STAT423 Final Project

Yuning Hu, Luna Lu, Yuhan Zhang

Part 0: Data Explore

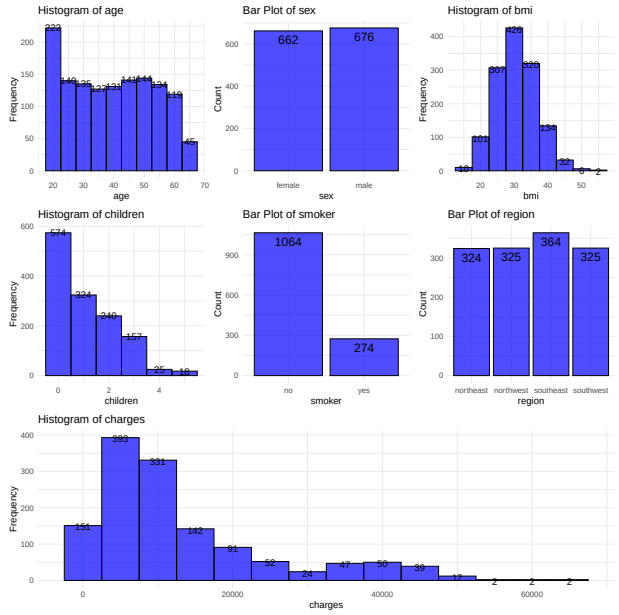
The dataset used in this study is the Medical Cost Personal Datasets [1]. This dataset contains 6 predictors, 1 response variable, and 1338 observations. This dataset contains the individual medical costs billed by health insurance, and 6 other descriptive variables about the individual.

Response variable:

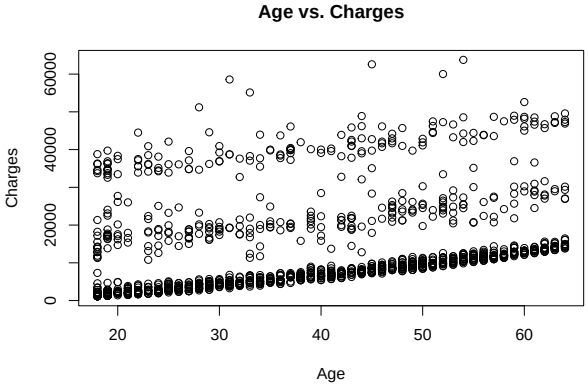
- charges (numerical): Individual medical costs billed by health insurance.

Input variables:

- age (numerical): Age of primary beneficiary.
- sex (categorical): Insurance contractor gender (female, male).
- bmi (numerical): Body mass index, a measure of body weight relative to height using the formula Weight (kg) / Height $(\text{m})^2$. ideally 18.5 to 24.9.
- children (numerical): Number of children or dependents covered by health insurance. It is recognized as a numerical variable in this study based on better model results.
- smoker (categorical): Whether the beneficiary has smoking behavior.
- region (categorical): The beneficiary's residential area in the US (NW, NE, SW, SE region).



We first checked there are no NA values in the data. Then we visualize each factor so that we can have a brief understanding of the data. We can see from the plot that the sex and region of in this data are evenly distributed. There are more data of patient under age 20 than other age intervals. For BMI, the data is approximately normally distributed with mean 30. For number of children, there are lesser data point with more children. Also, in the data, the ratio of smokers and non-smokers is approximately 4:1. For our target, the plot of charges is skewed to the right, the majority of charges is below 20000.

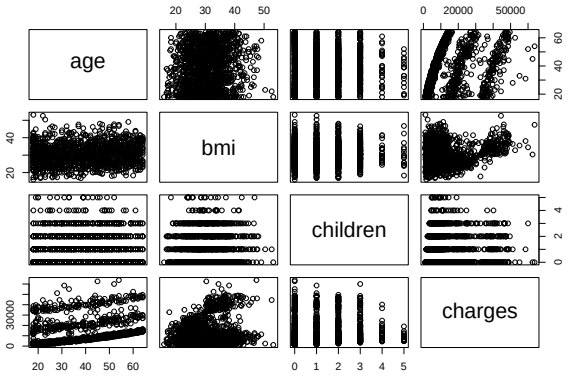


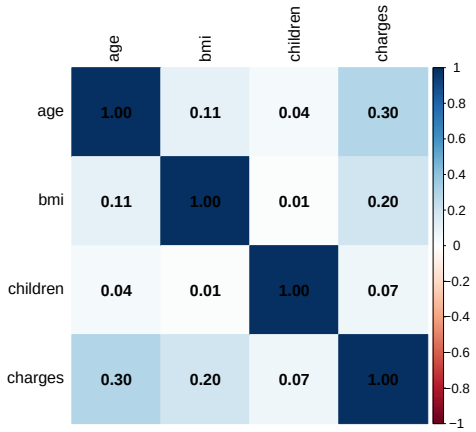
Then we explored the relationship between charges and age. As we can see from the plot, it is significant that the points are divided into three groups. We will investigate the division in later parts.

Part 1. Are there any correlated factors?

For the purpose of improving model performance, we would need to pay attention to the variables who are correlated, as it may result in unstable coefficients and hardship in interpretation.

To visualize their correlation, we can first check the scatter plot. However, categorical variables like "sex", "smoker", and "region" would not appear properly. From the scatterplot we can see that among the numerical variables, bmi has a weak positive correlation with charges and age has a weak to moderate positive correlation with charges. The correlation of other numerical variable pairs is not obvious just by eye-balling.





To get more detailed information on the correlation, we can check the correlation matrix.

From the correlation matrix we can see that it aligns with our observations from the scatterplot. Also we can see that all numerical variables' correlation to charges are positive. Since the largest correlation coefficient is 0.30, we would like to take a further step to decide whether to perform further data transformation. We would use the Variance Inflation Factor (VIF), which measures how much the variance of an estimated regression coefficient rises when the predictors are associated.

```
##              GVIF Df GVIF^(1/(2*Df))
## age      1.016822  1        1.008376
## sex      1.008900  1        1.004440
## bmi      1.106630  1        1.051965
## children 1.004011  1        1.002003
## smoker   1.012074  1        1.006019
## region   1.098893  3        1.015841
```

Since both general VIF and adjusted VIF for all variables are only slightly greater than 1, we can conclude that there is no problematic multicollinearity, and thus we would not perform further data transformation such as Principal Component Analysis (PCA).

Part 2. Do patients of one sex consistently pay more in medical costs compared to the other sex?

```
##                    Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)     -11938.5386  987.81918 -12.0857530 5.579044e-32
## age                256.8564   11.89885  21.5866552 7.783217e-89
## sexmale           -131.3144  332.94544  -0.3944020 6.933475e-01
## bmi                339.1935   28.59947  11.8601306 6.498194e-31
## children           475.5005  137.80409   3.4505546 5.769682e-04
## smokeryes        23848.5345  413.15335  57.7232020 0.000000e+00
## regionnorthwest   -352.9639  476.27579  -0.7410914 4.587689e-01
## regionsoutheast  -1035.0220  478.69221  -2.1621870 3.078174e-02
## regionsouthwest   -960.0510  477.93302  -2.0087563 4.476493e-02
```

First we fit a full linear model, and we can see that sex is not a significant factor that influences medical cost. However, based on Green and Pope's study [2], there is a difference between the medical cost based on different sex, where female costs more on the reproductive related diseases. We fit a linear model only consisting the sex term, and the summary output is:

```
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 12569.579   470.0717 26.739706 1.626108e-126
## I(sex)male   1387.172   661.3309  2.097547  3.613272e-02
```

Since we can identified smoking as a significant factor in the full model based on the p-value, we are concerned that the effect of sex on charges may be influenced by smoking patterns. If the proportion of smokers differs between sexes, the observed effect of sex might be due to its correlation with smoking. To test this, we formulate the following hypotheses:

H_0 : Proportion of smoker is the same between sex, H_1 : Proportion of smoker is not the same between sex

And we conduct a t test to test the hypothesis:

```
##
##  Welch Two Sample t-test
##
## data:  female.smoker and male.smoker
## t = -2.7961, df = 1324.9, p-value = 0.005248
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.10463411 -0.01834807
## sample estimates:
## mean of x mean of y
## 0.1737160 0.2352071
```

From the t test, we can see that the P-value is 0.005248, which is significant at level $\alpha = 0.05$, which indicates that there is a difference in the proportion of smoker between sexes. Given that smoking is a significant predictor, the previously observed effect of sex in our regression models may be confounded by its correlation with smoking.

To see how these two terms are interacted, we fit another linear model including both the sexes and smoker with the interaction between these two terms:

```
##                            Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)              8762.2973   318.5825 27.504014 2.726211e-132
## I(sex)male               -675.0926   457.0329 -1.477120  1.398794e-01
## I(smoker)yes            21916.6990   764.3671 28.673002 2.913341e-141
## I(sex)male:I(smoker)yes  3038.1023  1020.2005  2.977946  2.954255e-03
```

From the output, after including smoking and its interaction with sex, the coefficient for sex alone is no longer significant. However, the interaction term is statistically significant at level $\alpha = 0.05$, indicating that male smokers, on average, pay an additional \$3,038.10 in medical costs compared to non-smoker female.

We also apply LASSO regression that penalizes less important predictors by shrinking their coefficients toward zero.

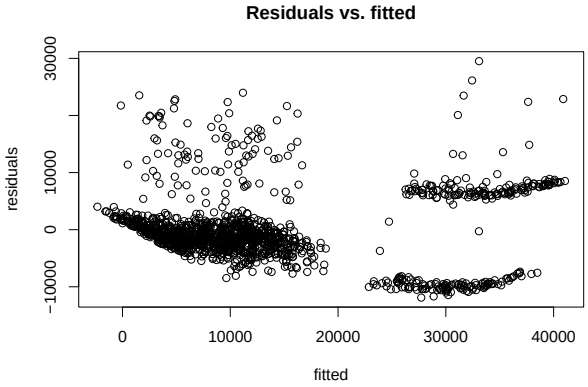
```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                            s1
## (Intercept)     -11566.9505
## age                252.9685
## sexmale              .
## bmi                322.5230
## children           418.4944
## smokeryes        23659.0533
## regionnorthwest      .
## regionsoutheast   -547.1304
## regionsouthwest   -513.6591
```

We can see that the coefficient for sex is shrunk to zero, indicating that sex alone does not significantly contribute to predicting medical costs after accounting for other factors. Thus, with these models, we conclude that there is a difference of charges in different sex-smoker groups instead of sex groups.

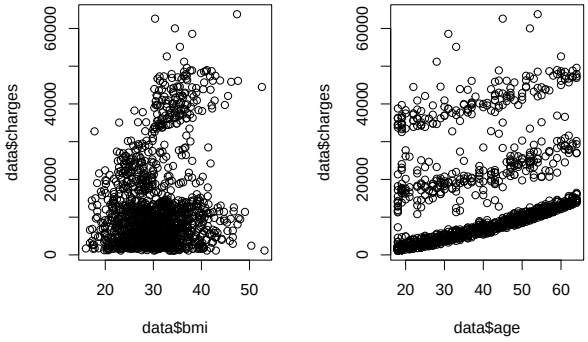
Part 3

We can see from the full model in part 2 that there are several terms in the summary with a large p-value, which means that we need to perform a parameter selection. We select the term `age`, `smoker`, `bmi` and `children` based on the P-value with significant level $\alpha = 0.05$, and fit a new linear model.

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -12102.7694  941.98394 -12.848170 1.051649e-35
## age            257.8495   11.89639  21.674608 1.748342e-89
## smokeryes    23811.3998  411.21971  57.904325 0.000000e+00
## bmi            321.8514   27.37763  11.755999 1.973987e-30
## children       473.5023  137.79167   3.436364 6.077158e-04
```



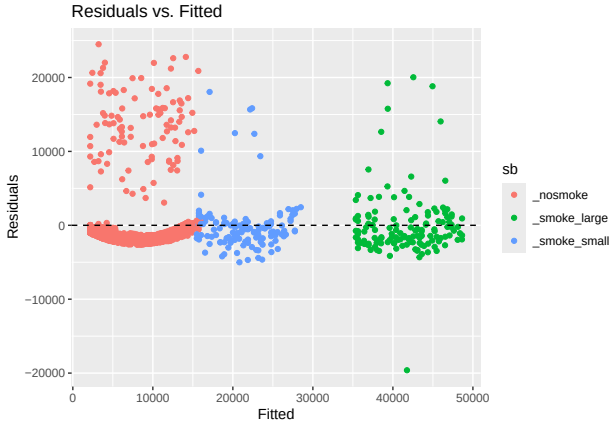
We can see from the residuals vs. fitted plot that even though the model summary shows a great fit, there is a split in the residuals, which means that there must be a split in the two groups we have (smoker vs. non-smoker). Also, from the scatter plots of charges vs. BMI and charges vs. age, it is clear that the data is divided into 3 different groups, and it is possible that the charges is different between three groups.

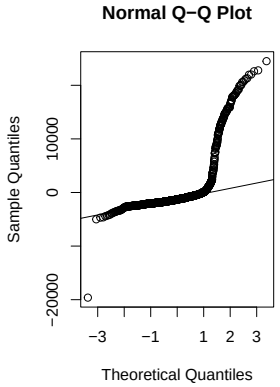


Based on the bmi vs. charges plot, we can see that the charges is splitted around BMI equals 30. From the previous model summary, we know that the group of smokers and non-smokers makes a significant difference in medical cost charges, then we can try to split the data into different groups based on both smoker and bmi. After testing different groupings and threshold using cross validation, we choose to split the data into three different groups: 1. smoker whose BMI is larger than 30; 2. smoker whose BMI is smaller or equal to 30; 3. non-smokers. Then we fit another linear model.

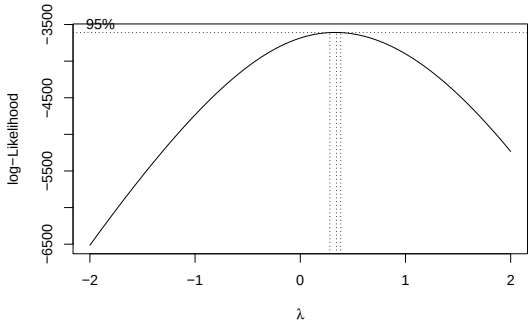
```
##
## Call:
## lm(formula = charges ~ age + children + sb, data = data)
```

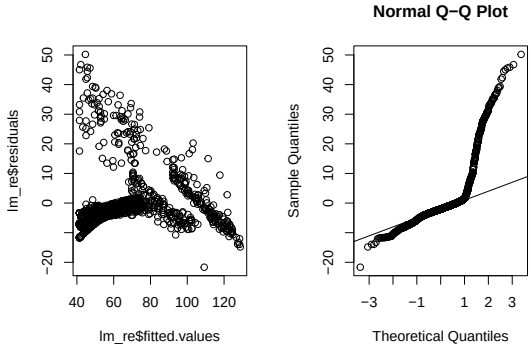
```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19615.6  -1880.4  -1312.8   -542.9  24499.3
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2616.221    385.769  -6.782 1.78e-11 ***
## age              266.329      8.811  30.227  < 2e-16 ***
## children         514.609    102.641   5.014 6.06e-07 ***
## sb_smoke_large 33190.031    400.124  82.949  < 2e-16 ***
## sb_smoke_small 13321.939    421.639  31.596  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4520 on 1333 degrees of freedom
## Multiple R-squared:  0.8611, Adjusted R-squared:  0.8607
## F-statistic:  2066 on 4 and 1333 DF,  p-value: < 2.2e-16
```





It is clear that the model is underestimated in all the groups. Since the assumptions of residuals are not well held, we do a Box-Cox test to transform the response variable and re-fit a linear model.





From the Box-Cox test, the optimal λ is 0.3434, and we transform the response value in the way $Y_{transform} = \frac{Y^{\lambda}-1}{\lambda}$. From the TA plot and the qqplot of the re-fitted model, the transformation does not do a really good job, and the response variable is still right-skewed. So the model of best fit we chose is:

$$Y = \beta_0 + \beta_1 age + \beta_2 I(smoker) + \beta_3 I(bmilarge) + \beta_4 children + \beta_5 I(smoker) * I(bmilarge)$$

References

1. Miri Choi. Medical Cost Personal Datasets. Kaggle. Available at: https://www.kaggle.com/datasets/mirichoi0218/insurance. Accessed March 5, 2025.

2. Green CA, Pope CR. Sex Differences in the Use of Health Care Services. *New England Journal of Medicine.* 1998;338(23):1678-1683. doi:%5B10.1056/NEJM199806043382307](https://www.nejm.org/doi/full/10.1056/NEJM199806043382307).

Appendix

Code

```r
# Select the BMI threshold which results in smallest MSE
set.seed(123)
train_index <- sample(1:length(data$charges), 0.8*length(data$charges))
train <- data[train_index,]
test <- data[-train_index,]
bmi_range <- c(30, 30.5, 31, 31.5, 32, 32.5, 33, 33.5, 34, 34.5, 35)
names(bmi_range) <- c(30, 30.5, 31, 31.5, 32, 32.5, 33, 33.5, 34, 34.5, 35)
test_mse <- numeric()
for (i in 1:length(bmi_range)){
  train$bmilarge <- as.factor(ifelse(train$bmi > bmi_range[i], 1, 0))
  test$bmilarge <- as.factor(ifelse(test$bmi > bmi_range[i], 1, 0))
  lm.1 <- lm(charges ~ age+I(smoker)+I(smoker):I(bmilarge)+children, train)
  pred <- predict(lm.1, newdata=test)
  test_mse[i] <- sum((test$charges - pred)^2)/nrow(test)
}
names(bmi_range)[which.min(test_mse)]
# "30"
```

Attribution

Part 0: All, Part 1: Yuning Hu, Part 2: Luna Lu, Part 3: Yuhan Zhang