Individual Health Insurance Forcast Report

Jiaying Ning, Meiju Chen, Qetsiyah Wang

5/11/2021

Contents

Introduction **2**

Data **2**

Explanatory Visualization Analysis **3**

Regression for the Repsonse Medical Costs **3**

Clustering Analysis **4**

Conclusion **4**

Introduction

Health insurance may be expensive. However, from the individual point of view, health insurances lessen the costs of unexpected high medical expenses, and provide preventive healthcare such as routine checkups and screening tests. On the other hand, from business point of view, health insurance companies make more profit by collecting more beneficiaries and lower medical spends on them. We might agree on some conditions are more prevalent for certain populations, however, medical spends are still difficult to predict. Thus, no matter what position we are in, customers or business owner, it is essential for us to understand what factors affect the charges to make informed decision.

In this report, we will predict insurance costs based on people's data, including age, gender, Body Mass Index(BMI), smoking status, etc. In addition, we will find predictive variables that influence individual medical costs billed by health insurance. Besides regression model, we also explore potential subgroup for deeply analyzing the association of medical costs with other variables.

Data

This dataset includes following variables:

- `age` - age of primary beneficiary
- `sex` - insurance contractor gender, female/ male
- `bmi` - Body Mass Index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- `children` - Number of children covered by health insurance / Number of dependents
- `smoker` - Smoking status, yes/no
- `region` - the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- `charges` - Individual medical costs billed by health insurance

```
##       age            sex           bmi           children     smoker
##  Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064
##  1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274
##  Median :39.00                Median :30.40   Median :1.000
##  Mean   :39.21                Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                Max.   :53.13   Max.   :5.000
##        region       charges
##  northeast:324   Min.   : 1122
##  northwest:325   1st Qu.: 4740
##  southeast:364   Median : 9382
##  southwest:325   Mean   :13270
##                  3rd Qu.:16640
##                  Max.   :63770
```

For categorical variables, the dataset includes similar number of people for each category, except for `smoker`. Around 20% of smokers and 80% of non-smokers. This make sense, since in the real world situation, non-smokers are way more that smokers. In this report, we are interesting in the health insurance cost, thus, the target feature will be the variable: `charges`. From the table above, `charges` ranges from around 1,000 to 64,000.

Explanatory Visualization Analysis

Scatter plots explore the relationships between the response `charges` and all the other variables. In Figure.1, `age` shows some potentially linear trend, smoking status(`smoker`) also affect individual medical cost. Gender(`sex`), Body Mass Index(`bmi`) and numbers of children(`children`) might also influence the medical cost but, comparing to the two former variables, they affected the cost more slightly. As for the resident area in the US(`region`), we cannot see significant differences.

To look for further information, we can dig deeper into categorical variables by using box plots. In Figure.2a, while there is a slightly increasing trend in charges as the `children` increases, `sex` and `region` have no noticeable differences in charges for each category. In addition, smoking status(`smoker`) has a significant difference to individual medical charges.

Since smoking status(`smoker`) influence charges in a significant way, we can take a look at the distribution of charges, categorizing them into smokers and non-smokers. In Figure.2b, the original distribution of `charges` , with no categorizing, is right skewed with a long tail to the right. There's a bump at around $40,000. After we categorizing charges into smokers and non-smokers, we can see the clear difference between two populations. Thus, we can say that smokers tends to have more individual medical charges than non-smokers.

Regression for the Repsonse Medical Costs

Linear Regression, Ridge and Lasso Regression

Five types of linear models would be utilized to predict 'charges'. 75% partition was performed to split data into train and test. Before establishing regression models, normality of data was firstly identified. Since the current distribution for outcome was skewed to the right, transformation method is recommended. Box-cox was performed to select the best transformation method.

In Figure.3 and Table.1, we observed that among all the significant predictors, age, bmi, children, and smoker-yes seemed to be the most significant. And smoker-yes status had the highest coefficient. Comparing to people who do not smoke, people who smoked tended to have an increase in 4 units in charges on average. In ridge regression, the best lambda was 0.0084. Final test error was 1.224. In lasso regression, the best tuning parameter was 3×10^{-4} and the test error was 1.2251. Lasso kept all variable for the final best optimal model. The differences in coefficients between models were subtle, meaning the effect of regularization method on linear regression was small for the current data, linear model was appropriate enough in this case. The difference in RMSE is also not dramatic. We had lasso with the smallest mean RMSE, meaning that lasso had relatively better fit.

Tree Regression and Random Forest

In Figure.4a and Figure.4b, we observed that a terminal nodes contained mean charges of 47000 with 4% of subjects. This meant that 4% of data fall into rectangle that have the mean charge of 47000 This rectangle contained the highest average mean of log value, and it was determined by those who did not smoke and had bmi <30 and age <45. The smallest mean charges was 2876 with 16% of subjects falling in to this category. This rectangle also contained the most subject, and it was determined by those who smoke, having age <33 and children <1.

When using the lowest cross-validation error, we obtained tree size of 12, and when applying 1se rule, we obtain tree size of 8. In here, the highest mean charges (47000)is determined by those who did not smoke and have bmi <30 and age <45, which is the same as the result obtained from the previous tree plot.

In the random forest regression shown in Figure.4c, similar to previous result, smoker is the most important predictor in predicting charges.

Clustering Analysis

Besides Linear Regression Model, we also explored homogeneity and located potential subgroups among observations for Age, BMI, Number of Children and corresponding Medical Charges. Within different subgroups, what statistical dissimilarities did each categorical indicator present? We would utilize K-Means CLustering for determining subgroups. The optimal cluster number was pre-identified as 4, shown as the Figure below.

Shown in Figure.5b, desired clustering results were clearly obtained for Children vs. Charges. There was no explicit subgroups observed within Age vs.BMI, meaning that no underlying relationship between BMI and age. Same conclusion could be addressed for BMI vs.Children. For Age vs.Charges or BMI vs.Charges, at low medical charges, three clusters got intertwined. The black subgroups at high medical charges was grouping explicitly, which would be considered as a desired clustering result. However, in Age vs.BMI, the black cluster showed the clustering across the whole age range, which was much too wide to initiate meaningful statistical analysis. In other words, there was no dissimilarity existed across all age at high medical charges. The black cluster in BMI vs.Charges could be regarded as more optimal clustering result even showing some outliers. Further analysis would focus on Children vs. Charges and BMI vs. Charges.

BMI vs. Medical Charges

Shown in the Figure.6a, cluster 1, at high medical charge and high BMI, presented high inter-dissimilarity with other clusters, even some outliers existed. People with high BMI tended to have high medical charges but people with low or mediate BMI did not indicate the association as explicit as high BMI. The silhouette score was 0.5813. While considering smoking status within BMI vs. Charges in subgroup 1, the association between BMI and medical charges was largely affected by positive smoking status, observing from the Figure.6d. For BMI that was large than 28, positive smoking would incline to get significant higher medical charge than negative smoking. The difference of median between two smoking status was -12093.05. For other two categorical indicators, sex and region, there was no significant impacts observed on the cluster 1 within BMI vs. Charges, meaning that the association between high BMI and high medical charges did not show heterogeneity across different regions and sex groups.

Number of Children vs. Medical Charges

Presented in Figure.7a, cluster 1 and 3 both could be considered as optimal clustering results with silhouette scores of 0.4409 and 0.4573. Other two clusters showed large overlapped clustering region, which would not be utilized for further analysis. Generally, two clustering result indicated that as number of children increased, the medical charge decreased. Considering the smoking status on analyzing associations between two clusters, same conclusion could be addressed for Children vs.Charges as BMI vs.Charges, that positive smoking status would be claimed with higher medical charges both for low and high number of children each person had. Noticeably, within the subgroup 1, the median difference of medical charges was 12093.05, which was lower than that of the subgroup 3, 13112.08. In other words, at high level of number of children, positive smoking status would indicate impact the medical claim charges stronger than low level of number of children. Furthermore, same observation was obtained for sex indicator. In Figure.7d, obtaining from more significant gap on medical charge between two sex groups in the subgroup 3 than the subgroup 1, female with high number of children tended to be charged higher than female with low number of children.

Conclusion

From visualization analysis, smoking Status presented strong association with medical charges. Explicit associations against charges were hard to be determined from Age and BMI. Boxplots of Sex and Region did not illustrate statistical significant associations with charges. Because the distribution of charges was not normal, the response was implemented with boxcox transformation. From ordinal linear regression, ridge regression and lasso regression model for fitting and predicting charges with all other variables, three regression models did not represent significant differences from RMSE analysis. Lasso kept all variables in the final optimal model. Noticeably, the estimated coefficient of smoking status was much higher than other

predictors, which was consistent with the result from explanatory visualization. Same conclusion could be made from Regression Tree and Random Forests. K-Means Clustering located subgroups for BMI vs.Charges, that high BMI groups tended to be grouped with high medical charges, for Children vs.Charges, low amount of children tended to be grouped with high medical charges. Furthermore, positive smoking status presented significant impacts on associations with Charges in these subgroups.

Table 1: Estimated Coefficients for Three Regression Models

	OLS	Ridge	Lasso
(Intercept)	9.8933	9.9094	9.8946
age	0.0840	0.0837	0.0840
sexmale	-0.1897	-0.1880	-0.1891
bmi	0.0324	0.0323	0.0323
children	0.2525	0.2518	0.2523
smokeryes	3.9989	3.9838	3.9981
regionnorthwest	-0.1797	-0.1769	-0.1776
regionsoutheast	-0.3992	-0.3942	-0.3968
regionsouthwest	-0.4079	-0.4038	-0.4057

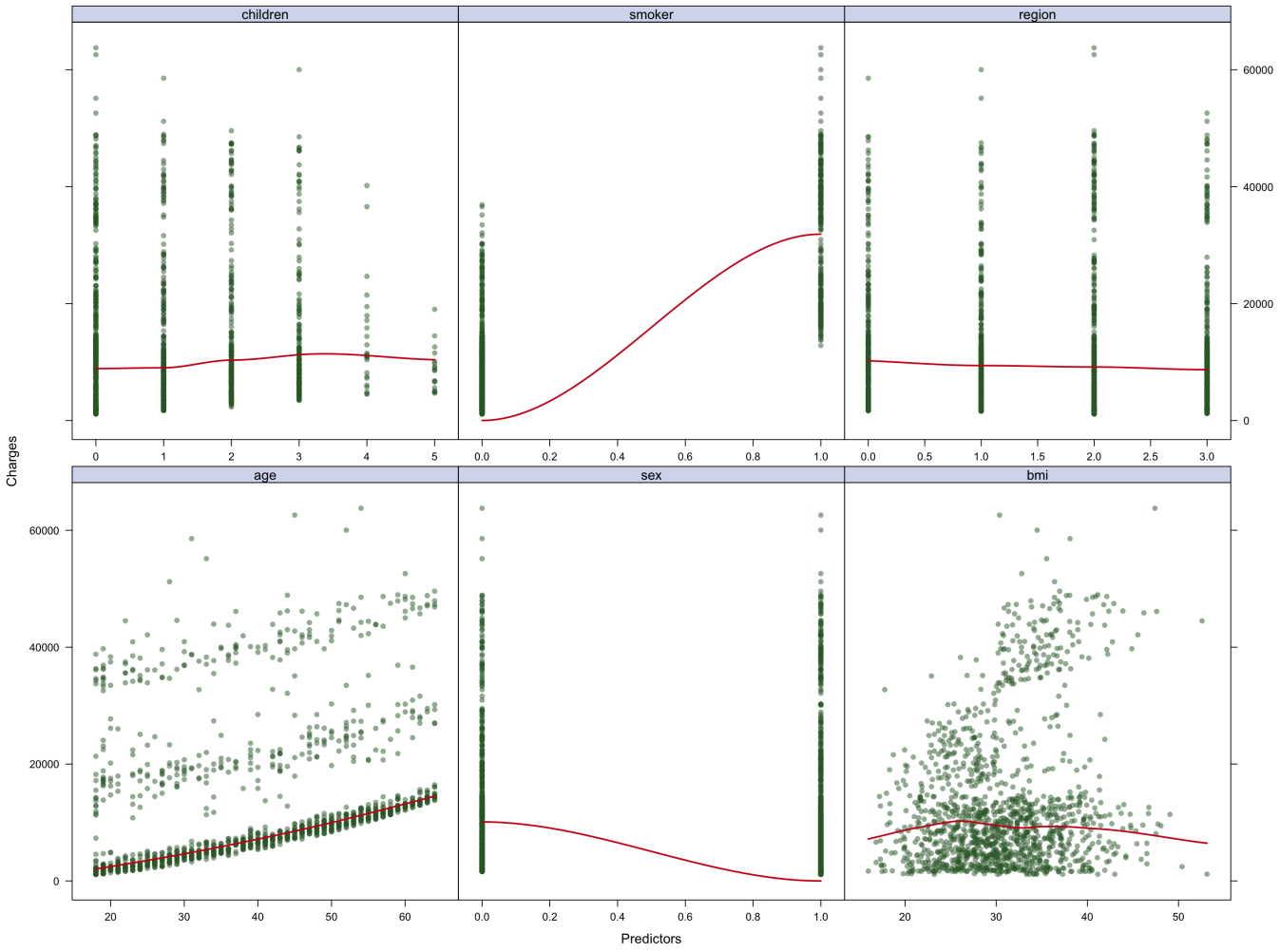
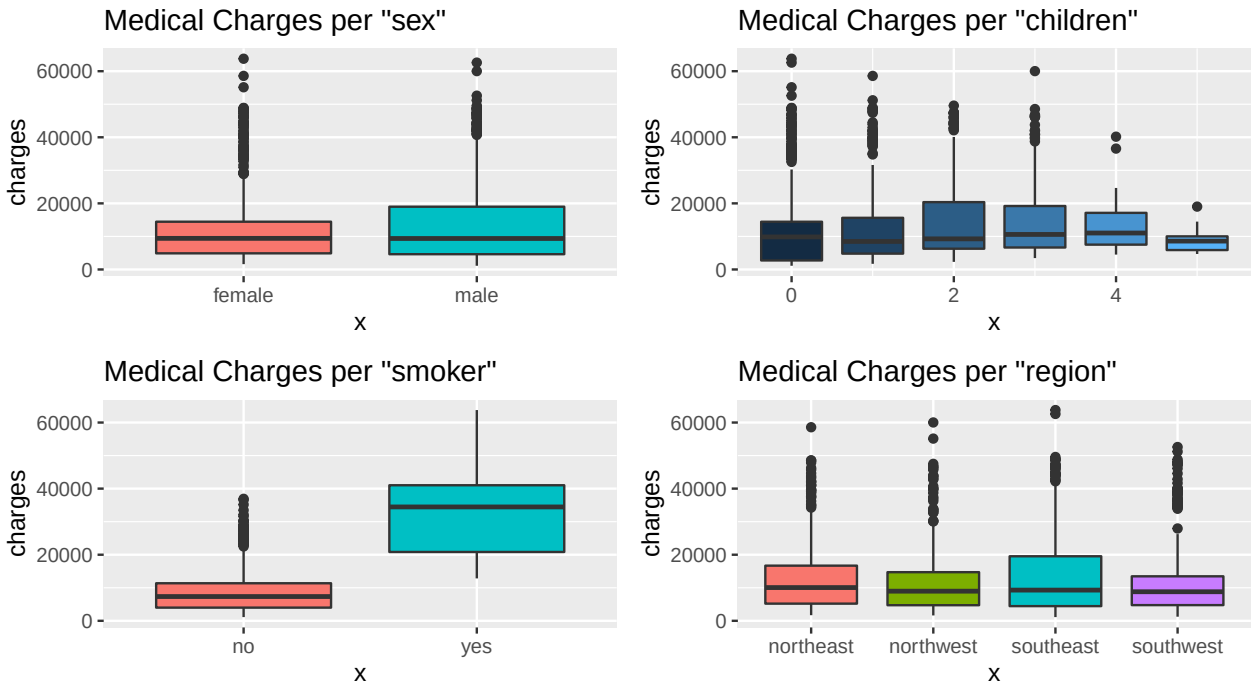
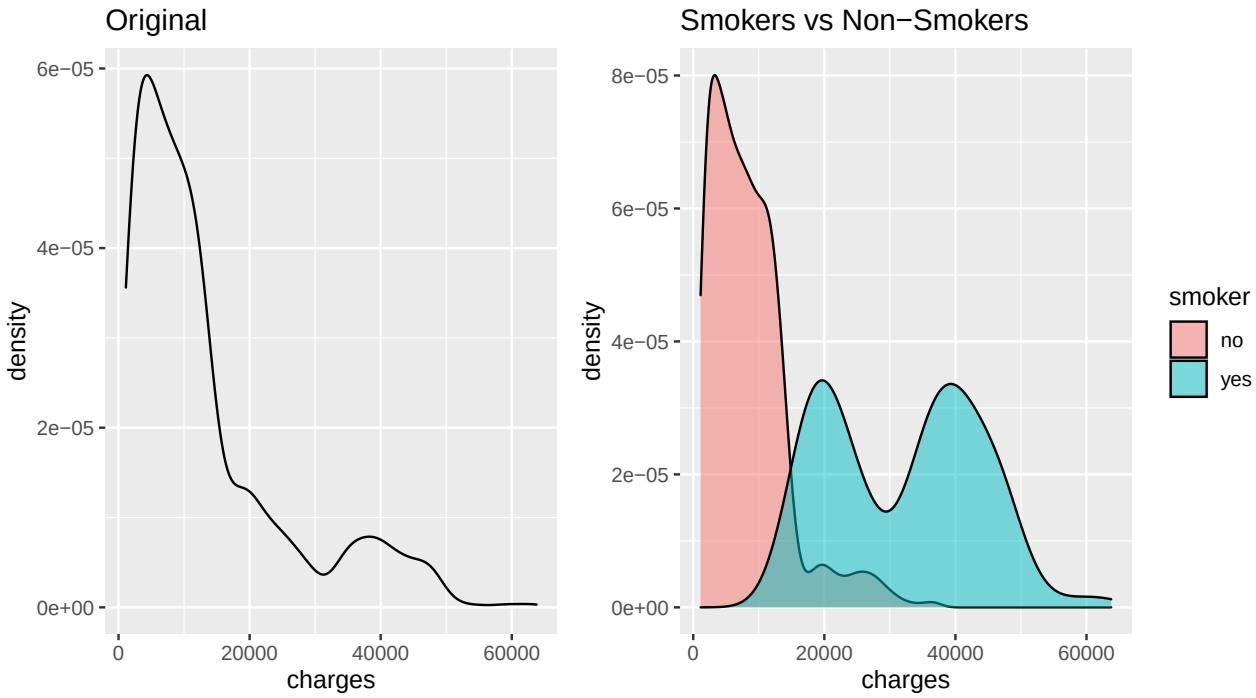


Figure 1: Scatter Plots of Medical Costs vs. Predictors

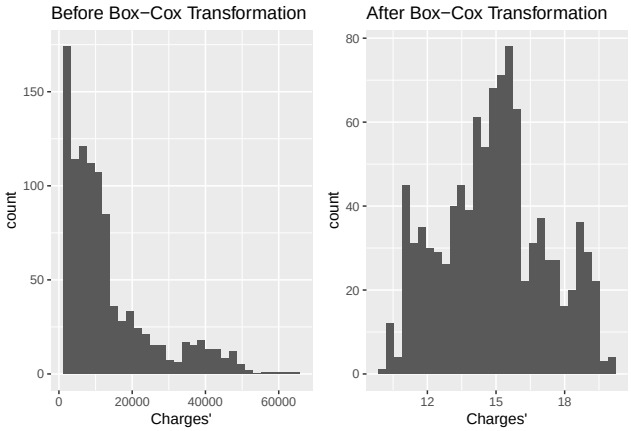


(a) Boxplot of Medical Costs vs. Predictors

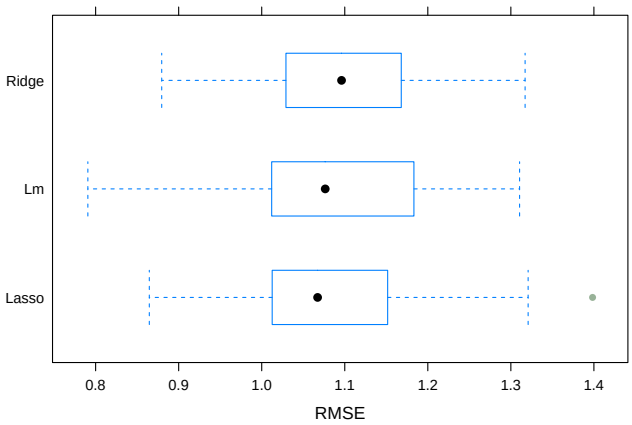


(b) Distribution of Medical Costs

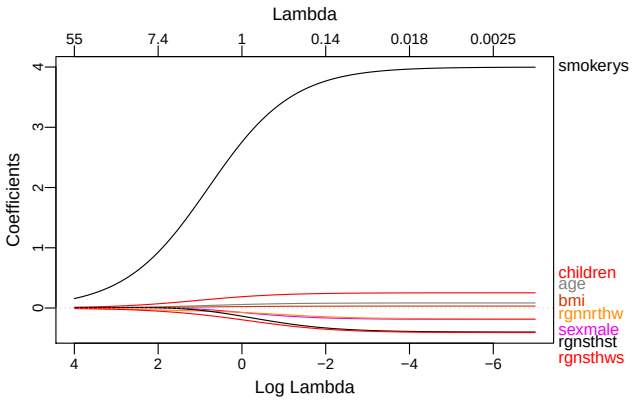
Figure 2: Explanatory Visualization of Medical Costs



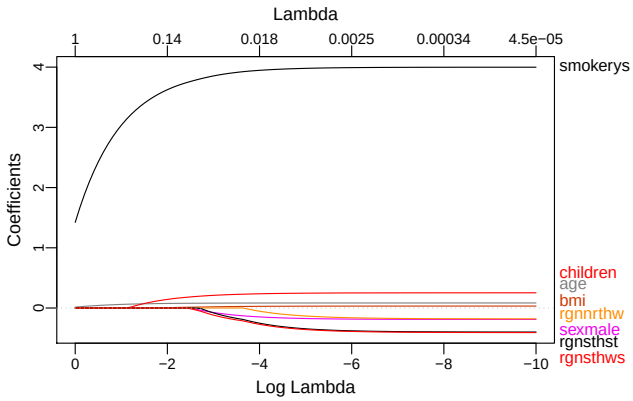
(a) Boxcox Transformation of the Response



(b) Ridge Regression

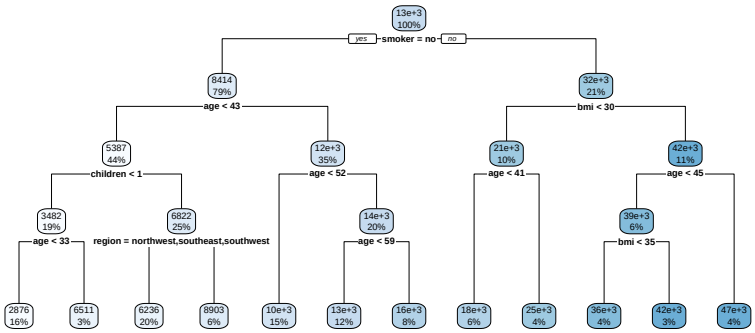


(c) Lasso Regression

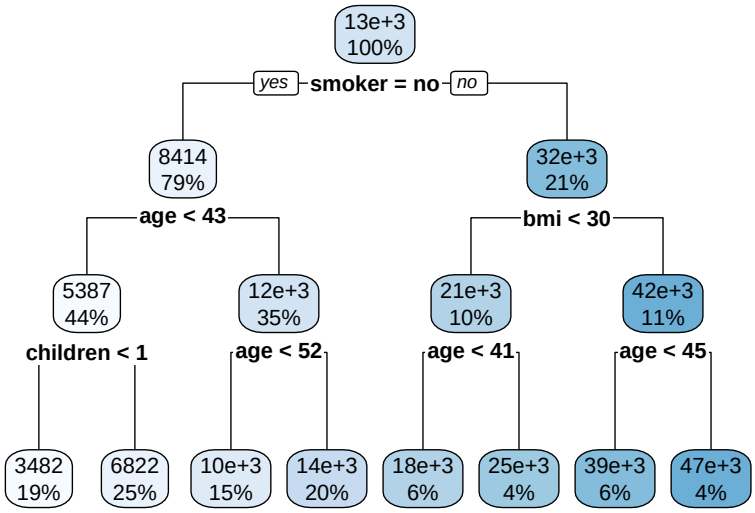


(d) Resampling of Three Linear Regression Models

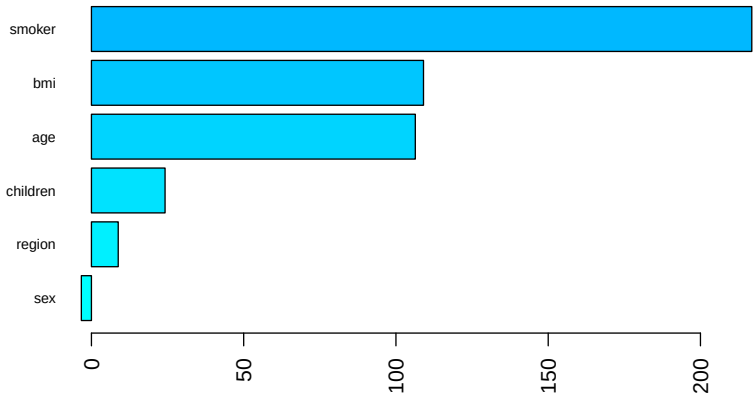
Figure 3: Cross-Validation of Ordinal LM, Ridge and Lasso Regression



(a) Regression Tree selected by Cross-Validation Error

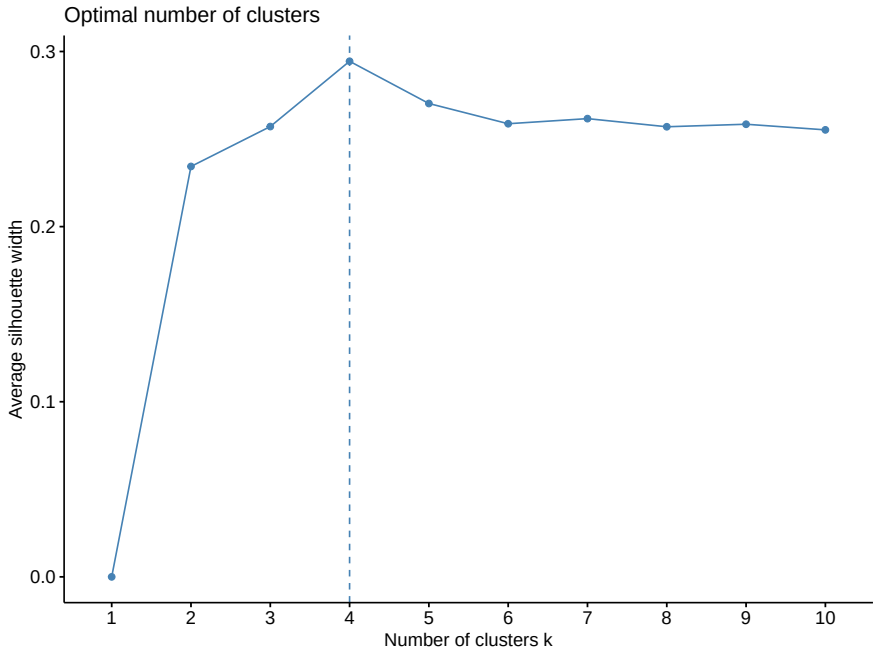


(b) Regression Tree selected by 1SE Rule

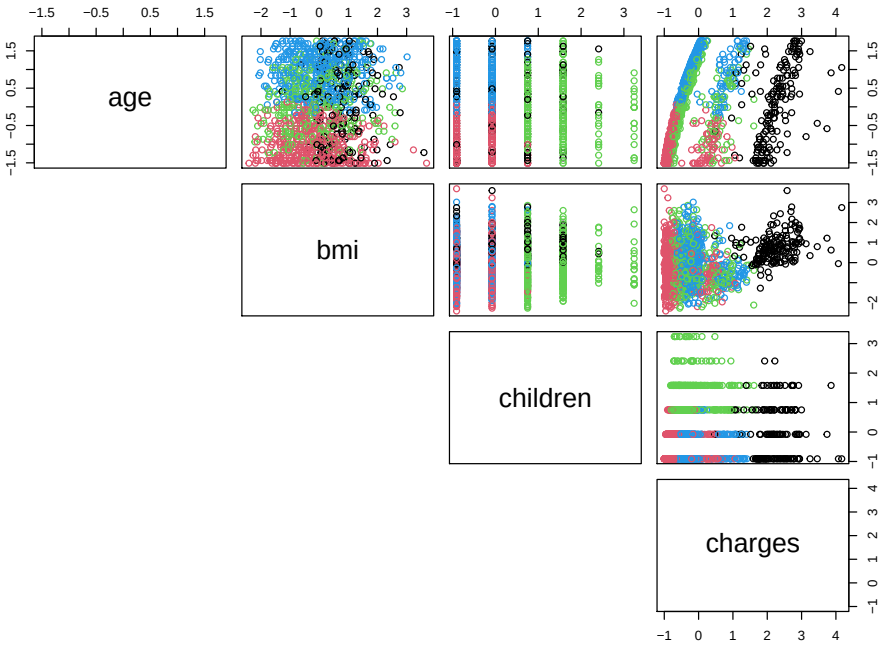


(c) Variable Importance by Random Forest

Figure 4: Tree Regression and Random Forest

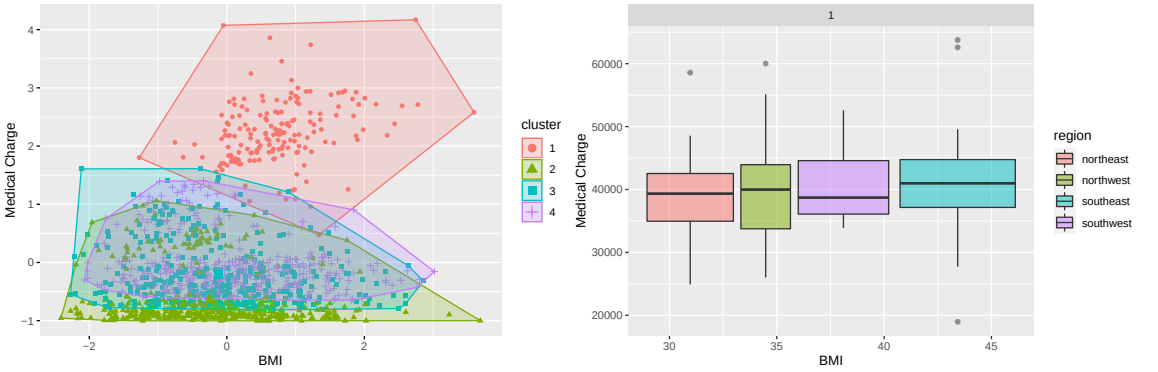


(a) Optimal Cluster Number



(b) Pairwise K-Means Clustering Results

Figure 5: K-Means Clustering



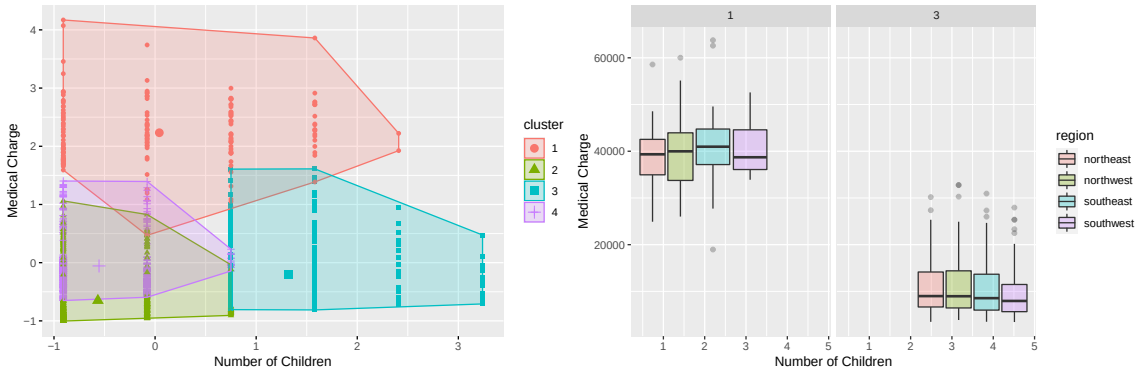
(a) K-Means Clustering

(b) Boxplot (Region)

(c) Boxplot (Sex)

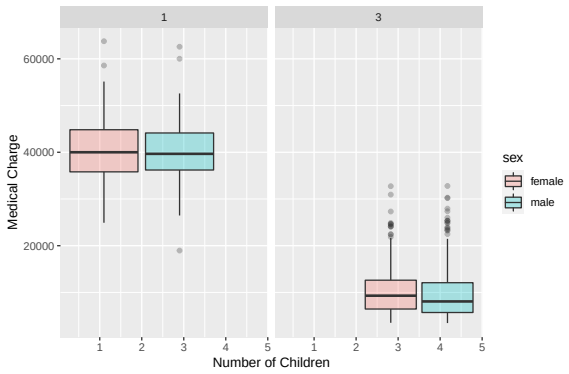
(d) Boxplot (Smoker)

Figure 6: BMI vs. Medical Charges for Subgroup 1

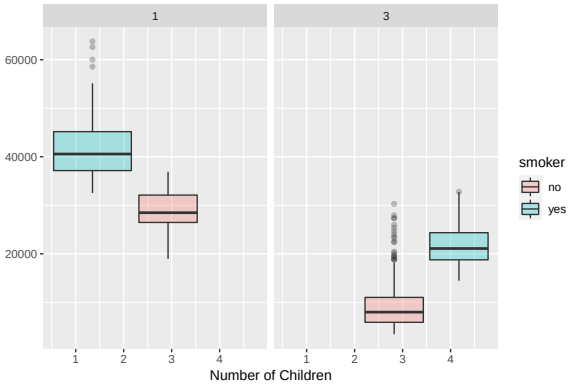


(a) K-Means Clustering

(b) Boxplot (Region)



(c) Boxplot (Sex)



(d) Boxplot (Smoker)

Figure 7: Number of Children vs. Medical Charges for Subgroup 1 and 3