STAT 306 Group Project Final Report

Group Mumber:
Xing Liu 92833748
Vishnu Rengan 42576561
Minkyung Yun 65258436
Amir Darijani 19233741

Introduction:

In recent years, personal health cost is a big topic all over the world, more and more people are starting to focus on their personal health and consequently the attention on the medical insurance costs has been highly raised. So our group wants to find out what factors could have some effects on the medical insurance costs. In this report, our dataset of interest will be the "Prediction of Insurance Charges" dataset from kaggle.

In this dataset we have the following variables:

- Age - the age of the customer [1]
- Sex - the sex of the customer[1]
- Bmi - the body mass index of the customer[1]
- Children - number of children the customer has[1]
- Smoker - whether or not the customer smokes[1]
- Region - where the customer lives[1]
- Charges - the insurance charges[1]

From preliminary analysis, we know that there are no missing values and that there are 1338 observations. Our variable of interest, or response variable, is the charges, while the rest of the variables are explanatory. In this dataset, the variables sex, smoker, and region are categorical while the variables age, bmi, children, and charges are numerical.

Recently, the validity of BMI as a measure of a person's health has been questioned.[2] For instance, many athletes with a high muscle mass will have "unhealthy" BMIs when in reality they are some of the healthiest people.[2] As such, we want to explore whether BMI is significant in predicting insurance charges of an individual by looking at several linear models.

Moreover, since age, number of children, region, smoking status and even sex could be related to the insurance charge of a customer, we want to find out the relationship between those factors and the response variable (insurance charge) and by any chance, if there exists any interaction between any of those variables.

Data analysis:

Relationship between response and explanatory variables:

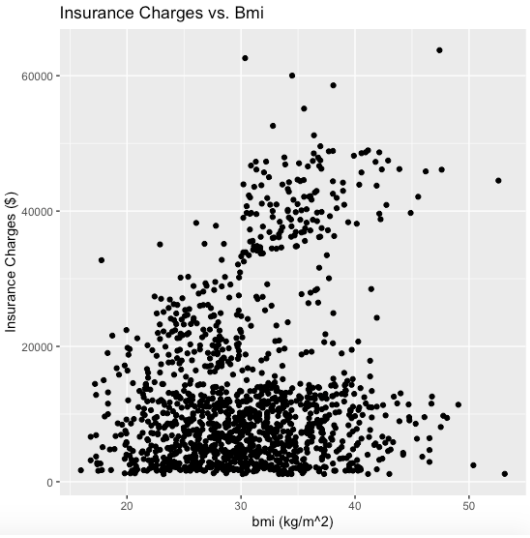


figure 1: Medical Costs vs. bmi

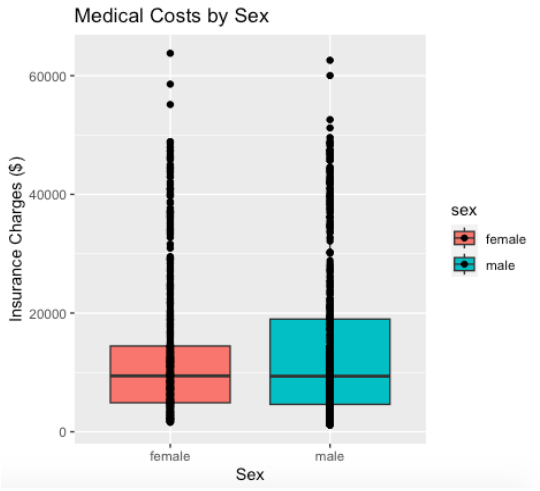


figure 2: Medical Costs by Sex

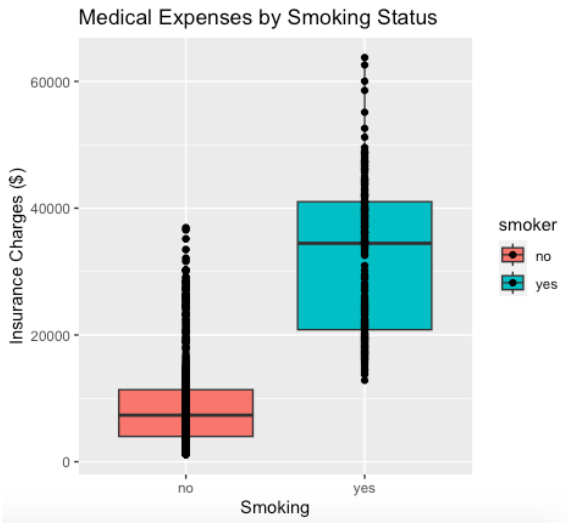


figure 3: Medical Costs by Smoking Status

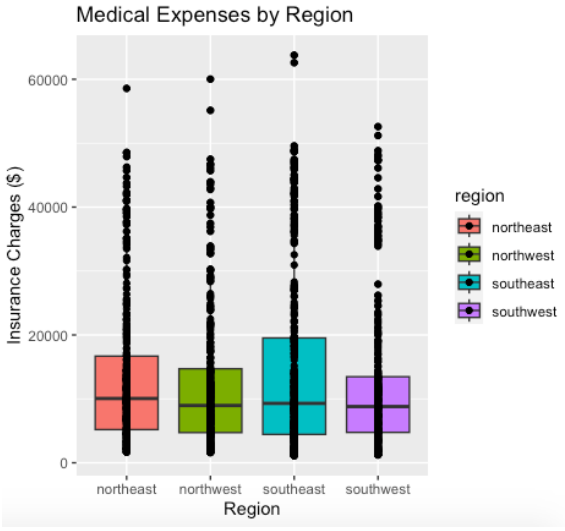
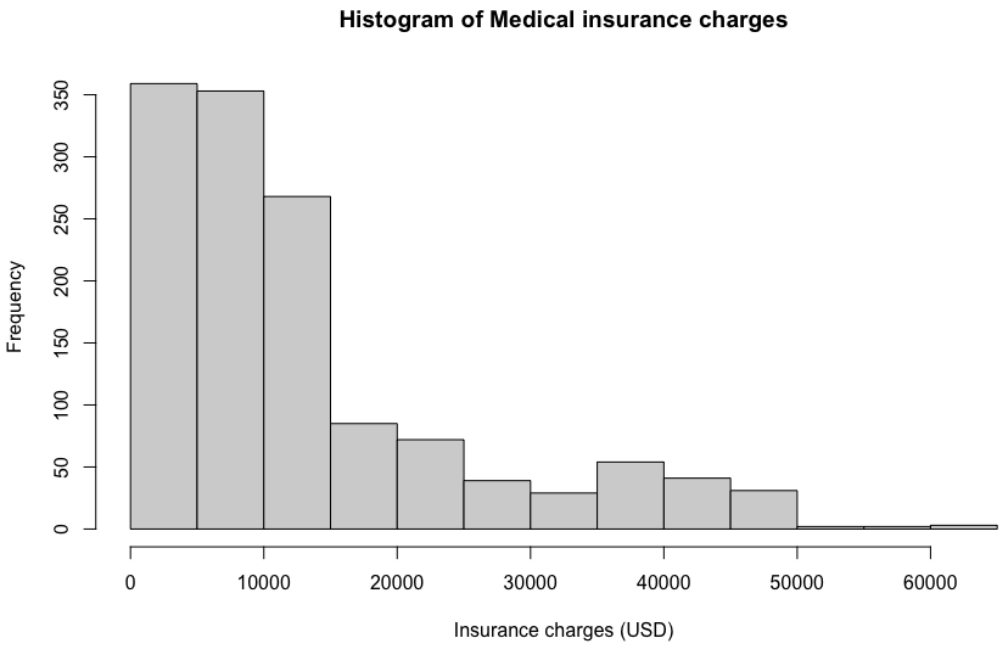


figure 4: Medical Costs by Region

Based on the scatter plot of medical costs and bmi from figure 1, there seems to be a slightly positive relationship, with a huge number of unusual observations. In figure 2, we can see that the median of medical expenses is similar between female and male. However, the box plot from figure 3 shows that medical expenses among the smoking group is considerably higher than that of non-smoking people. In terms of the relationship between medical costs and regions in figure 4, insurance charges appear fairly similar across all regions.

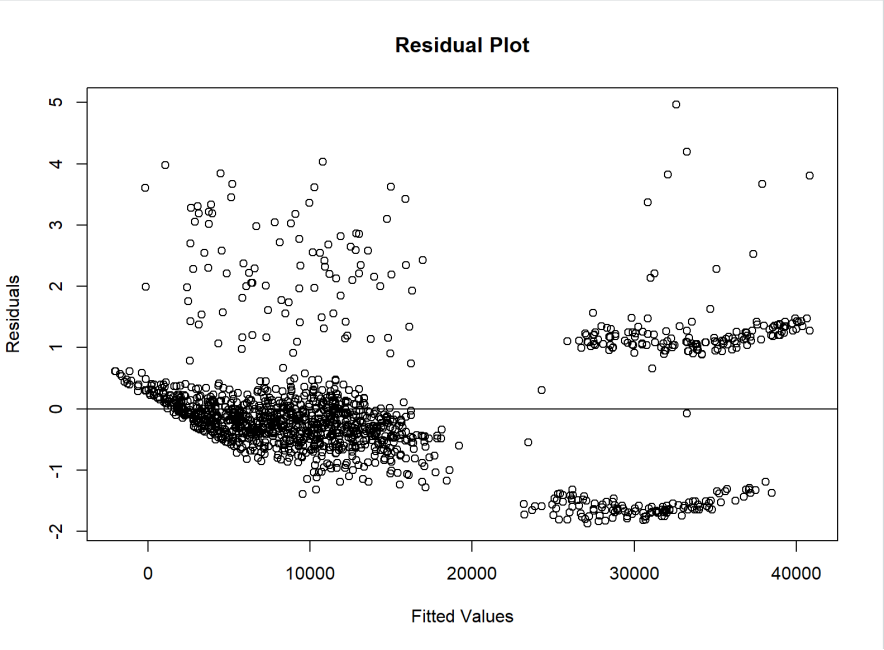
Data distribution:



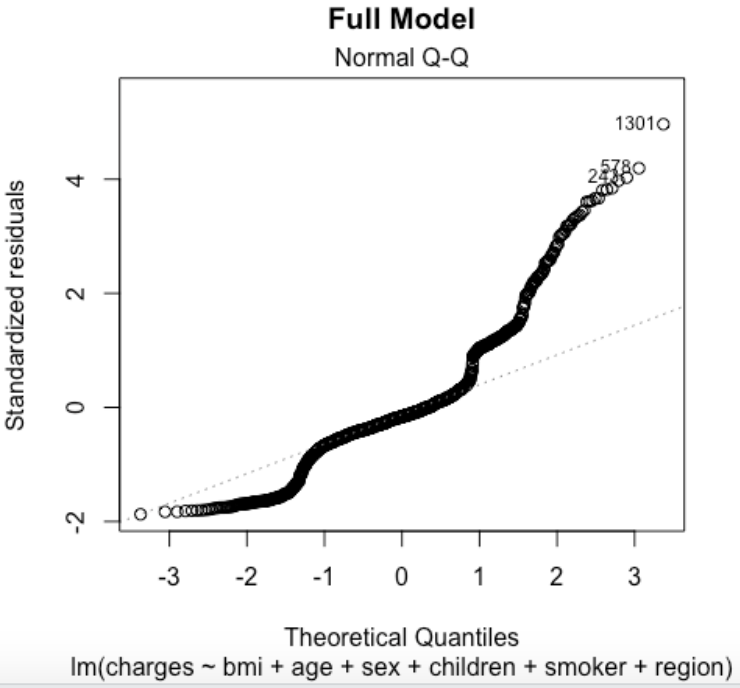
From the histogram we can see that our original data distribution of the insurance charge is right skewed, which means most customers in the USA get charged low on their medical insurance. However, this could affect our model construction later since some models need the assumption of normality.

Baseline Model:

We first fit a full model using all the variables



Based on the residual plot, we can see the right skewness of the data. In addition, our model does not seem to be doing as well for larger fitted values as the residuals are further away from the zero line. Lastly, there seems to be many outliers with standardized residuals above



In addition, the QQ-plot clearly shows that each end of the tail is in the opposite direction along the line. These heavy tails indicate that the assumption of normality is violated and there are a large number of outliers.

From the summary, we can see that sexmale and regionnorthwest have relatively large p-values. BMI has a very low p-value and seems significant. Overall, the adjusted R-squares is 0.7494 and residual standard error is 6062. Therefore, based on these results, our model will include age, bmi, children, smoker and region as explanatory variables for now. Our next step is to choose the best model and explore any interaction in this model.

Interaction term exploration:

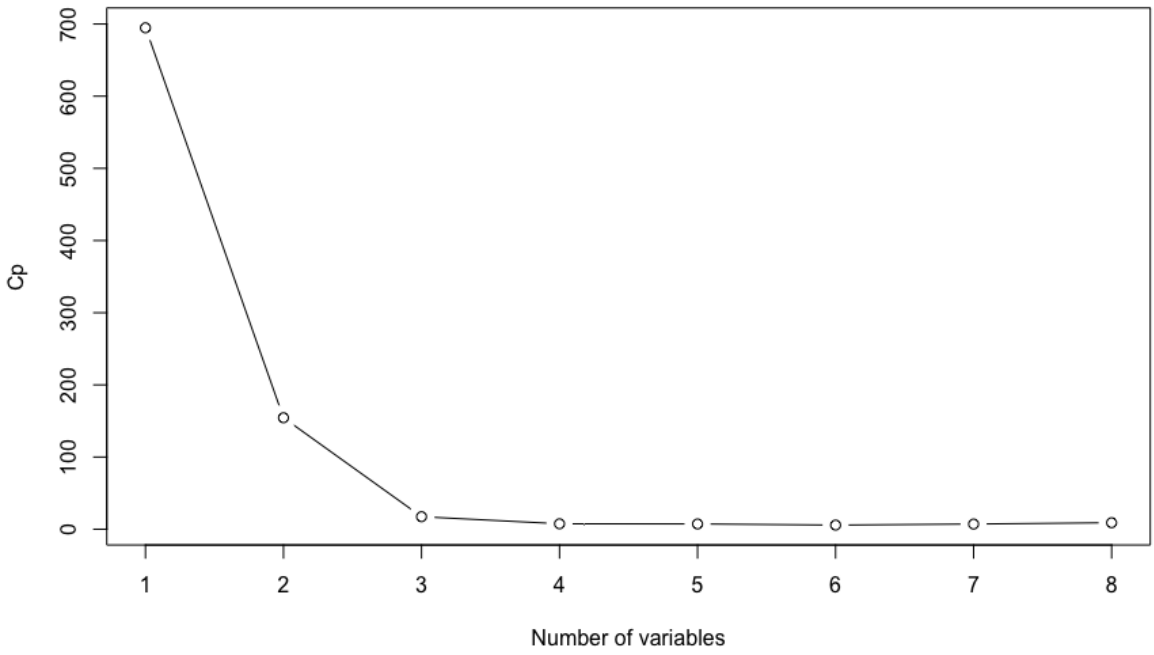
From the cdc paper, we are told that "older adults tend to have more body fat than younger adults for an equivalent BMI"[2] and that "women have greater amounts of total body fat than men with an equivalent BMI"[2]. As such, we decided to explore the interaction between BMI and Age as well as the interaction between BMI and Sex. So in order to testify our idea, we added two extra terms (bmi*age, bmi*sex) to our model. According to the output from R, p-values of both interaction terms are very large (0.929945 and 0.913838), which indicates that they are not significant in this model and we don't need the interaction terms here.

Model selection:

Since our goal for this project is to figure out the possible relations between medical insurance costs and some related factors, we decided to do a model selection in R.

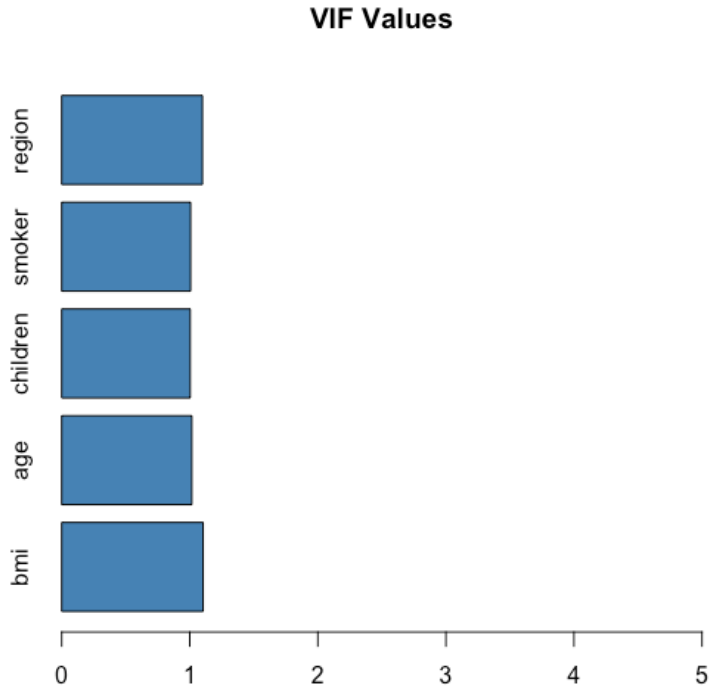
Using regsubsets we can look to see which variables are chosen as we fit models of varying sizes.

As we can see from the output, the first two variables chosen are smoker and age. After that all subsequent models contain BMI. This suggests that BMI is a relatively important variable in predicting the insurance charges for an individual.



From the Cp plot,we can also notice that Cp reaches the lowest when our model has 6 input variables in total. Therefore, combined with the previous results from regsubsets, our model will include the terms: age, bmi, children, smoker, region (two of the sub-category are chosen).

By applying the cross-validation (training and testing set), we split the data into 60% (training set) and 40% (testing set). Then, we build our model on the training set, and calculate the RMSE on the testing set and evaluate the performance or our model. The residual mean square root (RMSE) value of the model that we have been proposed from the model selection turns out to be 5725.86, which is smaller than that of the full model with a value of 5726.94. Based on both RMSE values, it suggests that our selected model would have better performance than the full model.



Furthermore, the results of the variance inflation factor (VIF) for our model indicate that there is no concerning presence of multicollinearity since all values are small, around from 1.00 to 1.10, which is far less than 10.

Final Model:

After fitting a final model which has now removed the variable sex, our model adjusted R squared has remained the same at 0.7496 while the residual standard error has decreased slightly to 0.6060.

Principal Components Analysis:

In this section, we will use principal components analysis to determine which numerical variables are the most important in explaining the variability in the data.

From R, we see that the first 2 components explain 70% of the variance in the data. Looking more closely at the components, it seems like age and bmi are both important for component 1 with bmi being more important for component 2.

Conclusion:

In summary, we will get rid of the sex variable in our final model and keep all the other factors. Therefore, the final model is : Medical insurance charge = -11990.27 + 338.66*bmi + 256.97*age + 474.57*children + 23836.30*smoker - 352.18*region-northwest - 1034.36*region-southeast - 959.37*region-southwest. From the model, we can see that bmi, age, and children have a positive effect on the medical insurance charge, namely, a customer with a higher bmi, older age and more children will have more insurance charge. Besides, among these factors, smoking status has the most influence on the charge since it is the largest coefficient and it is always chosen by the exhausted model selection process. In addition, according to the PCA and the exhausted model selection process, bmi is a relatively significant variable in terms of medical insurance costs prediction.

Also, since Region is a categorical variable with 4 levels, we treat the Northeast region as the baseline. We can see that for each dummy variable of the region term, their coefficients are all negative, which means that on average, the Northeast part of the USA has the highest medical insurance charge. In addition, according to the PCA and the exhausted model selection process, bmi is a relatively significant variable in terms of medical insurance costs prediction.

References:

1. Data from kaggle: https://www.kaggle.com/datasets/thedevastator/prediction-of-insurance-charges-using-age-gender

2. Department of Health and Human Services, Centers for Disease Control and Prevention. (2015). Body mass index: considerations for practitioners. https://www.cdc.gov/obesity/downloads/bmiforpactitioners.pdf