Predicting Medical Charges with R Regression

Voy Adamek

2023-03-15

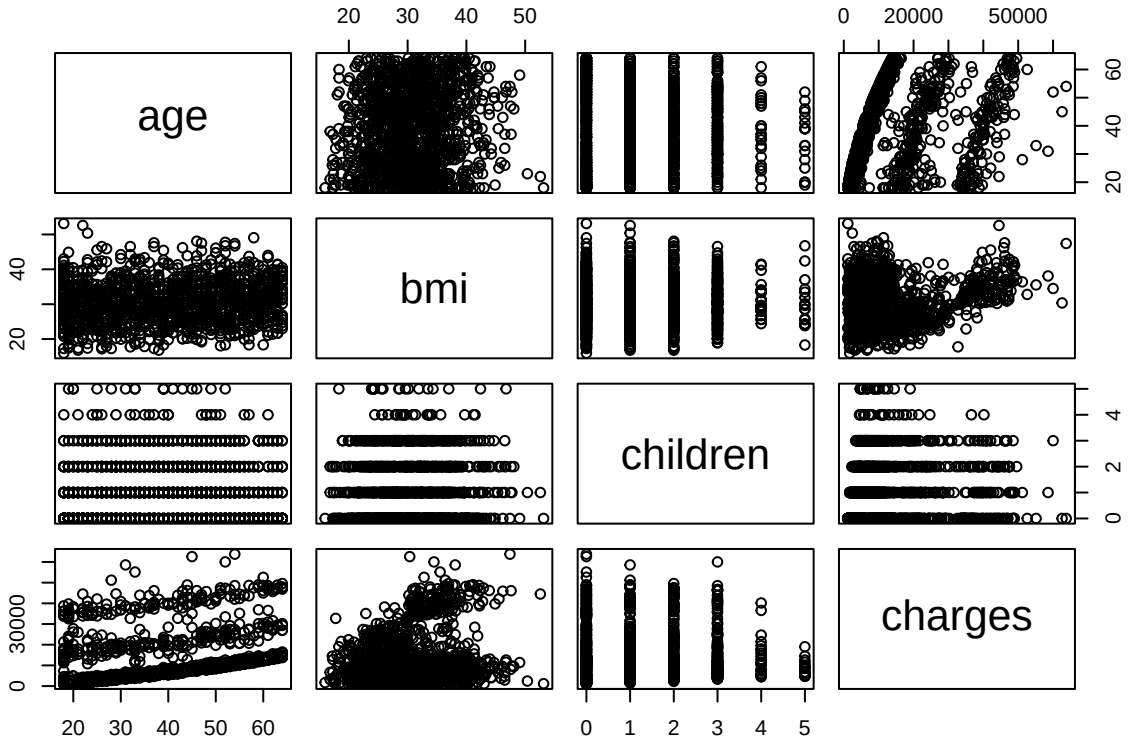
Introduction

In this project, we will explore a dataset of medical insurance charges and use linear regression to predict the charges based on various factors such as age, sex, BMI, smoker status, and region. We will also visualize the data to gain insights and better understand the relationships between the different variables.

Loading the data

First, we need to load the insurance dataset and check its contents. We use the `read.csv()` function to load the data into a dataframe and the `plot()` function to create a scatter plot of all the numeric columns in the dataframe.

```
df = read.csv('insurance.csv', header=TRUE)
num_cols <- unlist(lapply(df, is.numeric))
plot(df[,num_cols])
```



We can see from the plot that there is some correlation between the variables, but it's not immediately clear which variables are the most important. To get a better idea of the correlations, we can use the cor() function to create a correlation matrix.

```
round(cor(df[,num_cols]),2)
```

```
##             age  bmi children charges
## age        1.00 0.11     0.04    0.30
## bmi        0.11 1.00     0.01    0.20
## children   0.04 0.01     1.00    0.07
## charges    0.30 0.20     0.07    1.00
```

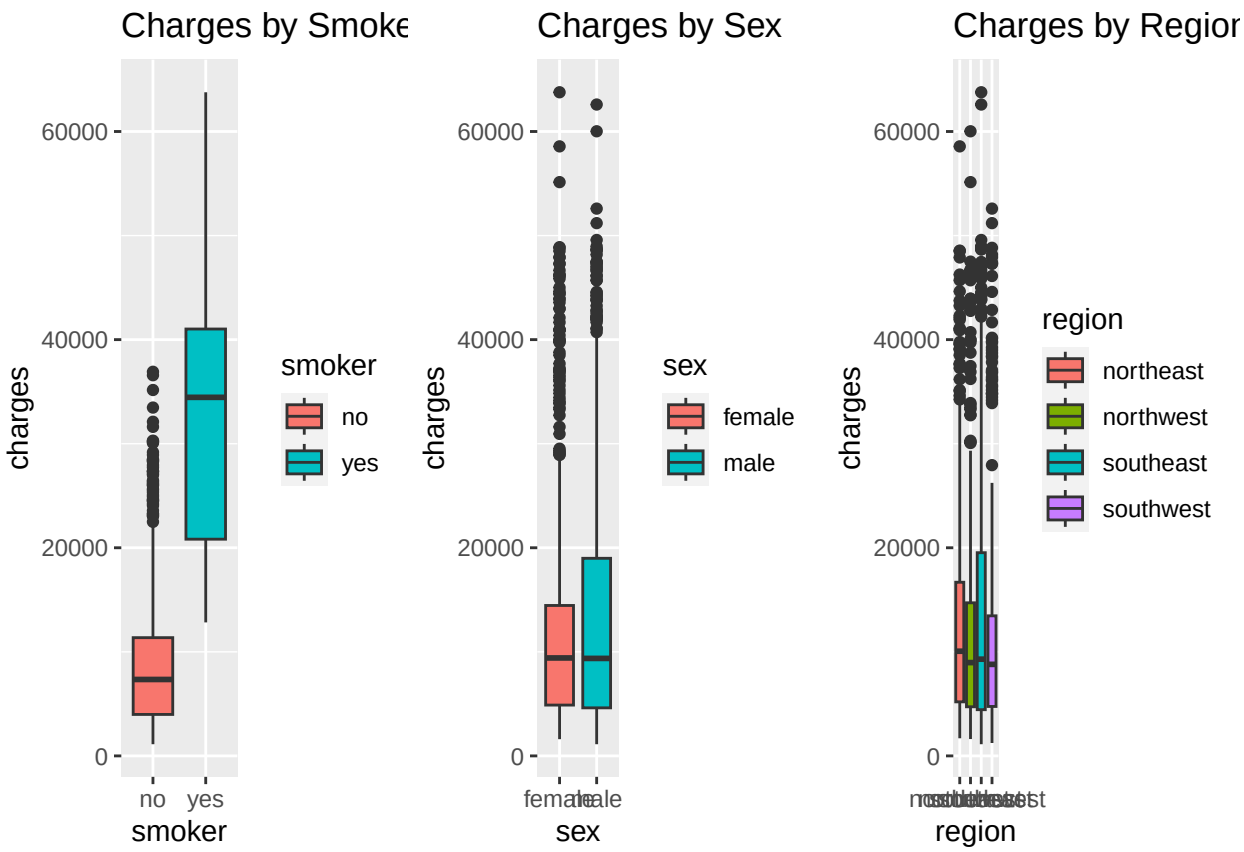
The cor() function calculates the correlation matrix between the numeric columns of the dataset. The resulting matrix shows that age and BMI have the strongest correlations with charges, while the correlation between children and charges is weaker. Specifically, age has a positive correlation of 0.3 with charges, and BMI has a stronger positive correlation of 0.2. The correlation between children and charges is weaker, with a value of 0.07.

Visualizing the data

Next, we can use boxplots to visualize the distribution of charges across different factors such as smoker status, sex, and region. This can help us see if there are any significant differences in charges based on these factors.

```
smoker = as.factor(df$smoker)
sex = as.factor(df$sex)
region = as.factor(df$region)

p1 <- ggplot(df, aes(x = smoker, y = charges, fill = smoker)) + geom_boxplot() + ggtitle("Charges by Sm
p2 <- ggplot(df, aes(x = sex, y = charges, fill = sex)) + geom_boxplot() + ggtitle("Charges by Sex")
p3 <- ggplot(df, aes(x = region, y = charges, fill = region)) + geom_boxplot() + ggtitle("Charges by Reg
gridExtra::grid.arrange(p1, p2, p3, ncol=3)
```



From the boxplots, we can see that smokers tend to have significantly higher charges than non-smokers, and there are also some differences in charges based on sex and region.

Linear Regression

To predict medical charges based on the available data, we built a linear regression model using the lm() function. In this model, charges were considered the dependent variable, while all other variables in the dataset were included as independent variables. The resulting model1 provides insights into the factors that significantly affect medical charges.

```
model1 = lm(charges ~. , data =df)
summary(model1)
```

```
##
## Call:
```

```
## lm(formula = charges ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

The summary() function provides a comprehensive overview of the model performance and highlights the coefficients for each independent variable. We observed that age, BMI, and smoker status were significant predictors of charges. Specifically, a one-year increase in age was associated with an increase of $257 in medical charges, while a one-unit increase in BMI corresponded to an increase of $332 in charges. Smokers had medical charges that were, on average, $23,857 higher than non-smokers.

In contrast, sex and region were not significant predictors of charges, as the p-values associated with their coefficients were higher than 0.05. This suggests that the effect of these variables on medical charges is not statistically significant, and other factors are likely to have a stronger influence. Overall, the linear regression model provides insights into the key factors driving medical charges and can inform healthcare policy decisions.

Conclusion

In this project, we have explored a dataset containing information about medical charges for different individuals. We began by loading the data and visualizing the relationships between the numerical variables. We then computed the correlation matrix to identify the variables that have the strongest associations with charges. Next, we created boxplots to explore the relationships between charges and the categorical variables of smoker, sex, and region. Finally, we built a linear regression model to predict medical charges based on the available data. The model summary indicates that age, BMI, and number of children are significant predictors of charges. Overall, this analysis provides insights into the factors that drive medical charges and can inform future healthcare policy decisions.