Estimation of Medical Cost using Multiple Linear Regression

Group Number: 6

Arib Alam (17MA20054), Ayush Suhane (17MA20005), Biswajit Ghosh (17MA20011)
Kunal Borse (17MA20053), Shreyas Kowshik (17MA20039) and Yash Sharma (17MA20050)

Problem Overview

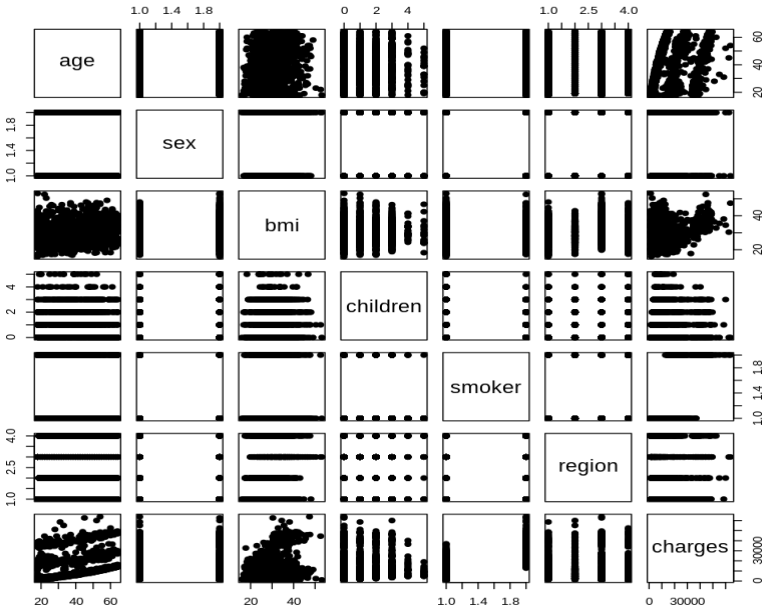
We have explored a dataset consisting of the cost of medical treatments and its relation with various factors (mostly non-medical). Insurance companies create models that accurately predict healthcare costs and then decide the insurance cover charges for a person. We attempt to make a conclusion about the health of patients and predict their medical costs. The cost of treatment depends on a multitude of factors such as diagnosis, city of residence, financial status etc. We use general information (age, sex), medical well-being (bmi), lifestyle (children, smoker) and region of residence to estimate the cost of medical treatment

Variables

Our dataset consists of 1,338 observations and 7 columns.
The columns are

- age: Age of the person
- sex: Gender of the person
- bmi (Body Mass Index): bmi of the person
- children: Number of dependents covered under the health insurance cover
- smoker: Whether the person smokes
- region: Persons' residential region (in the US)
- charges: Medical costs billed by health insurance

Pair Plots and Inferences



We can see that the variable **charges** show distinctive patterns with the variables **age**, **bmi**, **smoker** and **children**. For categorical variable **smoker**, we can observe some difference in the distribution of **charges** for each of the two categories, **smoker** and **non-smoker**.
No predictors appear to be highly correlated among each other, which can be further verified by calculating the Variance Inflation Factors. VIF of 1 indicates no presence of multicollinearity.

	GVIF	Df	GVIF^(1/(2*Df))
Smoker	20.202230	1	4.494689
age	1.003372	1	1.001684
children	1.003554	1	1.001775
region	1.028150	3	1.004638
smoke_bmi	20.277363	1	4.503039

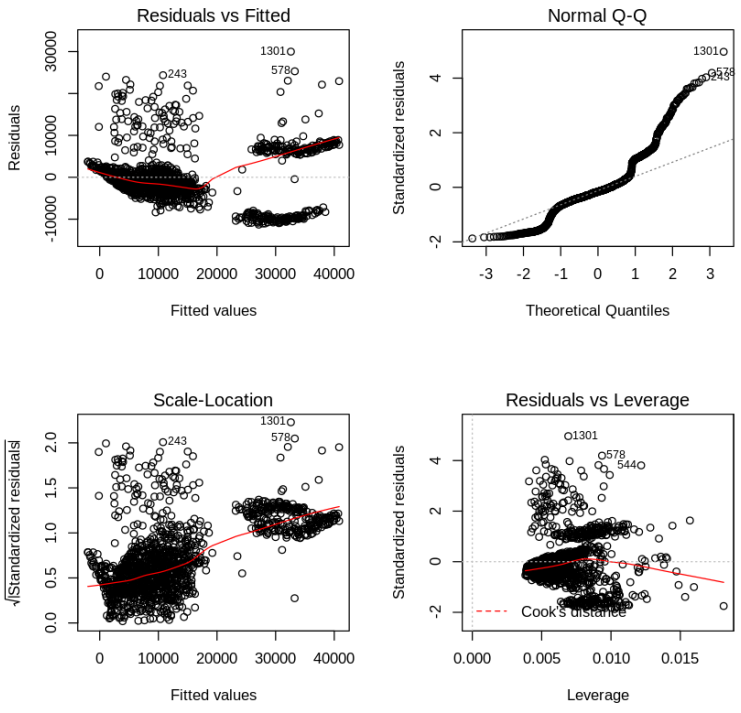
Building Preliminary Model with all variables

```
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
```

We first try building a Multiple Linear Regression model without any modifications, and using all the variables. Then we analyse the results and try to improve our model through various techniques.

As seen from the table, age, bmi, children and smoker seem to be the primary significant features in our dataset, as they have small p-values. We also observe that our value of Adjusted R- squared is 0.7494, which is a good value, though we will see that it can further be improved upon.

Analysis of Preliminary Model



As seen from the plot, the Residuals vs Fitted values plot shows a V- shaped pattern. This could suggest a non- linear relationship between the target and predictor variables, which will be explored in detail. Perhaps, the model we are building is not sufficient enough to capture variation in the data. We would try to modify some of the predictor variables and improve the model.

We also see that the Q-Q plot completely deviates from a straight line indicating non-normal distribution of the residuals.

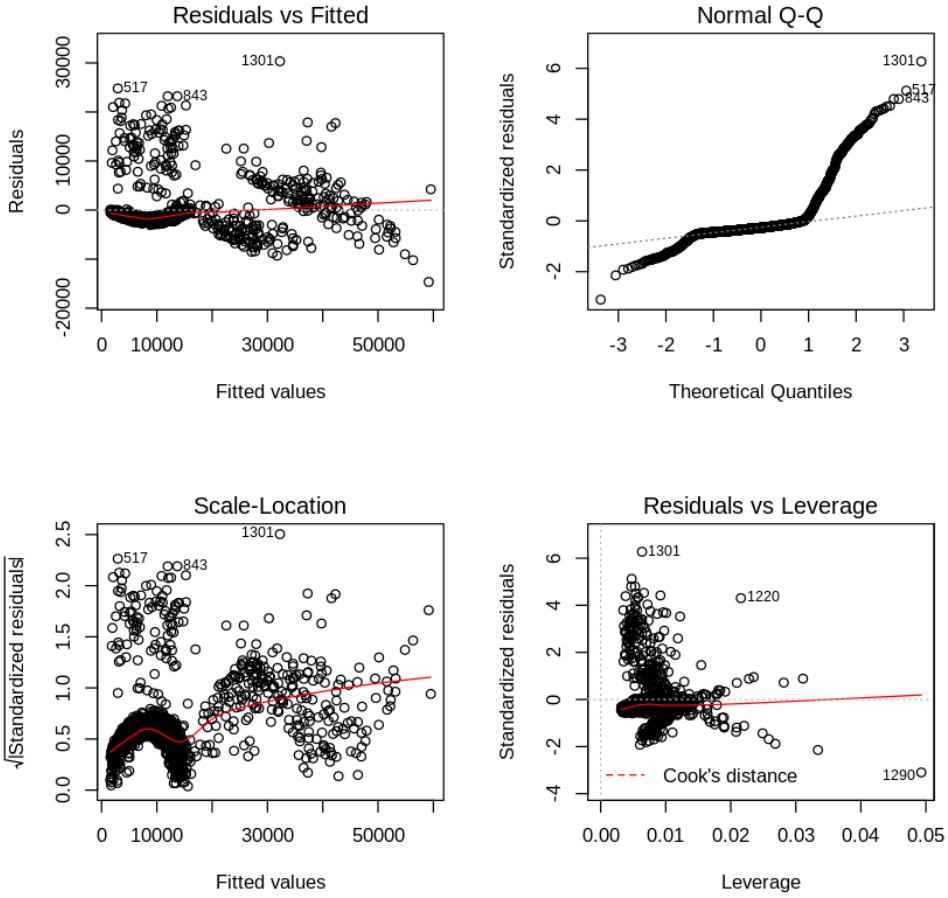
We see from the scale-location plot that the residuals are not appearing to be constant, this may point toward heteroskedasticity. Lastly, from the last graph, we conclude that there are no cases of influential data points, as the Cook's distance lines are barely even visible in the plot so there are no data points beyond the Cook's lines.

Lastly, we test our data for heteroskedasticity using the Breusch-Pagan Test.

BP = 121.74, df = 8, p-value < 2.2e-16

The null hypothesis for the above test is that there is constant variance. Since the p-value is quite low and is clearly less that alpha=0.05, we reject the null hypothesis and thus conclude that there is heteroscedasticity in our data, as confirmed from our plots above.

Combining ‘BMI’ and ‘Smoker’ variables



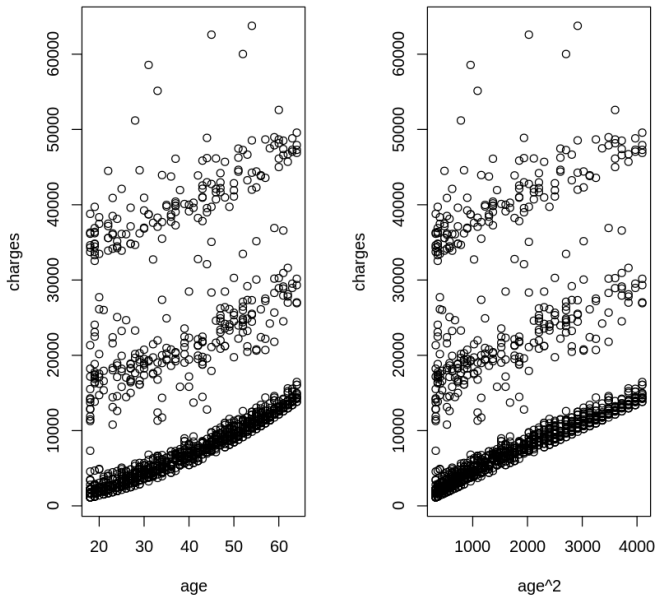
Since there is some non-linearity in the Residuals vs Fitted plot above, it makes sense that there are some other factors of variation, perhaps nonlinear that are not being captured.
For this purpose, a new variable is introduced : BMI * Smoker
Basically this is the value of BMI of only smokers while the other values are zero. The intuition behind this is that a person who has a high BMI and smokes is likely to spend more on his charges

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2453.564    857.695  -2.861  0.00429 **
Smoker          -20309.092   1648.861 -12.317  < 2e-16 ***
age                264.042      9.522  27.729  < 2e-16 ***
bmi                 22.615     25.620   0.883  0.37756
children           512.713    110.266   4.650 3.65e-06 ***
regionnorthwest   -581.704    381.215  -1.526  0.12727
regionsoutheast  -1207.011    383.109  -3.151  0.00167 **
regionsouthwest  -1227.601    382.576  -3.209  0.00136 **
smoke_bmi         1438.108     52.630  27.325  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4851 on 1329 degrees of freedom
Multiple R-squared:  0.8405,    Adjusted R-squared:  0.8395
F-statistic: 875.4 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Results of above model with new variable `smoke_bmi`

The adjusted-R-squared value has increased to 0.8395 with this inclusion. The new variable seems to be significant but renders BMI insignificant. Perhaps BMI purely now carries no information when used without smoker information, to capture any variance in the data.
From the residuals vs fitted plot, we can infer that a lot of the previous non-linearity has been removed (note the red line is almost horizontal). However, the scale location plot shows presence of non-constant variance for different fitted values.

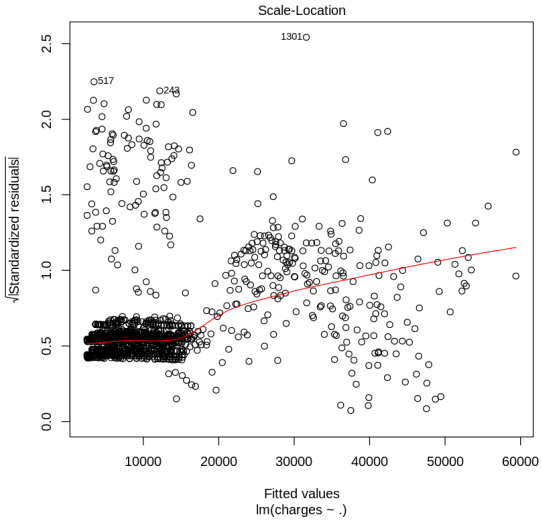
Let’s look at another plot for improving non-linearity.



Plot of Charges vs Age (left) and Charges vs Age^2 (right)

The above plot shows a slight non-linearity (curved tendency) which should be corrected with a square interaction. We thus replace Age by Age^2 (squared) for further analysis and results are below.

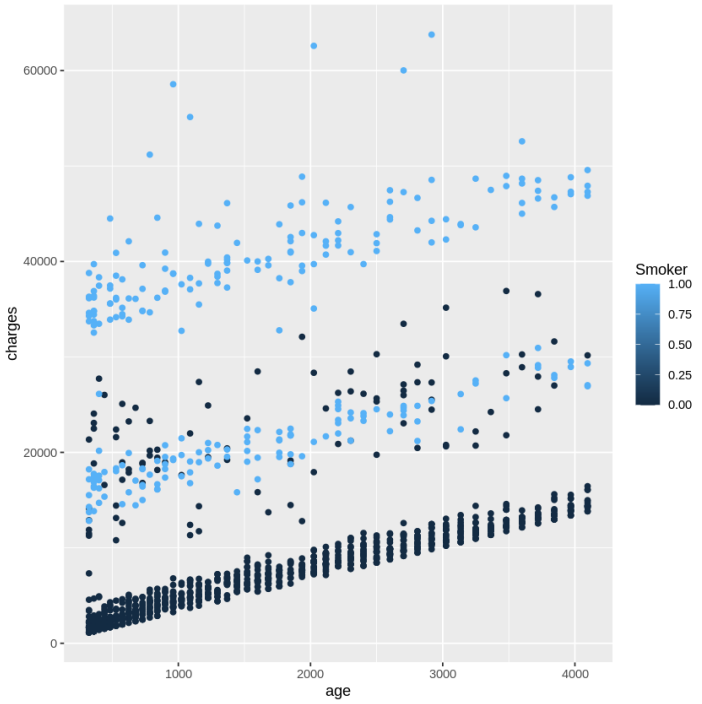
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.052e+03  8.117e+02   2.528  0.01158 *
Smoker          -2.025e+04  1.636e+03 -12.377  < 2e-16 ***
age              3.338e+00  1.179e-01  28.326  < 2e-16 ***
bmi              1.940e+01  2.543e+01   0.763  0.44566
children         6.539e+02  1.093e+02   5.982 2.84e-09 ***
regionnorthwest -5.939e+02  3.782e+02  -1.570  0.11658
regionsoutheast -1.203e+03  3.801e+02  -3.164  0.00159 **
regionsouthwest -1.225e+03  3.796e+02  -3.227  0.00128 **
smoke_bmi        1.436e+03  5.221e+01  27.510  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4813 on 1329 degrees of freedom
Multiple R-squared:  0.843,     Adjusted R-squared:  0.8421
F-statistic:   892 on 8 and 1329 DF,  p-value: < 2.2e-16
```

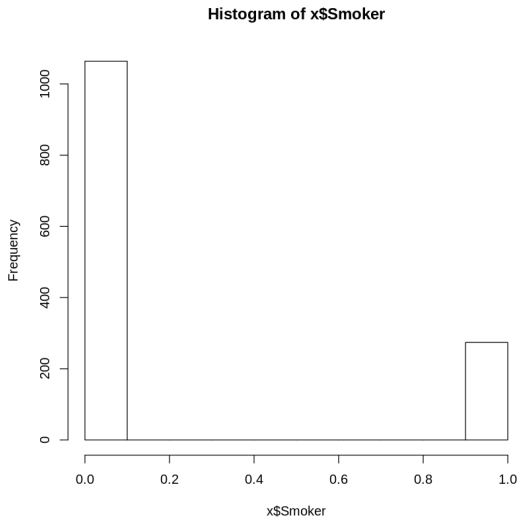


There is a slight increase in the Adjusted-R-Squared value to 0.8421.

However, the heteroscedasticity part is still problematic. We see that the data is effectively divided into two groups : One having very low variance (the black dense region in scale-location plot) and highly scattered values (more scattering as the fitted values increase).

Let’s see if there is any fundamental reason why this could be happening in the data.





Clearly, the above plot explains the groups formed in the scale location plot above.
First of all, the correlation seems linear between charges and age, and there is a clear grouping. People who don't smoke, have a low variance in terms of charges from the above plot while those who do smoke have a large variance. This corresponds to the two groups in the scale location plots.

So we can conclude that the variance in charges for smokers and non-smokers is different. This difference is violating our constant-variance assumption of the errors.

A frequency plot of the number of smokers clearly shows a huge imbalance in the number of samples. Perhaps, a lesser number of samples from the smoking population is capturing more noise and is less representative of the true distribution. Given a greater number of smoker samples, we can expect to probably improve on our fit.

We also tried to remove heteroscedasticity using non-linear transformations on the output variable (using log, square-root, square), but all only decreased the adjusted-r-squared value and also don't seem to be intuitive in general.

Conclusion

We started with all variables in the model with an adjusted-r-squared value of 0.75. We added more non-linear relationships by intuition and exploring plots in the data to further improve our value to 0.842 which is a significant improvement over the initial value.
The linearity assumption almost seems to hold but the non-constant variance assumption still seems to be violated from the plots. Since we could see distinctive grouping in the diagnostic plots and in that of smokers and non-smokers in relation to charges, we conclude that the difference in number of samples and the data-point being a smoker or non-smoker, is a significant cause of the non-constant variance violation. Given the small number of samples of smokers, this perhaps seems to be the best explanation of why there is non-constant variance in the given data.