STAT 306 Group Project Final Report

Group #: C8

Name: Muhammad Shahbaz Murtaza, Yang Wang, Zhenglin Wu, Dylan Tan
Student #: 61140299, 10260115, 26455675, 21323621

Part 1: Introduction

1.1 Motivation behind the choice of the dataset

In today's global scene, health has emerged as a top priority, heightened by the seismic impact of the COVID-19 epidemic, which killed millions and brought the globe to a halt. Despite the critical role that healthcare serves, a stark truth remains: a sizable section of the worldwide population is unable to afford the rising costs of medical care. This is where insurance companies play an important role, acting as financial guardians to bridge the gap between healthcare requirements and budgetary restraints.

Individuals and families can get vital medical services without carrying the full burden of high costs thanks to health insurance providers. Insurance companies contribute greatly to the overall well-being of their policyholders by providing coverage for a wide range of healthcare services, from routine check-ups to critical treatments.

Nonetheless, a cloud of uncertainty hangs over the health insurance pricing process. The complexities of how insurance companies estimate the cost of coverage continues to be a source of debate and investigation. Addressing this issue is critical to ensure that customers have access to fair, transparent, and affordable health insurance options.

In this project, we selected the dataset “Prediction on Insurance Charges” from Kaggle. The data is gathered from a variety of sources and contains information such as age, gender, region, and bmi for each customer. The question we pose for our report is the following:

"Which variables form the most efficient model for predicting insurance charges?"

1.2 Characterization of the Raw Data

```
> head(insurance)
  index age    sex    bmi children smoker    region   charges
1     0  19 female 27.900        0    yes southwest 16884.924
2     1  18   male 33.770        1     no southeast  1725.552
3     2  28   male 33.000        3     no southeast  4449.462
4     3  33   male 22.705        0     no northwest 21984.471
5     4  32   male 28.880        0     no northwest  3866.855
6     5  31 female 25.740        0     no southeast  3756.622
```

The dataset consists of various columns, including:

- `AGE (Numerical)`: Age of the insured individual.
- `SEX (Categorical)`: Gender of the insured individual (male or female).
- `BMI (Numerical)`: Body Mass Index, a measure of body fat based on height and weight.
- `CHILDREN (Numerical)`: The number of dependents covered by the insurance.
- `SMOKER (Categorical)`: Whether the insured is a smoker (yes or no).
- `REGION (Categorical)`: The geographic area of insurance coverage.
- `CHARGES (Numerical)`: Insurance charges the response variable associated with the insured individual.

1.3 Research question

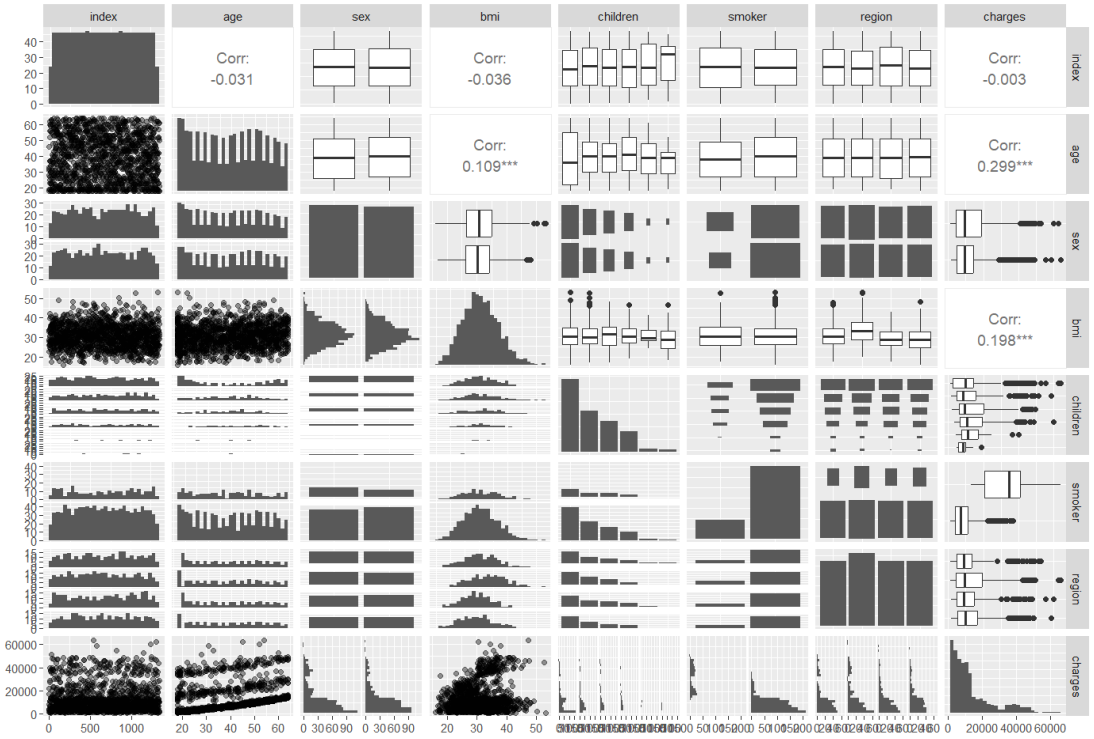
Given our explanatory variables, which variables form the most efficient model for predicting insurance charges?

Explanations and motivations behind our research question:

The research question, "Which variables form the most efficient model for predicting insurance charges?" was chosen due to its direct relevance to the insurance industry. Additionally, we found the dataset to be rich with over 1340 observations and diverse explanatory variables both categorical and continuous. This would be interesting to work with while model selection. Further, our study aims to serve as a stepping stone into more research regarding how insurance companies build their business models, exploring factors that impact their corporate decisions. It would also be intriguing to analyze if there are any biases that customers face depending on these explanatory variables. For instance: Is the demographic for older people charged higher than younger people? Are people with children more prone to incur these insurance charges? Insights to such questions could be provided by our analysis. In addition to that, from an insurer's point of view, this data could help us figure out our company's target demographic and in turn, help us curate our preexisting policies to optimize our company's profits/returns.

Part 2: Analysis

2.1 GGPAIRS/SCATTERPLOTS TO EXPLORE RELATIONSHIPS BETWEEN THE VARIABLES



Since the correlation between our variables is not too high, we do not need to be worried about multicollinearity between our predictors. The plot above allows us to be confident about that. Moreover, the distribution of each variable on the diagonal suggests that we have a diverse range of data points for each category, which is beneficial for a regression analysis as it provides a wide spectrum of values to learn from. The absence of strong linear relationships in the upper panels further reinforces the assumption that our model will not suffer from the adverse effects of highly interdependent variables. This sets a strong foundation for proceeding with the development of a robust predictive model.

2.2 MODEL FITTING:

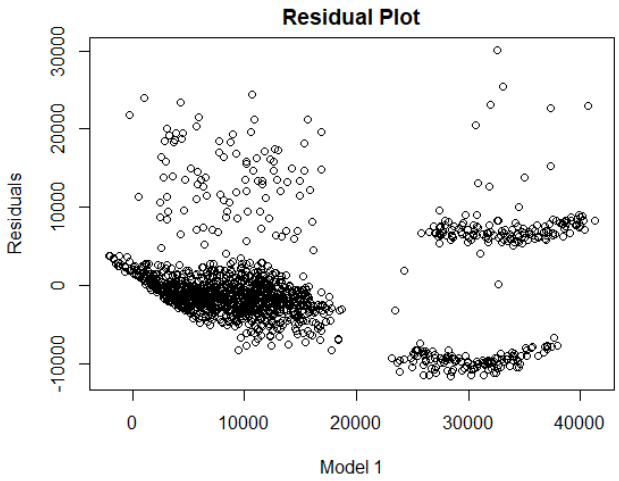
I. The Additive Model

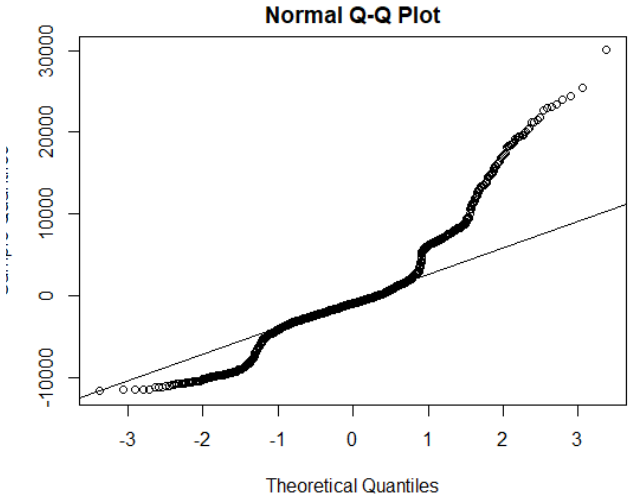
After exploring the associations between our model, we thought it would be best to fit a multiple regression model, an additive model at that, to serve as a starting point for our analysis. The first additive model is the sum of all our numerical and categorical variables. We used female (for the variable sex), 0 children (for the variable children), no (for the variable smoker), and northeast (for the variable region) as the baselines for our model. This meant that our regression coefficients for the dummies would compare each level with their respective baselines. So, our beta for sex, for instance, would compare how the charges differ for males and females - 0 meaning female, 1 meaning male.

Similarly, this one-hot encoding was used for the rest of the dummies too. Please find the following as our interpretation of the different variables:
SexMale {0, 1} - 0 suggesting the observation is female, 1 suggesting the observation is male.
SmokerYes {0,1} - 1 if the observation is a smoker, 0 if the observation is not a smoker.
Children1 {0,1} - 1 if the observation had one child, 0 otherwise
Children2 {0,1} - 1 if the observation had two children, 0 otherwise
Children3 {0,1} - 1 if the observation had three children, 0 otherwise
Children4 {0,1} - 1 if the observation had four children, 0 otherwise
Children5 {0,1} - 1 if the observation had five children, 0 otherwise
If all Children1 - Children5 would be 0, then that means the observation had 0 children. Further, either only one out of Children 1 - 5 could be 1, or they all would be 0. Similarly,
RegionNorthwest {0,1} - 1 if observation is from Northwest, 0 otherwise
RegionSouthwest {0,1} - 1 if observation is from Southwest, 0 otherwise
RegionSoutheast {0,1} - 1 if observation is Southeast, 0 otherwise
If all RegionNorthwest, Southwest, Southeast would be 0, then that means the observation was from Northeast.

Charges = -11927.2 + 257.2*(Age)* + 336.9*(Bmi)* - 128.2*(SexMale)* + 23836.4*(SmokerYes)* + 391.0*(Children1)* + 1635.8*(Children2)* + 964.3*(Children3)* + 2947.4(*Children4)* + 1116.0*(Children5)* - 380.0*(RegionNorthwest*) - 1033.1(*RegionSoutheast)* - 952.9(*RegionSouthwest)*

From our summary of this regression model, we found that our Adj-R^2 was 0.7497 which means the model fit the data well but the model fitted could be improved. To have a better look at the hidden patterns, we decided to take a look at the residual and the QQ-plot to gauge the linearity of the model.

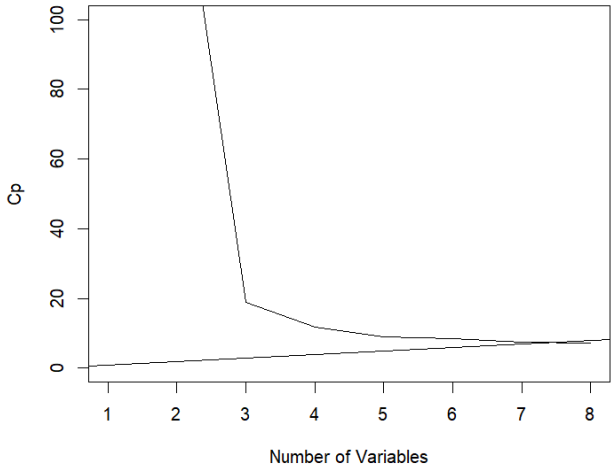




In our residual plot, we found a few issues which could be improved. First being that the residuals do not seem to be dispersed randomly, there seems to be some sort of clustering and pattern that may be preventing that from happening. Second, the residuals do seem to have some weak heteroscedasticity which could be looked upon. So, transformations may be a good idea to explore any non-linear patterns.

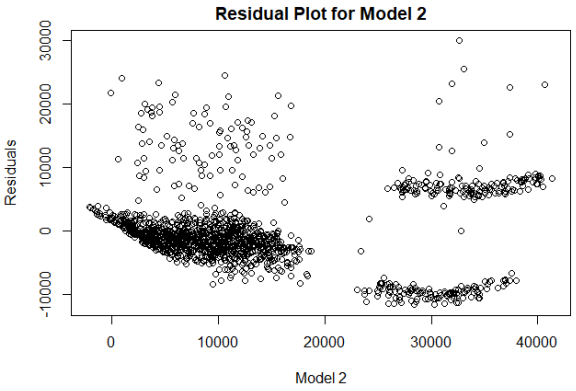
II. Exhaustive Selection Modeling (Model 2)

Before moving forward, we decided to use regsubsets() in R and specifically the exhaustive method of model selection to pick the best performing model and the number of parameters it would require. To help us make better calculated and informed decisions, we decided to look at Mallow's CP and Adj-R^2 statistics. The Mallow's Cp plot below helps us choose our best model:



Since our abline (0,1) intersects our Mallow's Cp curve at somewhere between the number of parameters 7 and 8, we decided to look at the Adj-R^2. We do that because that gives the lowest Mallow's Cp and it is also closest to the number of predictors we have.

The Adj-R^2 we found for the model with 7 parameters was 0.749 whereas the model with 8 parameters gave us an Adj-R of 0.750, so there was not much of a difference but we decided to go with the model with 8 parameters. However, the issue we were facing was that with 8 parameters only few of our levels/factors of our category Children were significant - the rest were not. Since we could not perform any filtering on rows of our dataset, we could either drop or keep the whole variable (not just a few factors/levels). Due to this, we kept all of our variables, but just dropped Sex as as a category as it was not significant in any of the models, nor was it significant in the additive model that we found in part I. To take further steps from this, we found the residual plot:



Since the pattern from Model 1 still remains, we can consider interaction terms between our categorical and numerical variables.

III. Interaction Analysis (Model 3)

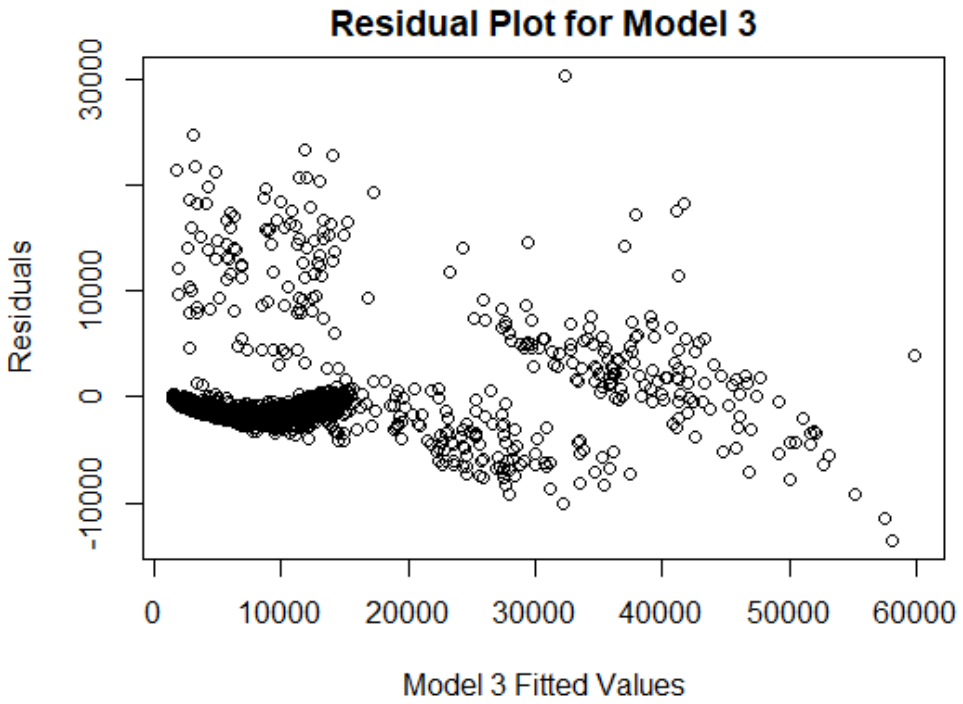
Since the pattern still remains, we can consider interactions with our terms. To see how each categorical variable is related to our response, *Charges*. So, we tried out three different models with all combinations of interactions that could occur for all three of our remaining explanatory variables: **Smoker**, **Children**, and **Region**.

After plotting each model, we decided to use the Adj-R^2 as our measure for deciding if we were improving from Model 2 or not. In our analysis, we found that the Adj-R^2 for the model with smoker interactions was 0.8401, which suggested better model fitting than we ever encountered before in our analysis. So, we decided to go ahead with that model.

However, upon further inspection of our model summary, we found that only one interaction term across all three variables had a very high statistical significance (with P-value around 2 x 10^-16) which was the interaction term between Smoker and BMI. With that in mind, we fit an improved model with all the terms from Model 2 with an additional **Smoker: BMI** term. Our results from this matched the results we achieved previously when including interactions with all numerical variables, implying that the pattern was caught with the **Smoker: BMI** interaction term. To corroborate

an association between the two, we found proven existing research suggesting a positive association between smoking and BMI. (2018, Taylor et. al)

For a deeper look into our model, we visualized the residuals and found the following results:

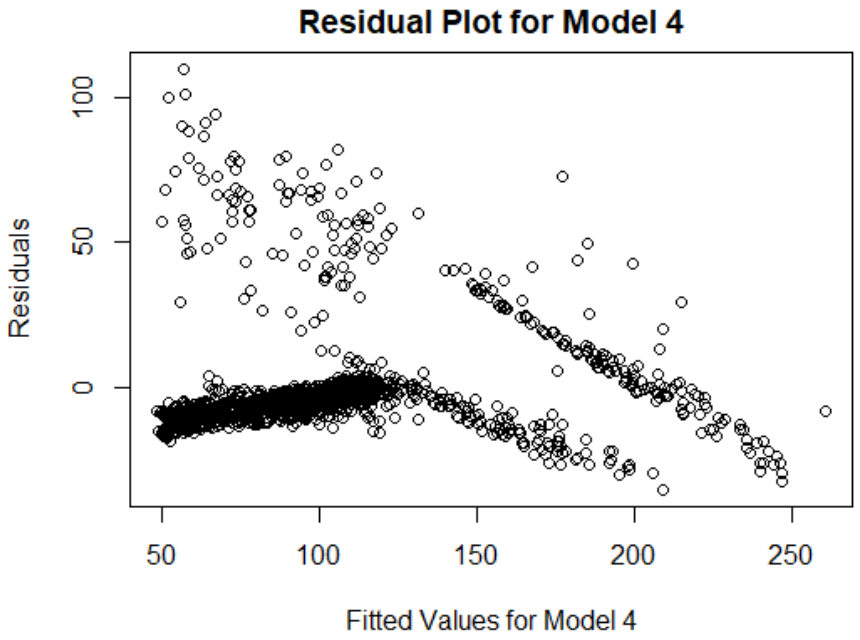


From Model 2 to Model 3, there seems to be variability and randomness in our residual plot but the clustering still exists. To catch any other implicit pattern, we perform transformations on our response in the next phase and explore if doing so leads to any improvement.

IV. Transformations (Model 4 & 5)

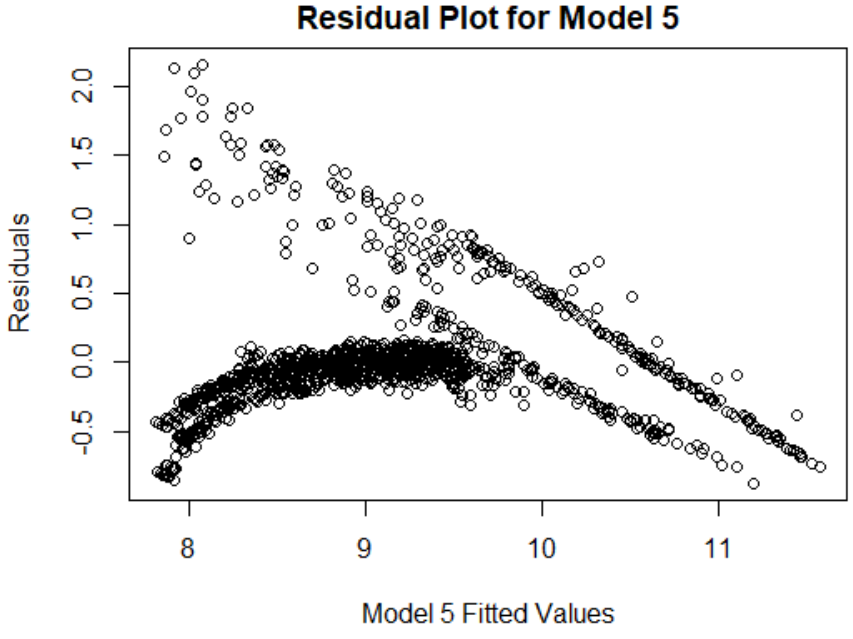
Since there seemed to still be some lurking pattern within our model and weak heteroscedasticity, we decided to transform our response, *Charges*.

For Model 4, we began with a square root transformation since the lower half of the residual plot resembles a square root curve. After fitting the same model from Model 3, but changing *Charges* to *sqrt(Charges)*, we found the Adj-R^2 to be 0.8238 which was still a good measure of fit, however not as good as we had earlier. This led us to not go ahead with the transformed response. We plotted the residual to support our decision:



From Model 3 to Model 4, the residuals divide themselves into two distinct sections which were not as distinct in Model 3. Due to the differences being so prominent, and there being a curve and a linear line, we decided to not move forward with this transformation, as it led to more questions than answers about the model fitted.

Similarly, we tried a logarithmic transformation due to the resemblance of the residuals with the log downside graph. Instead of using *sqrt(Charges)* as our response, we used *log(Charges)* as our response in this model. From our model summary, we found that our Adj-R^2 fell to a value of 0.7815 which meant the model fitted was worse than the one before. The following was the residual plot we obtained, to help make a decision:



We saw similarities in our residual plot from Model 4 and Model 5, hence as the Adj-R^2 was weaker for this Model, and the pattern still remained, we decided not to go ahead with the log transformed variable. However, it is important to note that the residual standard error value for Model 5 was significantly lower than what we had in Model 3, which brought forth another concern with our Model 3: a very high residual standard error. The aforementioned issue will be discussed later in the report.

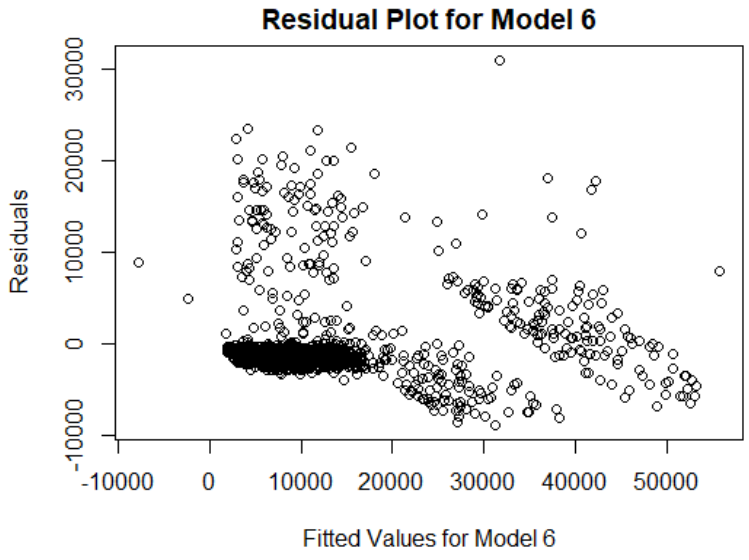
V. Adding Quadratic Terms (Model 6)

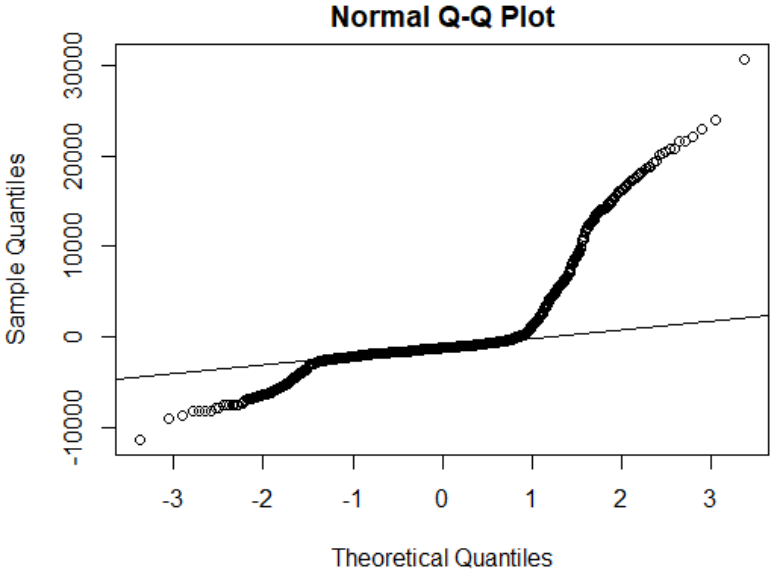
Since the curve still seemed troublesome, we thought that may be due to non-linearity. To help enhance model accuracy and prediction, we decided to test some quadratic and cubic terms by adding squared and cubed terms (of our numerical variables) and tested out multiple combinations of regression models. From the summaries, we plucked out the individual significant terms with very low p-values to add to our final model. The final model turned out to be:

Charges = 26178.842 + 3.324*(Age^2)* - 2653.283(*BMI*) + 93.916*(BMI^2)* - 1.054*(BMI^3)* + 1447.208*(SmokerYes)(BMI)* - 20556.762*(SmokerYes)* + 792.282*(Children1)* + 2001.558*(Children2)* + 1385.752*(Children3)* + 3777.566(*Children4)* + 2588.372*(Children5)* - 643.578*(RegionNorthwest*) - 1121.092(*RegionSoutheast)* - 1271.091(*RegionSouthwest)*

The Adj-R^2 for the aforementioned model turned out to be 0.8456 and Model 6 had lower residual standard error than Model 3 as well, so we decided this to be the best model we could find.

We visualized our results using a residual and a QQ-plot as we did for Model 1 to compare differences and improvements.



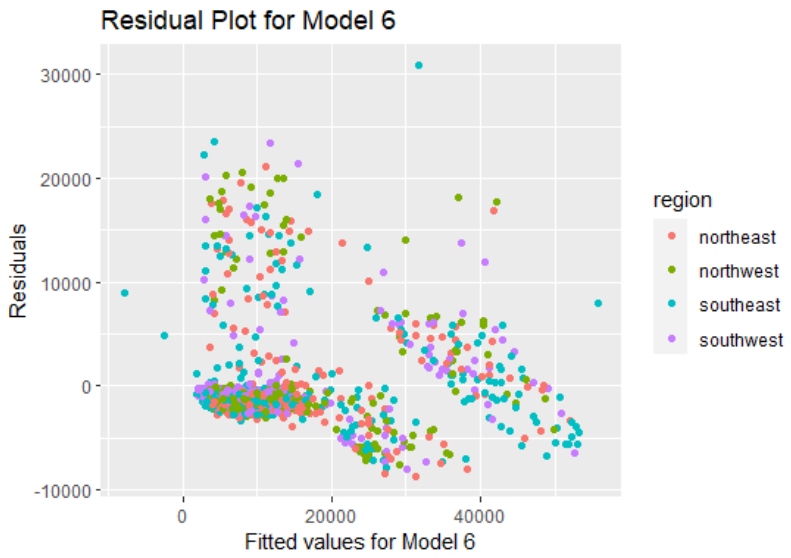


From the Normal Q-Q Plot, the normality of residuals seems clear. However, the upper tail seems heavier than the lower tail which may suggest some uncaught pattern. In addition to that, the residual plot seems more random than what we started with, suggesting an improved model.

Note: For future analysis, we looked at the residuals for our categorical variables separately since we suspected some clustering in our residual plot. To catch that pattern, we plotted the following residual plots with respect to all three of our categorical variables. From our findings below, there seem to be two separate clusters for our category - Smoker, which after further investigation could lead to more findings. (find the plots on pg. 12) However, due to concerns with overfitting, we decided against more exploration of the Smoker variable.

To improve upon our model, we also decided to conduct Principal Component Analysis, or more commonly known as PCA, for dimensionality reduction. Since three of our covariates were categorical, including them in PCA was not ideal. However, we transformed them through one-hot encoding for cohesive results. In our PCA plot, we found that all the components were almost equally divided in their explanation of the standard deviation for the data, which meant that our model could not be reduced further.

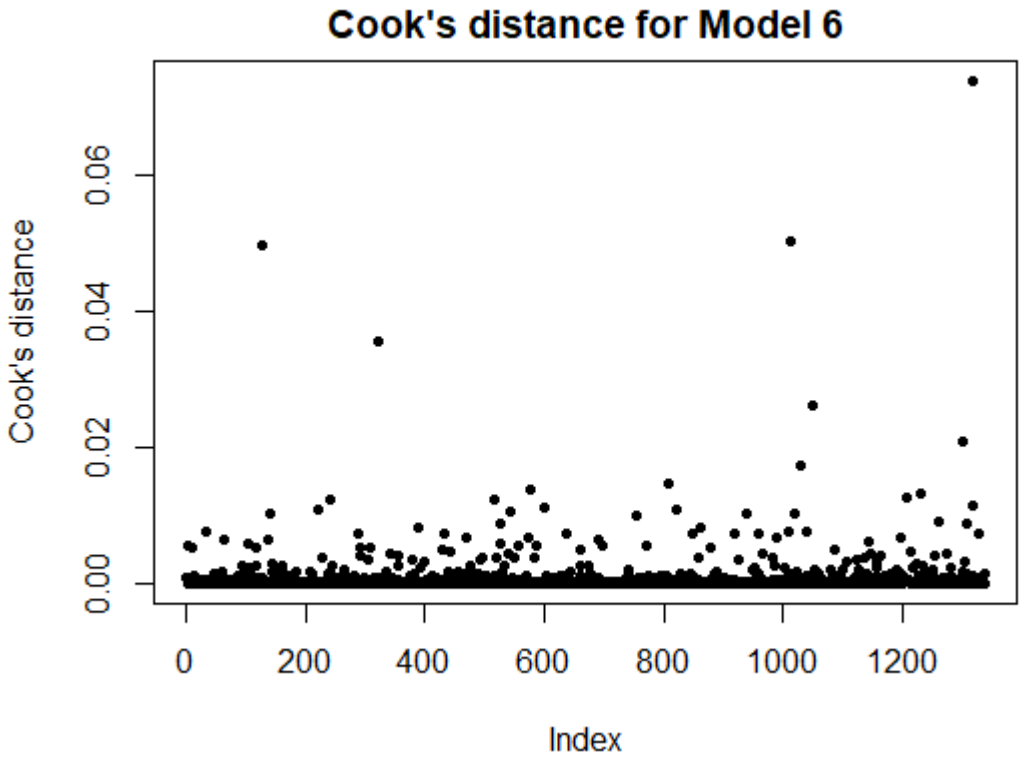
In addition to that, AIC and BIC were both examined throughout this model selection procedure, and Model 5 seemed to achieve the best results. However, as our main purpose for the model fitting was to explain the factors behind insurance charges rather than prediction, this metric was not utilized.


Residual Plot for Model 6
Residuals
30000
20000
10000
0
-10000
0
20000
40000
Fitted values for Model 6
region
northeast
northwest
southeast
southwest


Residuals
30000
20000
10000
0
-10000
0
20000
40000
Fitted values for Model 6
smoker
no
yes


Residuals
30000
20000
10000
0
-10000
0
20000
40000
Fitted values for Model 6
children
0
1
2
3
4
5

2.3 Examining Cook's Distance for our Final Model (Model 6)



From the plot above, it appears that most data points have a Cook's Distance value close to zero, which indicates that they are not influential with respect to the fitted regression model. This suggests that removing these points would not have a large effect on the coefficient estimates.

However, there are a few points with higher Cook's Distance values, particularly one point that stands out with a Cook's Distance well above 0.06. These points are potentially influential, meaning that they could have a significant effect on the parameter estimates or the prediction if they were to be removed.

Since our final model, Model 6, includes nonlinear terms for age and BMI, as well as an interaction term between smoking status and BMI, the presence of influential points could suggest that these terms are not capturing the relationship between the predictors and response variable adequately for all observations.

2.4 Final Model

Charges = 26178.842 + 3.324*(Age^2)* - 2653.283(*BMI*) + 93.916*(BMI^2)* - 1.054*(BMI^3)* + 1447.208*(SmokerYes)(BMI)* - 20556.762*(SmokerYes)* + 792.282*(Children1)* + 2001.558*(Children2)* + 1385.752*(Children3)* + 3777.566(*Children4)* + 2588.372*(Children5)* - 643.578*(RegionNorthwest*) - 1121.092(*RegionSoutheast)* - 1271.091(*RegionSouthwest)*

Part 3: Conclusion

Our study embarked on a detailed exploration of health insurance charges, utilizing advanced statistical methods in R. We meticulously evaluated various models, considering interactions and transformations to enhance understanding and predictive accuracy. Our final model, incorporating quadratic and interaction terms, achieved an Adjusted R-squared of 0.8456, signifying a robust fit to the data. Key findings highlight the significant impact of smoking status, BMI, and age on insurance charges, with the interaction between smoking and BMI being particularly influential. This analysis not only reinforces the intricate relationship between health-related variables and insurance costs but also underscores the complexity of insurance pricing mechanisms. Despite achieving a strong model fit, certain limitations, such as potential unaccounted non-linear relationships, were identified. Future research could explore these aspects further, perhaps considering additional variables or alternative modeling techniques such as Cross Validation could be helpful in achieving a better fitted model. This project has been an enriching experience in applying statistical concepts to real-world data, providing valuable insights into the dynamic field of health insurance.

REFERENCES:

1. Taylor, A. E., Richmond, R. C., Palviainen, T., Loukola, A., Wootton, R. E., Kaprio, J., Relton, C. L., Davey Smith, G., & Munafò, M. R. (2018). The effect of body mass index on smoking behavior and nicotine metabolism: A Mendelian randomization study. *Human Molecular Genetics*, *28*(8), 1322–1330. https://doi.org/10.1093/hmg/ddy434

2. The Devastator. (2023). *Prediction of Insurance Charges*. Kaggle.com. https://www.kaggle.com/datasets/thedevastator/prediction-of-insurance-charges-using-age-gender