finalproject

2024-04-07

Section 1: Introduction Provide a brief introduction of the goal of this final project. What is it all about? Where did you get the data from? What is the data framework? What are the main questions you want to answer with this data analysis?

The data comes from the Kaggle data sets. The link can be found here: https://www.kaggle.com/datasets/mirichoi0218/insurance/data

Section 2: Exploratory Data Analysis Include some graphical displays and numerical summaries of the data. Also comment on any patterns/characteristics of the data which you find interesting or anything relevant to your later analysis.

Provide a brief explanation/summary of variables you plan to include in your analysis. Here are some question you might ask:

Which variables are categorical (when applicable) and which are numerical? Should we remove any unusual observations? Should we add or remove some variables in our analysis? For categorical variables (when applicable), should we include any interactions? For numerical variables, any evidence supporting nonlinear trends?

Section 2: Exploratory Data Analysis

Here is a brief overview of all of the variables of this dataset:

Numerical Variables:

- age: age of primary beneficiary
- bmi: body mass index
- children: number of children/dependents covered by health insurance
- charges (response variable): medical costs billed by health insurance

Categorical Variables:

- sex: male or female
- smoker: yes or no
- region: residential area in the US, can be northeast, southeast, southwest, or northwest

The first step is to check for unusual observations by performing an outlier test. To do so, we obtained the studentized residuals for our data, and compared the largest values with the Bonferroni critical value.

This is the Bonferroni critical value we calculated:

```
## [1] -4.137174
```

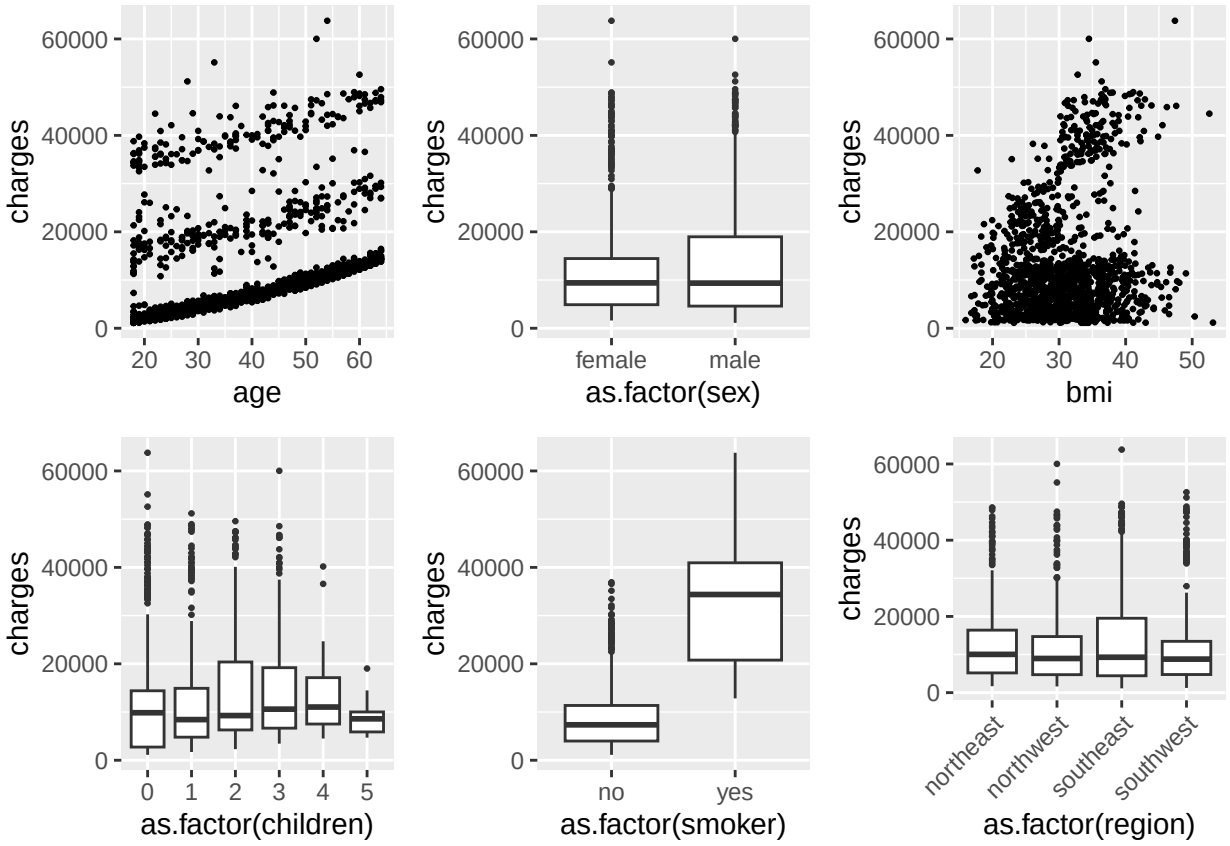
Next, here are the five largest studentized residuals, which we will be comparing to the Bonferroni critical value:

```
##     1301      578      243      220      517
## 5.009599 4.219800 4.053228 3.998326 3.863878
```

Since the absolute value of their respective studentized residual is greater than the Bonferroni critical value, we can conclude that Observation #1301 and Observation #578 are outliers.

We also checked Cook's distance, however we found that there is no point with Cook's distance greater than 1. We will simply remove the two outliers we found earlier.

After removing these two points, we will be analyzing the predictors to see if there are any variables we should add or remove.



From the plot of charges vs age, we notice a trend that the charges increase as age increases. We will not be removing 'age' as a predictor since it is clear that it has a significant impact on the response variable 'charges'. This makes sense because as people grow older, their health declines, and the insurance costs increase.

The charges vs sex boxplot tells us that males tend to have higher charges than females. We will conduct a t-test to test this difference.

We obtained this p-value from the t-test:

```
## $p.value
## [1] 0.03460436
```

Given that the p-value is less than 0.05, we conclude that 'sex' is a significant predictor, with males incurring higher insurance costs than females. A possible explanation for this is that males are usually more at risk of health conditions that result in increased charges compared to females.

From the plot of charges vs bmi, we can see a general trend that as bmi increases, the charges also increase. As a result, we will not be removing 'bmi' as a predictor since it appears to have an impact on the response variable. The graph also makes sense logically since people with a higher bmi tend to be overweight and subsequently, have worse health than people with a lower bmi, causing their insurance charges to be greater as well.

The boxplot of charges vs children conveys to us that charges tend to increase as the number of children grows from 0 or 1 to 2 or 3. While there are some data points for people with 4 or 5 children, it appears that

these groups have a very low amount of people, and it may be challenging to draw meaningful conclusions from it. We perform an ANOVA test to test the significance of 'children' as a predictor.

This is the p-value we obtain:

```
## [1] 0.008620164
```

This p-value is less than 0.05, thus signifying that the predictor 'children' is significant. We will be keeping as a predictor in our model. A larger number of children/dependents covered leads to higher charges, as there are more individuals who need to be covered.

From the boxplot of charges vs smoker, it is evident that smokers generally face significantly higher charges than non-smokers. Consequently, we will not be removing 'smoker' as a predictor. People who smoke typically incur higher insurance costs due to the health risks posed by smoking.

The boxplot of charges vs region shows us that the southeast region seemingly has higher charges than the other three regions. We will conduct an ANOVA test to make sure that the predictor 'region' is significant.

This is the p-value we obtained from the ANOVA test:

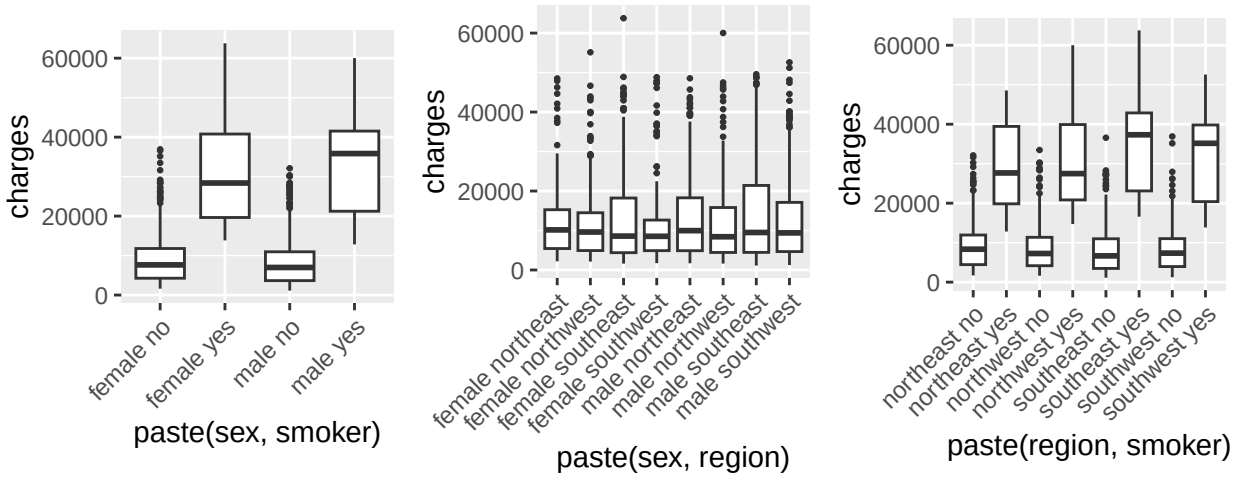
```
## [1] 0.04525914
```

The p-value is less than 0.05, indicating that the predictor 'region' is significant, and thus we will retain it as a predictor. This confirms that there are regional differences in the insurance costs.

Now, we will explore whether or not we should include any interactions between categorical variables. We will be analyzing all possible first-order and second-order interactions between our categorical predictors, including:

- 'sex' and 'smoker'
- 'sex' and 'region'
- 'region' and 'smoker'
- 'sex', 'smoker', and 'region'

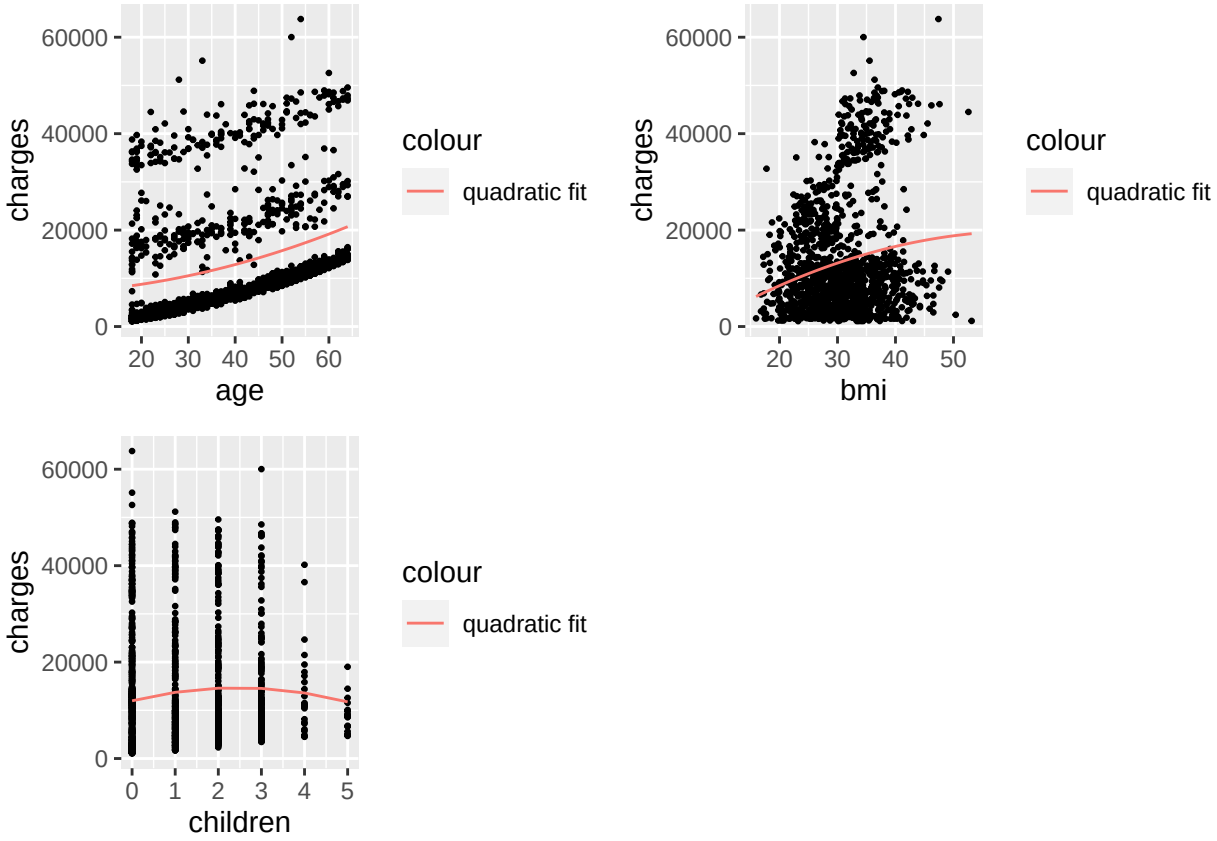
Here are some plots that show the interactions between these predictors.



From the plots, it appears as though smoking increases the charges more for males than females. Additionally, it seems like smoking has a greater impact on charges in the southeast and southwest, when compared to northeast and northwest. However, it does not appear that sex and region is a significant interaction.

We applied sequential F-tests with the anova function to assess the significance of the interaction terms in our model. Initially, we included all potential interactions among the three variables 'sex', 'smoker', and 'region'. We first found that the second-order interaction sex:smoker:region was not significant, and removed it from the model. Subsequently, we tested the first-order interactions and found that the interaction sex:region was not significant. Finally, after removing sex:region as a predictor, we concluded that only the interactions sex:smoker and smoker:region are statistically significant, and we will retain them in the final model.

Now, we will examine the numerical variables and see if there is evidence of any nonlinear trends.



Based on the plot of the quadratic fit for charges vs age, we see there is evidence of non-linearity for the predictor 'age' since the quadratic fit has some upward curvature.

Similarly, based on the plot of the quadratic fit for charges vs bmi, we see there is some evidence of non-linearity for the predictor 'bmi', although maybe not as clear as for 'age'. This time, the quadratic fit curves downwards.

Finally, the plot of charges vs children also suggests some non-linearity. However, the non-linearity is not as evident as for 'age and 'bmi' since although the quadratic fit does appear to curve downwards as the number of children increases, there is not many data points for 4 and 5 children, so we may not be able to draw a strong conclusion.

Section 3: Methodology

You are required to build at least two prediction models using the methods covered in this class. For each model or method you are using, include a brief description of the methodology and a description of the R implementation (R coding steps).

Your should consider the following sub-sections: Section 3.1: Start with a simple model, a model that doesn't require much training, for example, a linear regression model built after appropriate variable selection and model diagnostics. Section 3.2: Use the Linear Regression model built in 1 to make predictions on a testing set. Section 3.3: Fit a different kind of model like a non-parametric regression, regularized regression models or others, and make a prediction on the same testing set. Section 3.4 (Optional): You can also try with other methods learnt in other classes if applicable (like for example Random Forests).

Section 4: Discussion and conclusions In this part you should make a summary of your results by comparing the different modeling approaches and discussing the impact of your analysis. You should also write the main conclusions of your analysis in three to four bullet points.