Multiple Linea Regression Project

Yifan Jiang

Introduction

I will be using a dataset that is specifically dedicated to the expenses incurred for treating different patients. The cost of treatment depends on a number of factors, including the patient's age, sex, number of children, type of clinic, and more. While we lack data on the patient's diagnosis, we possess other relevant information that can help us draw conclusions about their overall health and conduct regression analysis.

The detail of the variables in this data set:

- 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
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

I will be using age, sex, bmi, children, and smoker variables as explanatory variables and charges as response variable. I will look into the linear relationship between these explanatory variables and the response variable.

Data Description

This data set contains 1338 observations and total of 7 variables. I will be only using 6 of them which are age, sex, bmi, children, smoker, and charges. I first change the two categorical variable into numerical variables with value of 0 and 1. For variable “sex”, I assigned value 1 to male and 0 to female. For variable “smoker”, I assigned value 1 to yes and 0 to no.

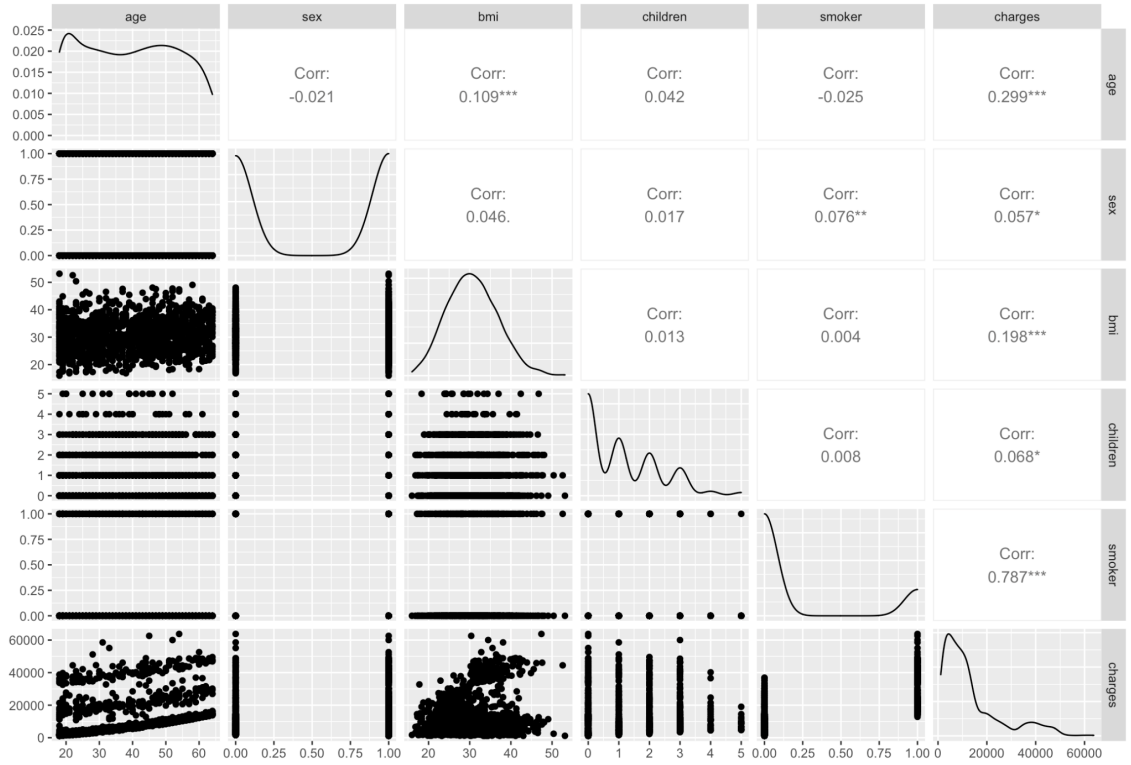
- Summary statistic

```
> summary(dataSet)
      age             sex              bmi            children        smoker          charges     
 Min.   :18.00   Min.   :0.0000   Min.   :15.96   Min.   :0.000   Min.   :0.0000   Min.   : 1122  
 1st Qu.:27.00   1st Qu.:0.0000   1st Qu.:26.30   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.: 4740  
 Median :39.00   Median :1.0000   Median :30.40   Median :1.000   Median :0.0000   Median : 9382  
 Mean   :39.21   Mean   :0.5052   Mean   :30.66   Mean   :1.095   Mean   :0.2048   Mean   :13270  
 3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:34.69   3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:16640  
 Max.   :64.00   Max.   :1.0000   Max.   :53.13   Max.   :5.000   Max.   :1.0000   Max.   :63770  
```

- Correlations

```
> correlation_matrix
                age         sex         bmi    children       smoker     charges
age       1.00000000 -0.02085587 0.109271882 0.04246900 -0.025018752 0.29900819
sex      -0.02085587  1.00000000 0.046371151 0.01716298  0.076184817 0.05729206
bmi       0.10927188  0.04637115 1.000000000 0.01275890  0.003750426 0.19834097
children  0.04246900  0.01716298 0.012758901 1.00000000  0.007673120 0.06799823
smoker   -0.02501875  0.07618482 0.003750426 0.00767312  1.000000000 0.78725143
charges   0.29900819  0.05729206 0.198340969 0.06799823  0.787251430 1.00000000
```

- The distribution of each variable and relationships among the variables



Results and interpretation

Created the linear model with the 6 variables.

```
Call:
lm(formula = charges ~ ., data = dataSet)

Residuals:
     Min       1Q   Median       3Q      Max
-11837.2  -2916.7   -994.2   1375.3  29565.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12052.46     951.26 -12.670  < 2e-16 ***
age            257.73      11.90  21.651  < 2e-16 ***
sex           -128.64     333.36  -0.386 0.699641
bmi            322.36      27.42  11.757  < 2e-16 ***
children       474.41     137.86   3.441 0.000597 ***
smoker       23823.39     412.52  57.750  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6070 on 1332 degrees of freedom
Multiple R-squared:  0.7497,    Adjusted R-squared:  0.7488
F-statistic:   798 on 5 and 1332 DF,  p-value: < 2.2e-16
```

Based on the summary table of the linear model, we see that the Multiple R-squared value is 0.7497 which tells us that around 75% of the variation in the charges can be explained by the explanatory variables. The F-statistic and p-value suggests that the model is a pretty good fit of the data. However, we can see that the variable "sex" is not significant. Thus, I created a reduced model without the variable "sex" to see which model is better.

```
Call:
lm(formula = charges ~ age + bmi + children + smoker, data = dataSet)

Residuals:
     Min       1Q   Median       3Q      Max
-11897.9  -2920.8   -986.6   1392.2  29509.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12102.77     941.98 -12.848  < 2e-16 ***
age            257.85      11.90  21.675  < 2e-16 ***
bmi            321.85      27.38  11.756  < 2e-16 ***
children       473.50     137.79   3.436 0.000608 ***
smoker       23811.40     411.22  57.904  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6068 on 1333 degrees of freedom
Multiple R-squared:  0.7497,    Adjusted R-squared:  0.7489
F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

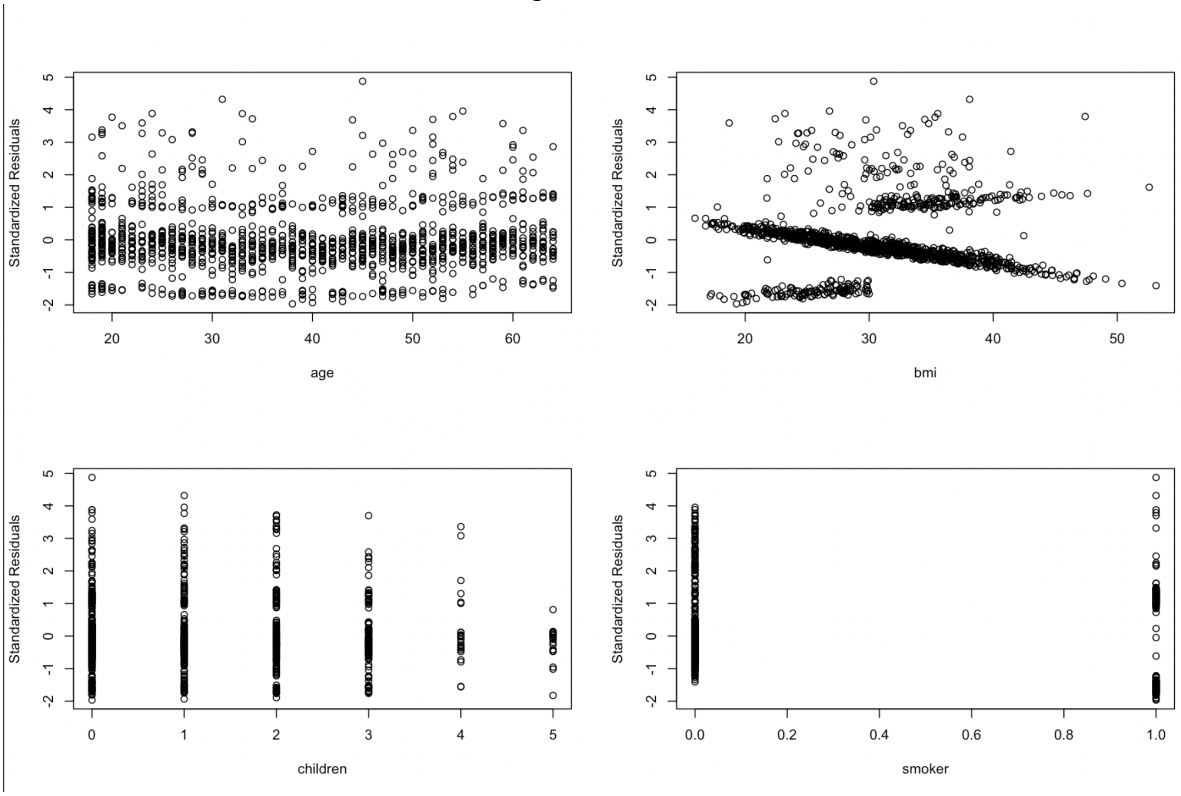
```
> anova(reducedModel, linearModel)
Analysis of Variance Table

Model 1: charges ~ age + bmi + children + smoker
Model 2: charges ~ age + sex + bmi + children + smoker
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1   1333 4.9078e+10
2   1332 4.9073e+10  1   5486063 0.1489 0.6996
```

The 0.6996 p-value in the Anova table tells me that we fail to reject the null hypothesis(the reduced model) due to the p-value, hence the reduced model is the better fit. Therefore the prediction function is:

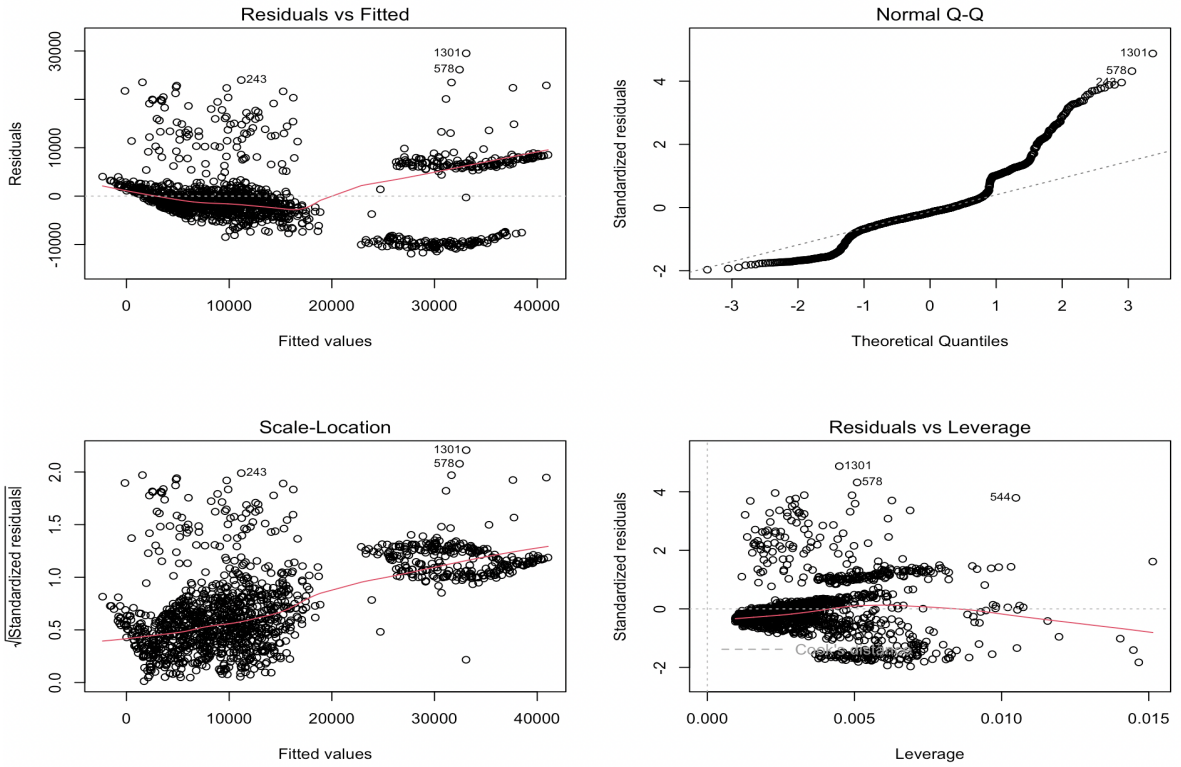
*Predicted charges = 257.85 * age + 321.85 * bmi + 473.5 * children + 23811.4 * smoker*

Then I looked at the Standarized residual plots of the data set.



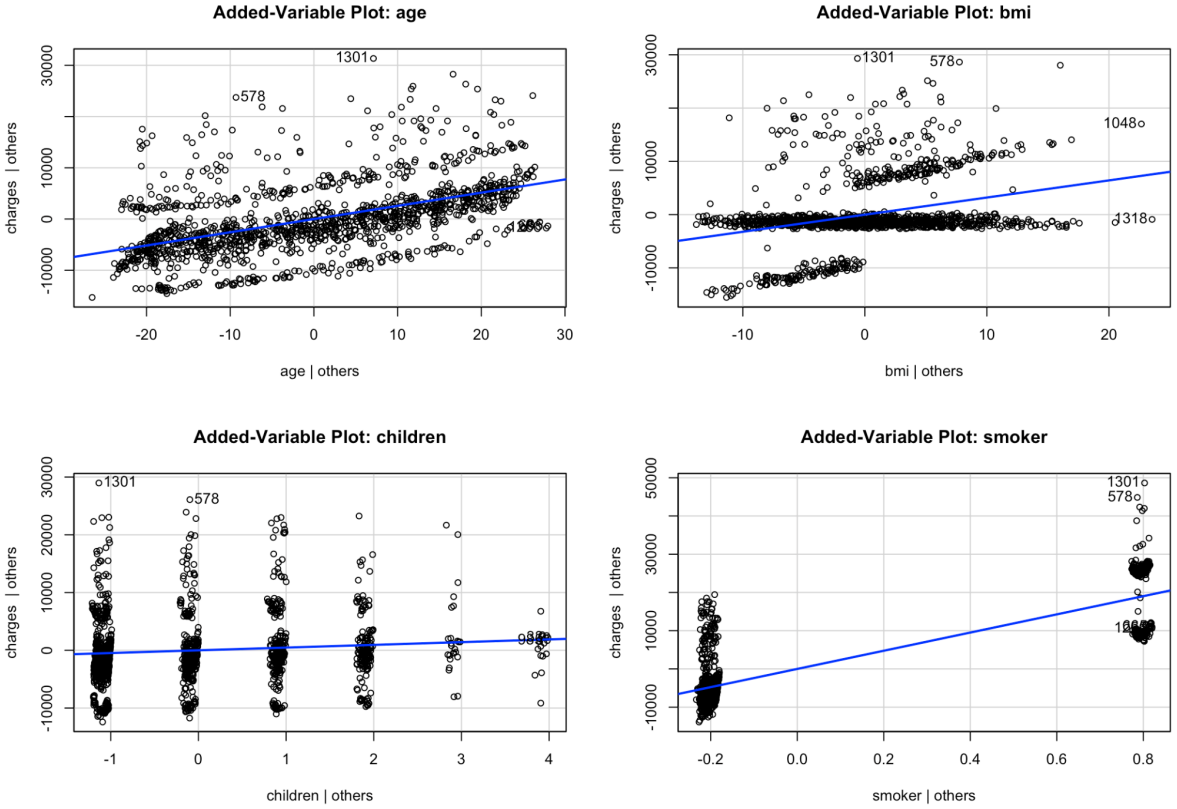
The fairly random nature of these plots is indicative that the model is a valid model for the data.

I also assessed the model using some diagnostic tools.



From the Residual vs fitted graph, we can see that the relationship is not perfectly linear.From the Normal Q-Q plot, we can see that the graph is heavily tilted on the right top corner. From the Scale-Location graph, it's showing the variance of the error is not perfectly constant. For the Residual vs Leverage plot, I first calculate Leverage points = 2*(4+1)/1338 = 0.007473842, we can see a lot of points outside of [-2,2] and > 0.0075 range. From the result of diagnostic tools, we can tell that the data set contains many outliers and leverage points which could effect the fitness of the linear model and the accuracy of the predictions.

I also looked at the Added-Variable Plots and found that the results(slopes) are consistent with our prediction function.



I then checked for possibility of multicollinearity using VIF.

```
> vif(reducedModel)
     age      bmi children   smoker
1.014498 1.012194 1.001950 1.000745
```

The values of the variables are not greater than 5, which imply that the slopes for the predictors are estimated pretty accuratly, thus there are no significant sign of multicollinearity.

Transformation

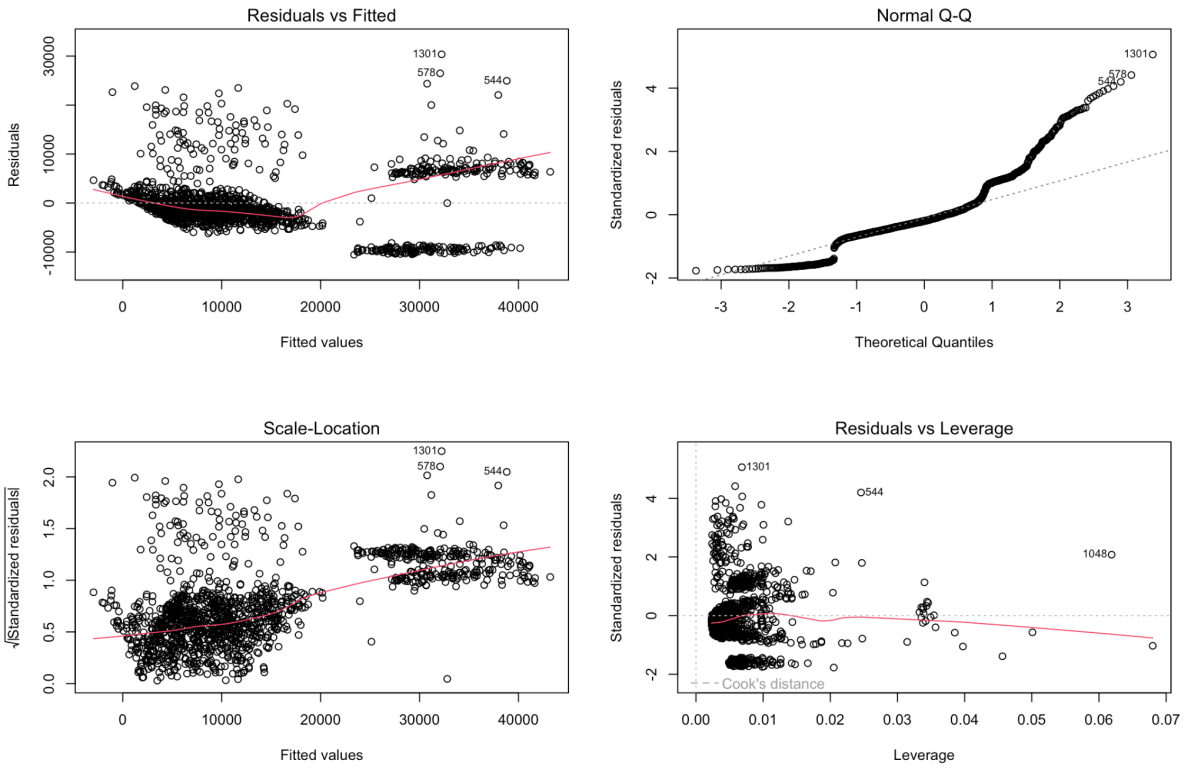
I decide to try a polinomial(quadratic) transformation to see if there are any improvement because from the shape of the Residual vs fitted graph I see a little curve.

```
Call:
lm(formula = charges ~ age + I(age^2) + bmi + I(bmi^2) + children +
    I(children^2) + smoker + I(smoker^2), data = dataSet)

Residuals:
   Min     1Q Median     3Q    Max
-10551  -3114  -1196   1702  30358

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -13518.329   3498.607  -3.864 0.000117 ***
age              -87.357     82.479  -1.059 0.289726
I(age^2)           4.322      1.028   4.204  2.8e-05 ***
bmi              792.804    206.940   3.831 0.000134 ***
I(bmi^2)          -7.542      3.251  -2.320 0.020496 *
children        1272.677    371.985   3.421 0.000642 ***
I(children^2)   -185.366    100.799  -1.839 0.066142 .
smoker         23813.533    408.529  58.291  < 2e-16 ***
I(smoker^2)           NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6021 on 1330 degrees of freedom
Multiple R-squared:  0.7541,    Adjusted R-squared:  0.7528
F-statistic: 582.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```



However, it did not really help.

Conclusion

In this project, I looked into the relationship between some different factors of people and their corresponding medical charges. I first built a linear model to see if the relationships are linear. Then I checked the statistic significant of all the variables to see if all of them are significant to the model. After that, I build the reduced model with one less variable. Then I vaildated the linear model with plots, VIF and diagnostic tools. I then see there is a small curve in the Residual vs Fitted graph, thus I tried a quatratic transformation to see if there will be an improvement. And the result I got did not improve much. Overall, the reduced linear model is a pretty vaild model for this data set. From the prediction function, we can see that whether a person is a smoker has the most influence on the cost of medical bill. I found this article from National Library of Medicine https://pubmed.ncbi.nlm.nih.gov/9321534/ that talks about how a person's smoking status will affect one's medical costs. I brings up a very important idea that I did not think of. People who do no smoker usually live longer, therefore have overall larger amount of medical costs compared to smokers. And this showed me the limitations of my research project that the data we have limited, to fully research into this topic, we need more precise data and include larger ranges for each variable.

Citation Page

Barendregt, J. *"The health care costs of smoking"*. National Library of Medicine.