Final Project
STAT 515 01
13 May 2019
Rahul Sonti

Simulated total medical expenses charged to insurance in 2013

Introduction

The data set our group chose to evaluate is a simulated data set that contains total medical expenses charged to the insurance company and various patient characteristics. These data were based off the U.S. Census Bureau demographic statistics in 2013. The data source was found on Zach Stednick's GitHub[11] and it was in .csv format. Inpatient data was simulated to depict real world circumstances for 1,338 beneficiaries in the U.S. that are enrolled in the insurance plan.

The data set contained over 1,338 observations and 7 different fields. These fields recorded the age of the beneficiary, the sex of the beneficiary, tobacco smoking habits, the BMI of the beneficiary, the number of children dependents, the region of the United States that the beneficiary lives in and the total medical expenses charged to the insurance company.

Data Definitions[1]

Age refers to the age of the primary beneficiary.

Sex refers to the insurance contractor gender, either female or male.

Smoker refers to if the beneficiary is a tobacco smoker or not.

BMI is the body mass index; it provides an understanding of the body. Objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.4 to 24.9 in adults.

Children refers to the number of dependents or children that are covered by health insurance.

Region is the beneficiary's residential area inside the United States; these include northwest, northeast, southeast, and southwest.

Charges refers to the individual medical costs billed by the health insurance company.

Questions and Analysis

With this supplementary information, we considered the relevant questions we could ask that the insurance company would be most interested in. Predicting the average medical costs for various demographics in the population is the most sought out piece of information for the insurance company. Using the associated categorical fields (like smoker, sex and geographic region) and the numeric fields (like age, BMI and children), we hypothesize that we can make these predictions. **We want to know "Can this data help us predict average medical costs?"**

To begin, we conducted a preliminary analysis of the data in R[9] to determine what possible approaches would be most appropriate. There were no missing values in the data set, or any additional formatting needed to pre-process the data set. Using a summary view, we were able to quickly see the spread of the data. The preliminary analysis was conducted using the tidyverse[6], caret[8] and Rcpp[1,2,3] packages.

The total medical charges have a minimum of $1,122, a mean of $13,270, and a maximum of $63,770. There are 662 females and 676 males. The BMI of the beneficiary had a minimum of 15.96, a mean of 30.66 and a maximum of 52.13. There are 1,064 non-smokers and 274 smokers in the data set. The number of observations in each region of the United States is fairly equal; 324 in the northeast, 325 in the northwest, 364 in the southeast and 325 in the southwest. Lastly, the minimum age is 18, the mean age is 39.21 and the maximum age is 64.

In order to go a little deeper in the examination of the categorical data, we created various bar charts [Fig. 1] with ggplot2[5] to explore the frequency of the gender, sex and smoking habits of the beneficiary. The underrepresentation of smokers may mean our final model may not be able to predict them as accurately as non-smokers due to the lack of data to draw on.

For the numeric data, we used histograms to get an idea of the distribution of this data [Fig. 2]. The age and gender of the beneficiary were mostly uniformly distributed. The distribution of total charges is skewed right. We may want to focus on the data only towards the left of the chart or consider different probability distributions other than normal to correctly fit this data. Moreover, the distribution of BMI of the beneficiary is bell curve shaped. We also attempted to plot the numeric data against each other to understand if there were any relationships that could be easily spotted [Fig. 3]. Comparing total charges versus the age of the beneficiary gave a clear vertical pattern, hinting that there may be a positive relationship between total charges and age. When assessing the relationship between BMI and charges, it is a bit more random and not as intuitive. It seems that there may be a positive relationship between BMI and total charges, but further investigation is needed. Lastly, the relationship between children and total charges is quite horizontal. From the summary plot, it seems that there are higher total charges the less children that you have, which is counterintuitive. However, this may be because we have many more data points for 0, 1, 2 and 3 children than we do for 4 or 5 children.

These findings helped guide our decision on the types of models we wanted to use to predict our desired variables. We chose to focus on the unsupervised learning technique principal component analysis (PCA) to further examine the relationships between the variables and the supervised learning techniques of regression decision trees, random forest, and logistic regression to predict the outcomes we were interested in.

Fig. 1 Bar charts that illustrates frequency of categorical variable; Top left: Region of residency of beneficiary. Top Right: Gender of beneficiary. Bottom Center: Smoking habits of beneficiary.

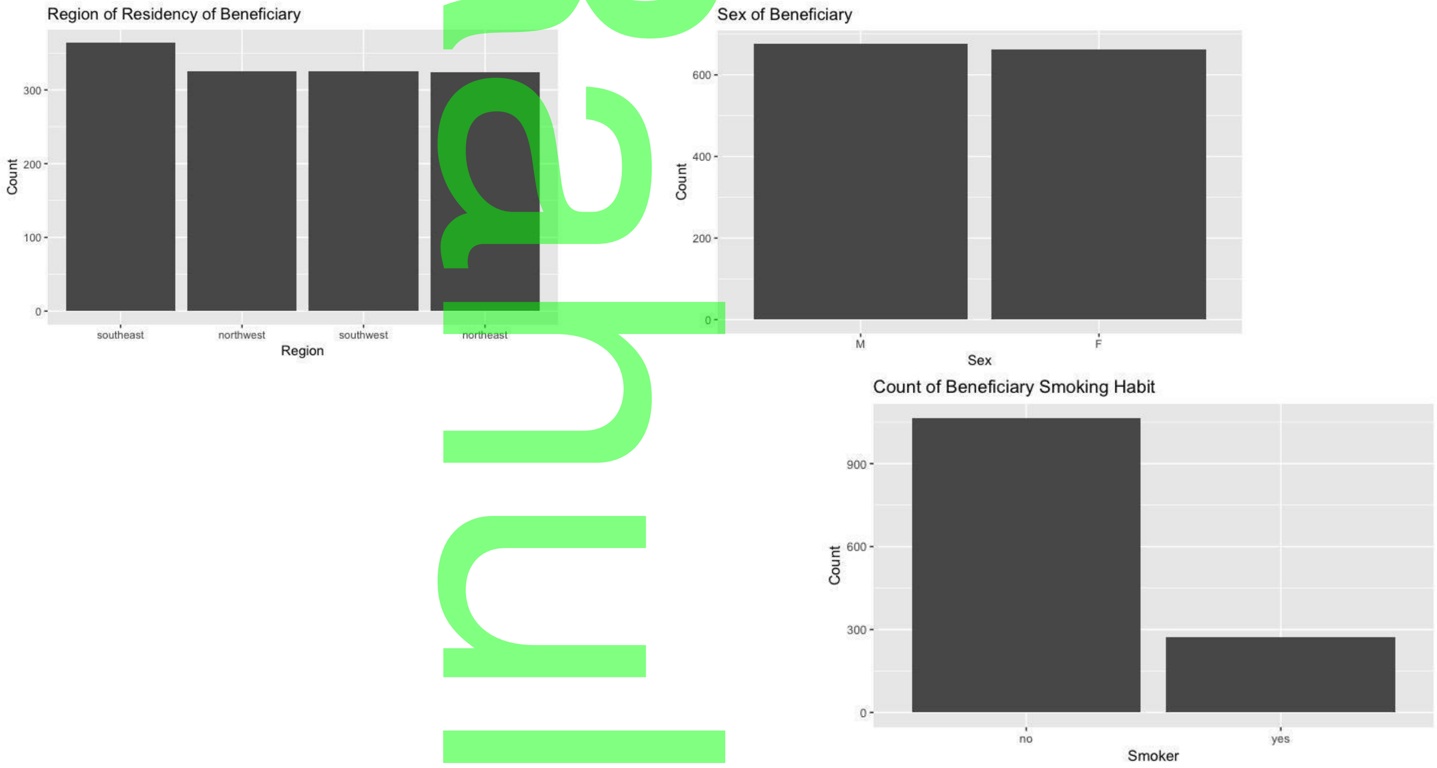


Fig. 2 Histograms that illustrate the frequency distribution of the total charges, age of beneficiary, BMI of beneficiary and number of children or dependents on the insurance plan (left to right, top to bottom).

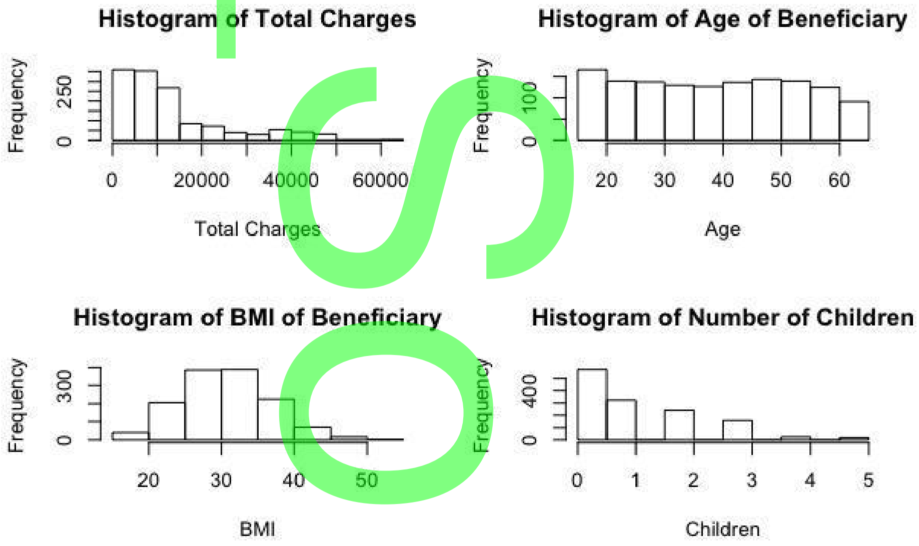
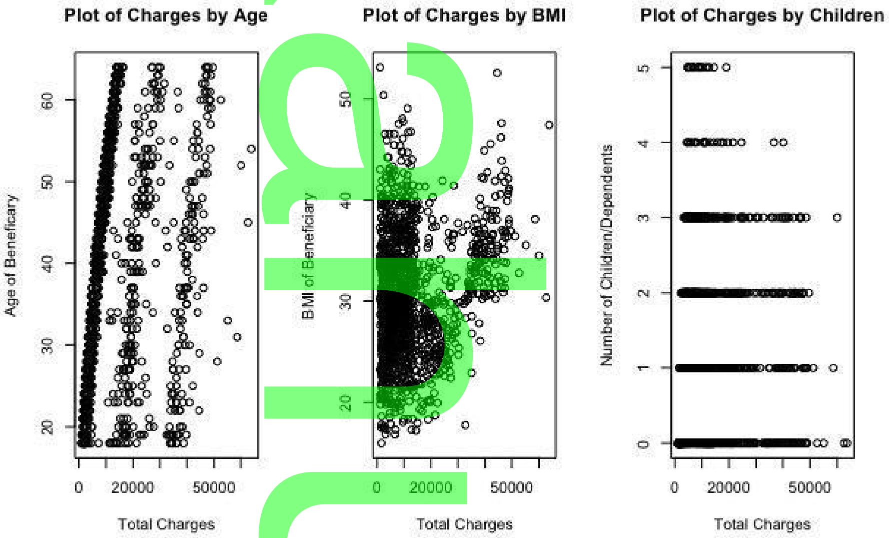


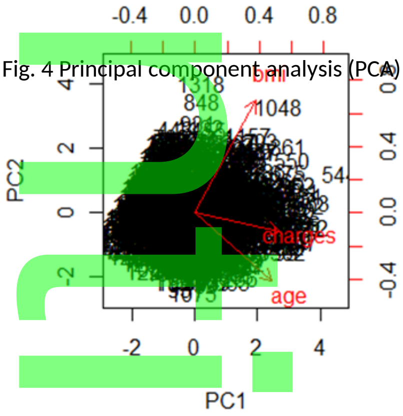
Fig. 3 Plots that compare total charges to age of beneficiary, BMI of beneficiary, and number of children or dependents on the insurance plan (left to right).



Principal Component Analysis (PCA)

We used unsupervised learning and conducted a principal component analysis (PCA)[4,5] on the insurance data set (Fig. 4) to explore the variables before beginning supervised learning. We completed the PCA by omitting the sex, region, smoker, and number children variables as there were only limited number of values that they can take. We performed PCA by considering age, charges, and body mass index (BMI). From the bi-plot we can see that age and charges have large positive loadings the first principal component (PC1), so this tells us that as the age increases the insurance charges that are billed are also increasing. The BMI has a positive loading towards the first principal components, meaning that the insurance charges mostly increase when the BMI increases. We can also see that the charges and ages are negatively correlated with the principal component 2 (PC2) which means that there are less people of high age who pay insurances with respect to the other ages and the BMI is positively correlated with the principal component 2. Observations 544, 1318, and 848 appear to be outliers.

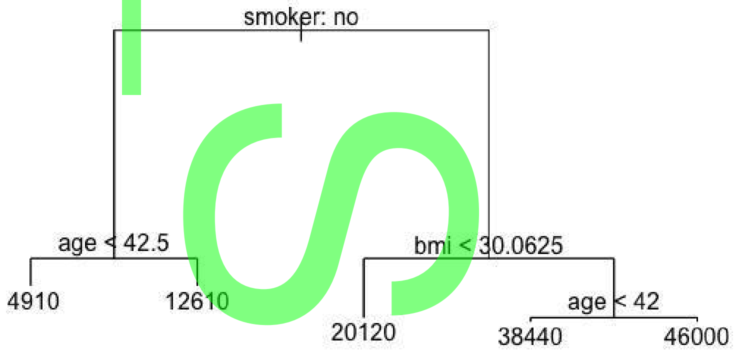
Fig. 4 Principal component analysis (PCA)



Regression Decision Tree

One approach to answering our questions of can we predict total charges is by using a regression decision tree in the MASS[12] and tree[10] package to predict the average total charges for an individual. This analysis was chosen because it has the ability to predict numeric values and can capture non-linear relationships better than a linear model. There was no manipulation to the data set prior to analysis. The data set was broken up into a training data set and a testing data set. A tree (Fig. 5) was created using the training data set and cross validation was conducted. The cross validation determined that the tree did not need to be pruned. Therefore, a prediction was made on the test set using the unpruned tree. The mean squared error (MSE) for the regression decision tree is 5,026.48, meaning that the prediction of the actual total insurance charges could vary by $5,026.48.

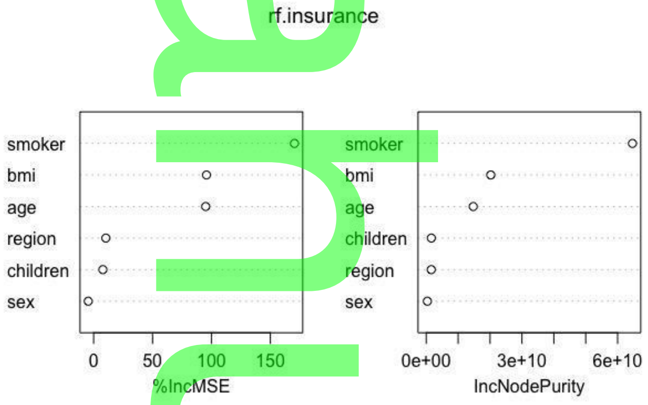
Fig. 5 Regression Decision Tree



Random Forest

To take it a step further, a random forest was constructed in the randomForest[7] package to see if the accuracy of predicting the actual total insurance charges could be increased. Random forest tends to be more accurate because it builds numerous decision trees on the bootstrapped training data set. Additionally, while building the forest, the algorithm decorrelates each tree by using a random set of the variables to build each tree. This distinct difference stops strong variables from being the top split in every tree and all of the trees being extremely correlated. To begin the analysis, the data frame was again split into a training and test data set. The MSE of the random forest is 4,709.07, meaning that the prediction for the actual total insurance charges can vary by $4,709.07. Lastly, an importance plot (Fig 6.) was created to detail which variables are the most influential to changing the MSE of the prediction. The most influential variables were smoking habits, the BMI, and the age of the beneficiary.

Fig. 6 Importance variable plot from Random Forest.



Logistic Regression

Our final strategy was to employ logistic mixed effects regression modeling[13] and categorize insurance charges in an alternate way. The packages used in developing this model were tinyverse[6], ggplot2[4], and Sleuth3[14]. Logistic regression is performed for binomially distributed dependent variables. The model predicts the outcome of a dichotomous dependent variable in terms of log odds as a linear combination of a set of independent variables. In the current model, the dichotomous variable is created for insurance charges which are greater than 80 percent of the total sample charges, approximately $20,000. This different bucketing of insurance charges will help us examine the more extreme cases in our skewed dataset. This dependent variable is checked with log odds as a linear combination of Age, Sex, Smoker, BMI, Children and Region.

The model results showed that age, BMI, and smoker status (where smoker is “yes) are significant variables affecting the charges, all with p values under .0001. The predicted insurance charges greater than $20,000 are a total of 128 observations and less than $20,000 are 446 observations. The predicted charges with error for being more than $20,000 are a total of 28 observations while the predicted charges for being less than $20,000 are a total of 20 observations. The MSE for the model is 0.083 and the accuracy of the prediction is 0.916.

Conclusion

Our goal was to answer the question “Can we predict charges to the insurance company?” By using a combination of summary statistics, data exploration with unsupervised learning, and prediction with supervised learning, we were able to provide insights on the variables surrounding this question and create a model that would help us answer this question within a certain tolerance for error.

Our summary examination revealed skews in the insurance charge amounts and the number of children. A skew in the discrete count of children dependencies isn’t concerning and this field would likely only be used as a categorical bin for decision trees. We found out smokers tended to be underrepresented in the data set, which originally caused us to consider them a more difficult group to predict. However, this variable ended up being highly influential and a good decision tree branch. Using PCA helped us confirm variables of interest as they relate to charges, including age and BMI.

The first model we chose was a decision tree, due to its success at using both categorical and numeric data to predict a numeric result. This suited our data well due to the combination of patient information, including demographics and health metrics, alongside our desired variable charges. The model could predict results with an MSE of 5,026.48, insinuating a possible variability of $5,026.48 in our final charge prediction. Since the mean charge was $13,270, we figured reducing this error would be in our best interest. We followed up with the random forest model, which reduces error by utilizing multiple combined decision trees with randomly bootstrapped data and randomly selected predictor variables. Our MSE improved to 4,709.07. Finally, our logistic regression model helped us identify insurance charges above and below $20,000 with over 90% accuracy.

By reviewing the data set, identifying important variables, and attempting multiple models, we decided that the random forest model would be the best at answering our final question and predicting what the average charge will be for the insurance company based on the accompanying patient information. PCA allowed us to examine related variables and logistic regression helped us categorize and predict insurance charges in the 80th percentile of our sample. Potential room for improvement could be to acquire more smoker data, which could possibly help us retrain the model to better identify smokers and non-smoker insurance chargers and reduce MSE.

References

1. Dirk Eddelbuettel and Romain Francois (2011). Rcpp: Seamless R and C++ Integration. Journal of Statistical Software, 40(8), 1-18. URL http://www.jstatsoft.org/v40/i08/.
2. Dirk Eddelbuettel and James Joseph Balamuta (2017). Extending R with C++: A Brief Introduction to Rcpp. PeerJ Preprints 5:e3188v1. URL https://doi.org/10.7287/peerj.preprints.3188v1.
3. Dirk Eddelbuettel (2013) Seamless R and C++ Integration with Rcpp. Springer, New York. ISBN 978-1-4614-6867-7.
4. Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
5. Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse
6. A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.
7. Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2019). caret: Classification and Regression Training. R package version 6.0-84. https://CRAN.R-project.org/package=caret
8. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
9. Brian Ripley (2018). tree: Classification and Regression Trees. R package version 1.0-39. https:// CRAN.R-project.org/package=tree
10. Stednick, Z., Caballero, N., & Zanoni, R. (2015, February 18). Stedy/Machine-Learning-with-R-datasets. Retrieved May 6, 2019, from https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/insurance.csv
11. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
12. UCLA. *LOGIT REGRESSION | R DATA ANALYSIS EXAMPLES*. n.d. 11 May 2019. <https://stats.idre.ucla.edu/r/dae/logit-regression/>.
13. Ramsey , F. L., & Schafer, D. W. (2013). Sleuth3: Data Sets from Ramsey and Schafer's "Statistical Sleuth (3rd Ed)." Retrieved from Sleuth3: Data Sets from Ramsey and Schafer's "Statistical Sleuth (3rd Ed)" website: https://cran.r-project.org/web/packages/Sleuth3/index.html