the facilities in our dataset can be attributed to the demographic and economic characteristics we identified in our analysis.

The resulting model can be used by our client, an emergency department staffing company, to target its business development efforts toward those facilities located in zip codes which indicate more patients who have commercial health insurance. According to our final linear regression model, facilities with the most commercial patients tend to be privately owned or nonprofit facilities located in zip codes with higher wages, more Asian/Asian-American residents, and a large service and hospitality industry.

The analysis was performed on a dataset consisting of all emergency department facilities in the state of California for which payer mix, demographic, and economic data was available -- a total of 346 facilities. We obtained demographic and economic data on each facility's zip code from the 2010 Census and the latest American Community Survey. Because of the relatively small size of the dataset, we see this as a preliminary analysis, and our conclusions should be tested on a more complete dataset. We believe our findings prove the value of this analysis and recommend that our client purchase or create a larger dataset to test the findings and, if necessary, train a new model that is more general to the entire country.

# 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 stabilize 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.

### Final Dataset

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...
```

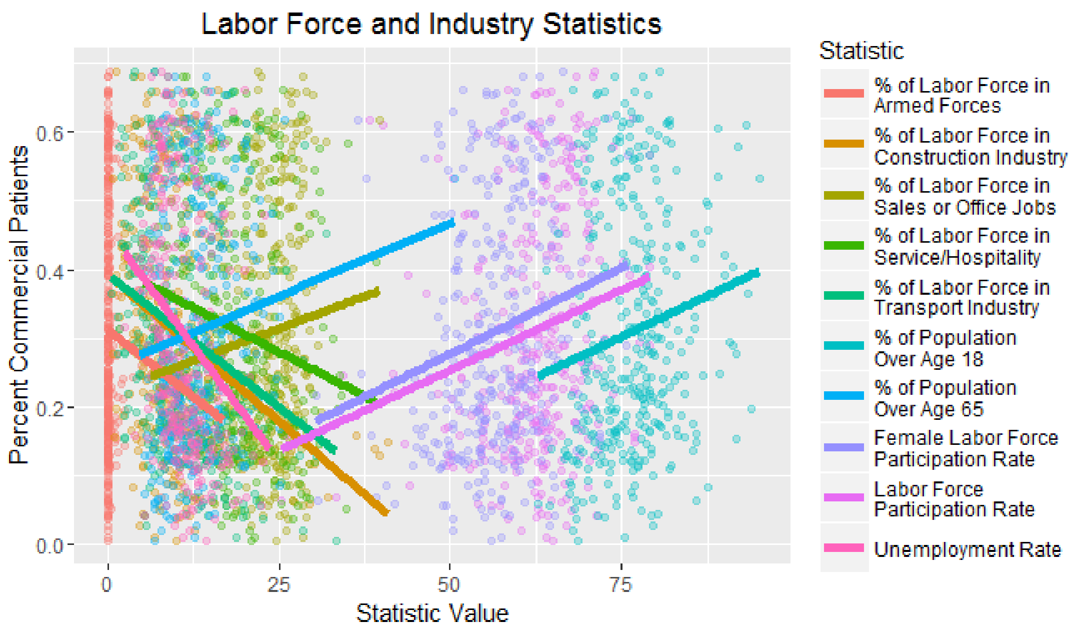
```
## $ mn_househ_income       <int> 64511, 47170, 115405, 91969, 36011, 904...
## $ pct_wSSI               <dbl> 8.2, 8.2, 3.5, 3.7, 14.6, 2.3, 3.6, 3.6...
## $ pct_wcash_assist       <dbl> 3.2, 6.5, 0.9, 2.7, 5.4, 2.2, 1.8, 1.8,...
## $ pct_SNAP               <dbl> 8.6, 13.7, 1.4, 5.3, 17.8, 1.3, 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. Each point represents a single facility's proportion of patients with commercial insurance plotted against an economic or demographic statistic for the zip code in which that facility is located.

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.

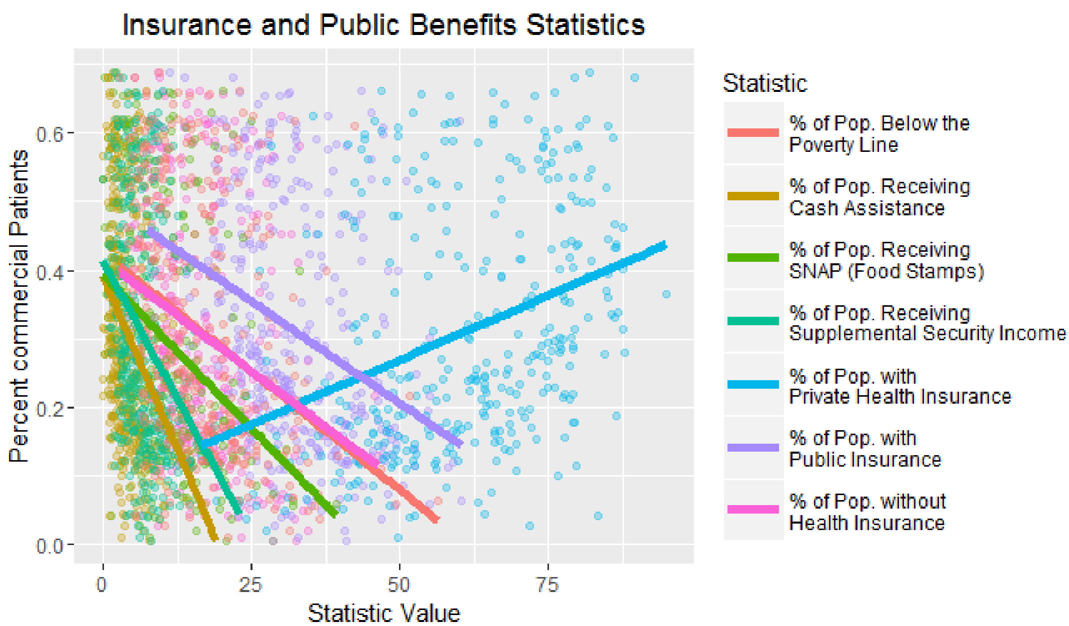
The graph below shows that a higher labor force participation rate indicates more commercial patients at emergency departments located in the same zip code. Because health insurance in the United States is so often tied to employment, more people in the labor force likely means more people with employer-sponsored health insurance. 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, likely because these types of jobs may pay more or provide better health benefits than the other industries considered.




*Correlation with % Commercial Patients:*

| | % Commercial Patients |
|---|---|
| % of Labor Force in Armed Forces | -0.0568932 |
| % of Labor Force in Construction Industry | -0.2931384 |
| Labor Force Participation Rate | 0.1998524 |
| % of Labor Force in Sales or Office Jobs | 0.0809252 |
| % of Labor Force in Service/Hospitality | -0.1690078 |
| % of Labor Force in Transport Industry | -0.2540939 |
| Female Labor Force Participation Rate | 0.1902173 |
| % of Population Over Age 18 | 0.1570656 |
| % of Population Over Age 65 | 0.1242494 |
| Unemployment Rate | -0.2823910 |

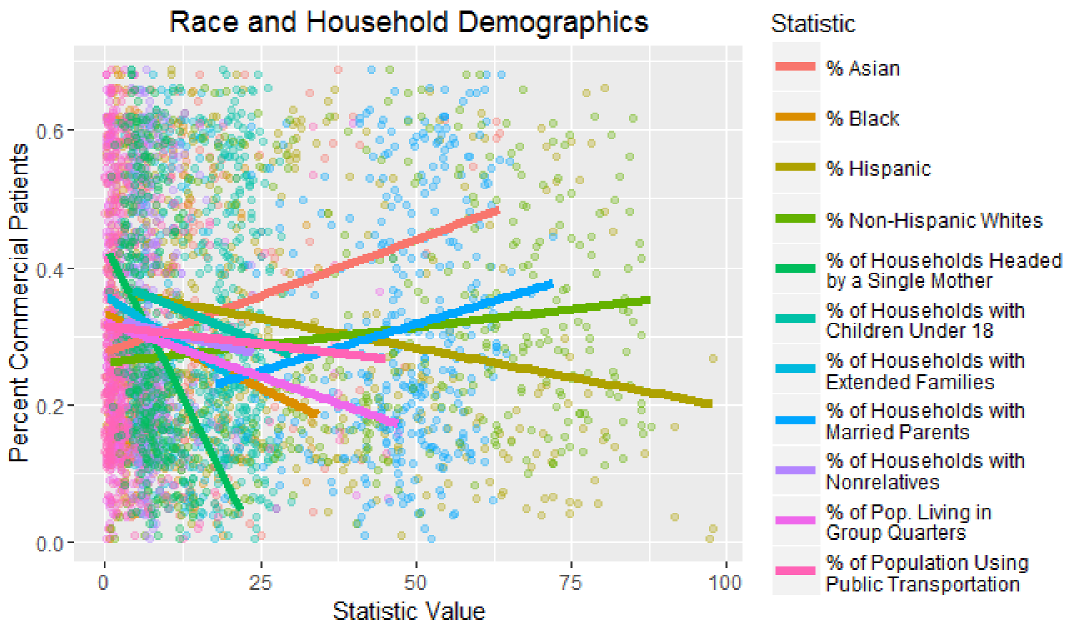
The following graph shows that more people on public assistance indicates fewer commercial patients at facilities in the same area. People who are eligible for public assistance may not be employed or may have incomes low enough to be eligible for Medicaid, in which case they would not need commercial health insurance. 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, which is the entire basis of our analysis.




*Correlation with % Commercial Patients:*

| | % Commercial Patients |
|---|---|
| % of Pop. without Health Insurance | -0.2655239 |
| % of Pop. with Public Insurance | -0.3270782 |
| % of Pop. with Private Health Insurance | 0.3405461 |
| % of Pop. Receiving SNAP (Food Stamps) | -0.3484687 |
| % of Pop. Receiving Cash Assistance | -0.3344520 |
| % of Pop. Receiving Supplemental Security Income | -0.3029456 |
| % of Pop. Below the Poverty Line | -0.3487996 |

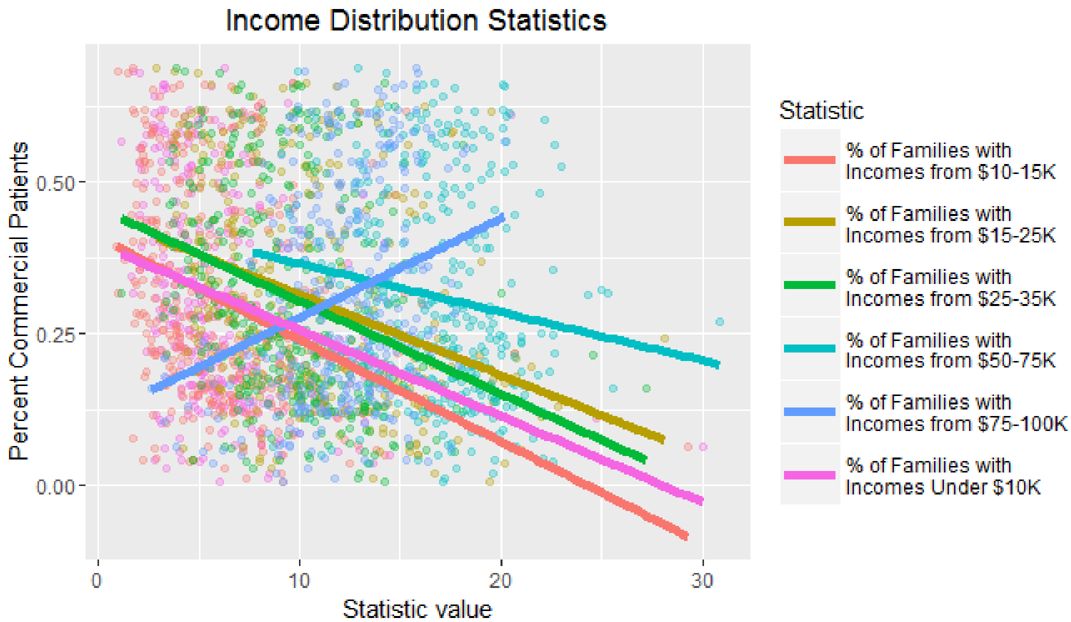
The Race and Household Demographics graph below shows that populations with more Asians, and to a lesser extent more whites, may have more commercially insured individuals. Also, it appears that the proportion of family households headed by a single mother may have a strong inverse relationship with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




*Correlation with % Commercial Patients:*

| | % Commercial Patients |
|---|---|
| % Asian | 0.2069825 |
| % Black | -0.1495464 |
| % of Households with Extended Families | -0.1388121 |
| % of Pop. Living in Group Quarters | -0.1149497 |
| % Hispanic | -0.2211891 |
| % of Households with Children Under 18 | -0.0946915 |
| % of Households with Married Parents | 0.1615811 |
| % Non-Hispanic Whites | 0.1466676 |
| % of Households with Nonrelatives | -0.0401824 |
| % of Population Using Public Transportation | -0.0426155 |
| % of Households Headed by a Single Mother | -0.3060861 |

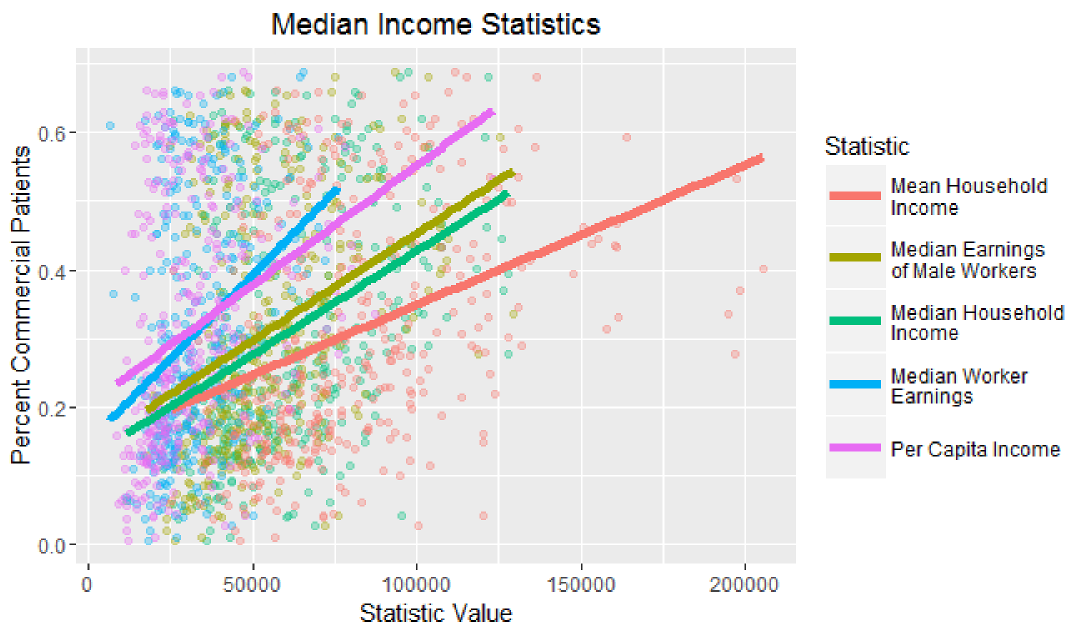
When observing the income distributions of a population, a positive correlation to commercially insured patients is not reached until household incomes exceed $75,000.




*Correlation with % Commercial Patients:*

| | % Commercial Patients |
|---|---|
| % of Families with Incomes Under $10K | -0.2714347 |
| % of Families with Incomes from $10-15K | -0.3084706 |
| % of Families with Incomes from $15-25K | -0.3319653 |
| % of Families with Incomes from $25-35K | -0.3061599 |
| % of Families with Incomes from $50-75K | -0.1425221 |
| % of Families with Incomes from $75-100K | 0.2632964 |

As expected, higher median incomes for an area indicate more ER patients with commercial insurance. Higher wages likely means more people that can afford the health benefits offered by their employers and fewer workers eligible for Medicaid.




*Correlation with % Commercial Patients:*

| | % Commercial Patients |
|---|---|
| Median Household Income | 0.3798722 |
| Median Earnings of Male Workers | 0.3469115 |
| Median Worker Earnings | 0.3352430 |
| Mean Household Income | 0.3539993 |
| Per Capita Income | 0.2999576 |

## 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 $C_p$, a metric that is less biased toward overfitting by adding more variables than other model performance metrics. We then filtered the dataset to only those variables which are included in the best model. Finally, we added dummy variables for the factor values of owner, which was one of the predictors chosen by regsubsets.


The model that minimizes Mallow's $C_p$ 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 Righly 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.

```
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.

```
set.seed(50)
inTraining <- createDataPartition(model_filtered3$pct_comm, p = .7, list =
FALSE)
training <- model_filtered3[inTraining,]
testing <- model_filtered3[-inTraining,]
```

## Impute Missing Values

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 $R^2$ is similar to the model $R^2$, 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 $R^2$ and the Adjusted $R^2$. It also resulted in a lower $R^2$ 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 $R^2$ 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 $R^2$ of 0.214 but a much lower $R^2$ of 0.017 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:

```
##    shrinkage interaction.depth n.minobsinnode n.trees      RMSE  Rsquared
## 18      0.01                 1              5     950 0.1566044 0.2273668
##        RMSESD RsquaredSD
## 18 0.01658431  0.1453709
```

Test set prediction results:

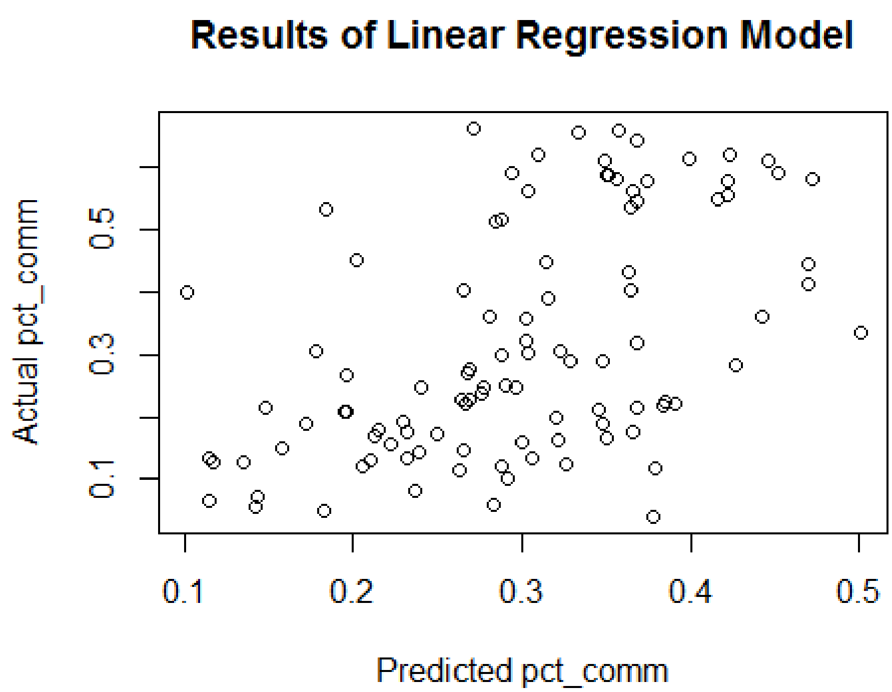
```
##      RMSE  Rsquared
## 0.1763890 0.1029211
```

The $R^2$ of the best tune increased in this model, as did the $R^2$ from predicting the test set, but the testing $R^2$ 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.

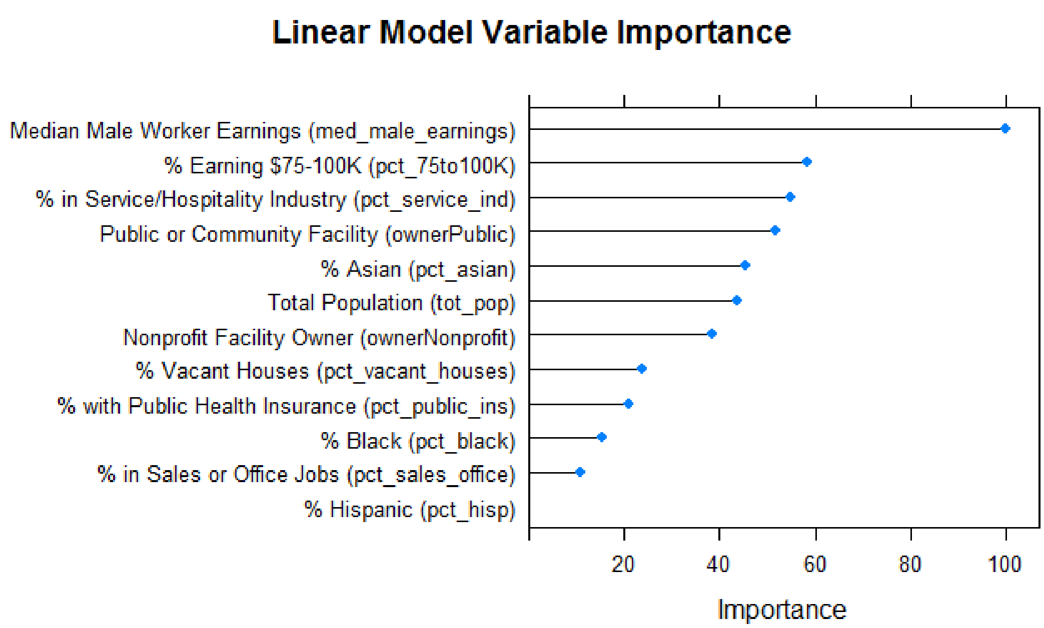



At the cross validated $R^2$ of 0.226, the model only explains a fraction of the variance in the proportion of commercial patients at California emergency rooms. Even so, it can provide the client with a starting point of where to focus its business development and sales efforts.

According to the model, facilities with the most commercial patients are likely privately owned or nonprofit facilities located in zip codes with high wages, more Asian/Asian-American residents, and a large service and hospitality industry. The following coefficient summary and variable importance chart show the impact each variable has on the final model:

*Linear Model Coefficient Summary:*

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -0.0689347 | 0.1603153 | -0.4299944 | 0.6676006 |
| ownerNonprofit | -0.0478559 | 0.0272140 | -1.7585038 | 0.0799860 |
| ownerPublic | -0.0795775 | 0.0363282 | -2.1905143 | 0.0294871 |
| pct_service_ind | 0.0056271 | 0.0024576 | 2.2896805 | 0.0229426 |
| pct_sales_office | 0.0023551 | 0.0027397 | 0.8596160 | 0.3908920 |
| pct_75to100K | 0.0105072 | 0.0043663 | 2.4064234 | 0.0168961 |
| med_male_earnings | 0.0000036 | 0.0000010 | 3.7537581 | 0.0002204 |
| pct_public_ins | -0.0019233 | 0.0016176 | -1.1889821 | 0.2356675 |
| tot_pop | -0.0000012 | 0.0000006 | -1.9249413 | 0.0554648 |
| pct_black | -0.0018589 | 0.0018368 | -1.0120660 | 0.3125654 |
| pct_asian | 0.0020633 | 0.0010383 | 1.9871357 | 0.0480887 |
| pct_hisp | 0.0003600 | 0.0007148 | 0.5036645 | 0.6149771 |
| pct_vacant_houses | 0.0016752 | 0.0013119 | 1.2769683 | 0.2028955 |




***Median Male Earnings:*** The most significant predictor is the median earnings of male workers - the positive coefficient on this variable indicates facilities in zip codes with higher median male earnings have more commercial patients. Similarly, the proportion of workers earning $75,000 to 100,000 also indicates more commercial patients. This result is not surprising - more people in an area with higher paying jobs likely means more people who receive and are able to afford the health benefits offered by their employers.

***Service or Hospitality Industry Workers:*** An initially counter-intuitive predictor of higher commercial patient volume is the percent of the labor force employed in the service or hospitality industry. More workers in an industry that is not known for high wages or great benefits would seem to predict fewer emergency department patients with commercial health insurance; instead, the opposite is true. It is possible that the proportion of service industry workers tells us more about other residents in the area than about those indicated by the statistic. Perhaps zip codes with vibrant restaurant and hotel industries are desirable areas for higher-income individuals and families to reside, and it is these residents presenting at emergency departments with commercial insurance.

***Facility Ownership:*** The only predictor in our model that is related to the facility itself rather than the zip code in which it is located is the ownership of the facility. The model's baseline identifies an investor-owned facility, while non-profit and public ownership each negatively affect the expected commercial patient volume compared to that baseline. Public or community hospitals tend to serve a lower-income population and, along with non-profit hospitals, may have more generous charity policies. Uninsured patients with less-emergent health issues may decide to visit these emergency departments rather than privately owned facilities located closer to their homes if they are concerned about the cost of services.

***Asian Population:*** Another significant predictor is the proportion of the population that self-identifies as Asian or Asian-American. In this model, a higher proportion of Asians indicates more commercial patients. According to the 2010 Census, 30.9% of Asians or Asian-Americans in the United States reside in the state of California. As recommended in our conclusion, further study should be done to determine whether all of our findings still stand when applied to facilities outside of California.

## Conclusion

We would recommend that the client take the following action as a result of this analysis:

1. Due to the limited scope and small sample size of this publicly available dataset, which includes only emergency facilities in California, the client should identify a larger, more complete dataset of emergency department facilities in the United States. A dataset may be available for purchase, or the client could create its own dataset using data from its own client facilities. The client should test the current model's predictions on this larger dataset and, if necessary, train a new predictive model using the larger dataset.

2. The client should consider whether the results of this or any predictive model 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. In states that expanded Medicaid, it is possible that a significant number of workers who are offered commercial health benefits through their employer could be eligible for and elect to enroll in Medicaid instead. This would replace commercial patient volume with Medicaid volume at emergency department facilities.

3. Further study should be done to determine whether the population residing in a facility's zip code is the best available representation of its patient 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.