Homework 5

Chance Kang / A13605546 / csk025@ucsd.edu
Jayden Kim / A16271107 / s0k003@ucsd.edu
Sia Sheth / A16357789 / snsheth@ucsd.edu
Math 189
Spring 2023

Conceptual Question

1.

Linear regression has some assumptions such as:
1. an approximately linear relationship between two sets of variables x and y (predictor and outcome)
2. constant variance
3. the errors must be independent and normally distributed

To ensure our statistical inferences are valid and reliable we should be checking all 3 of these assumptions. If there is any violation to these assumptions, it can lead to invalid statistical inferences. For example, prediction intervals, confidence intervals, hypothesis tests. These may result in incorrect conclusions about the significance of predictor variables.

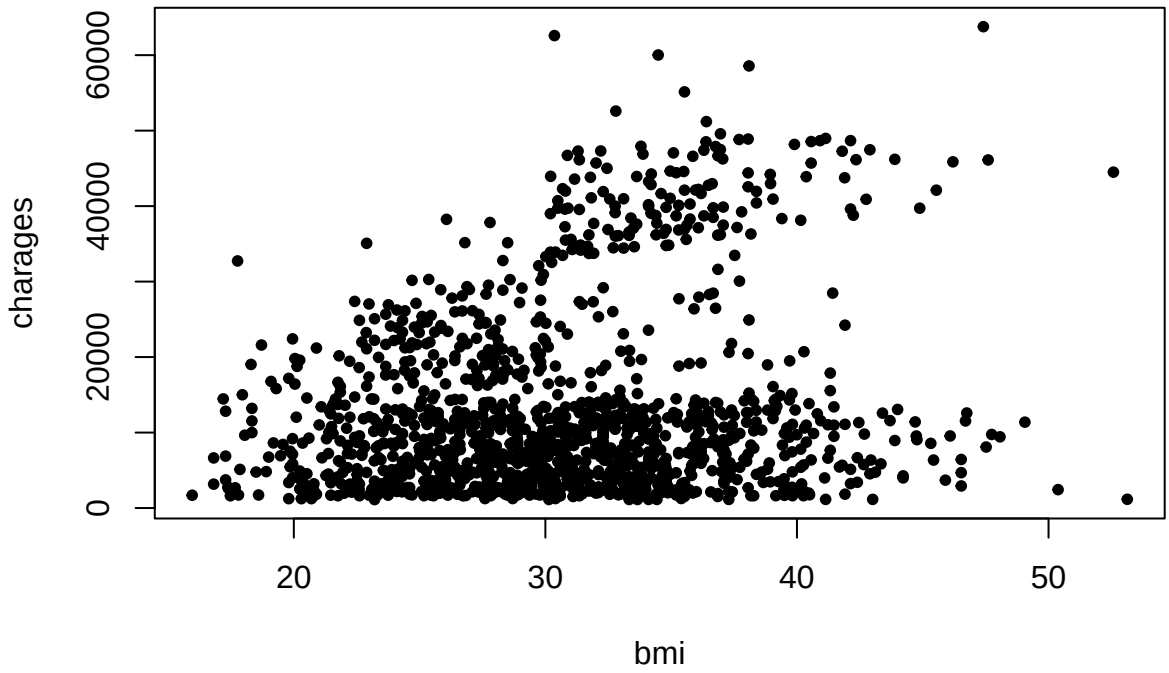
By checking the modeling assumptions, it helps us determine the suitability of the linear regression model. It also enables us to make any necessary modifications to improve the overall performance of the model. Additionally, it can help us identify any potential outliers or influential observations that could change the model's fit and give us insight into the data generation process too.

Application Question

```
hw5<-read.csv("./homework4_insurance.csv")
```

2.

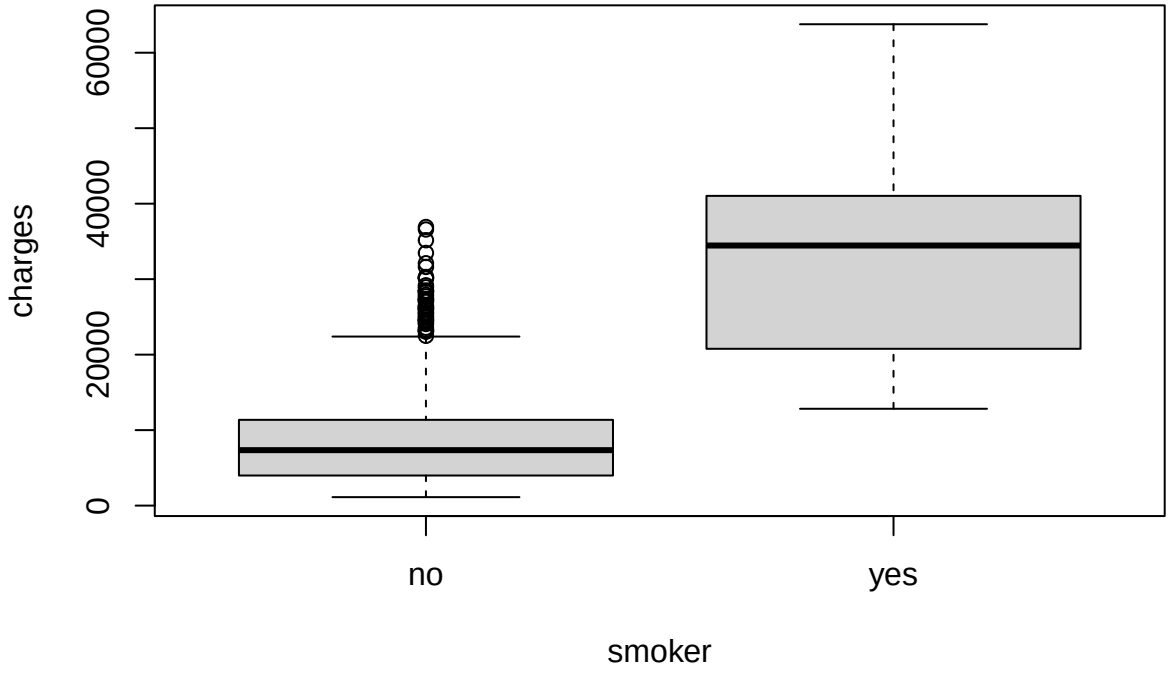
```
plot(hw5$bmi, hw5$charges, xlab = "bmi", ylab = "charages", pch=20)
```



There is a concentration of points in the area bmi 20~40 and charages 0~15,000. Aside from the concentration, higher charges are associated with higher bmis: charges 20,000~30,000 with bmi 15~30, and charges 30,000~50,000 with bmi 30~40.

3.

```
boxplot(hw5$charges~hw5$smoker, xlab = "smoker", ylab = "charges")
```



Though we see some considerable outliers for high charges associated with nonsmokers, it does appear that smokers have higher medical bills in general.

4.

```
lm_c_bs <- lm(charges ~ bmi + smoker, data = hw5)
```

a.

```
summary(lm_c_bs)
```

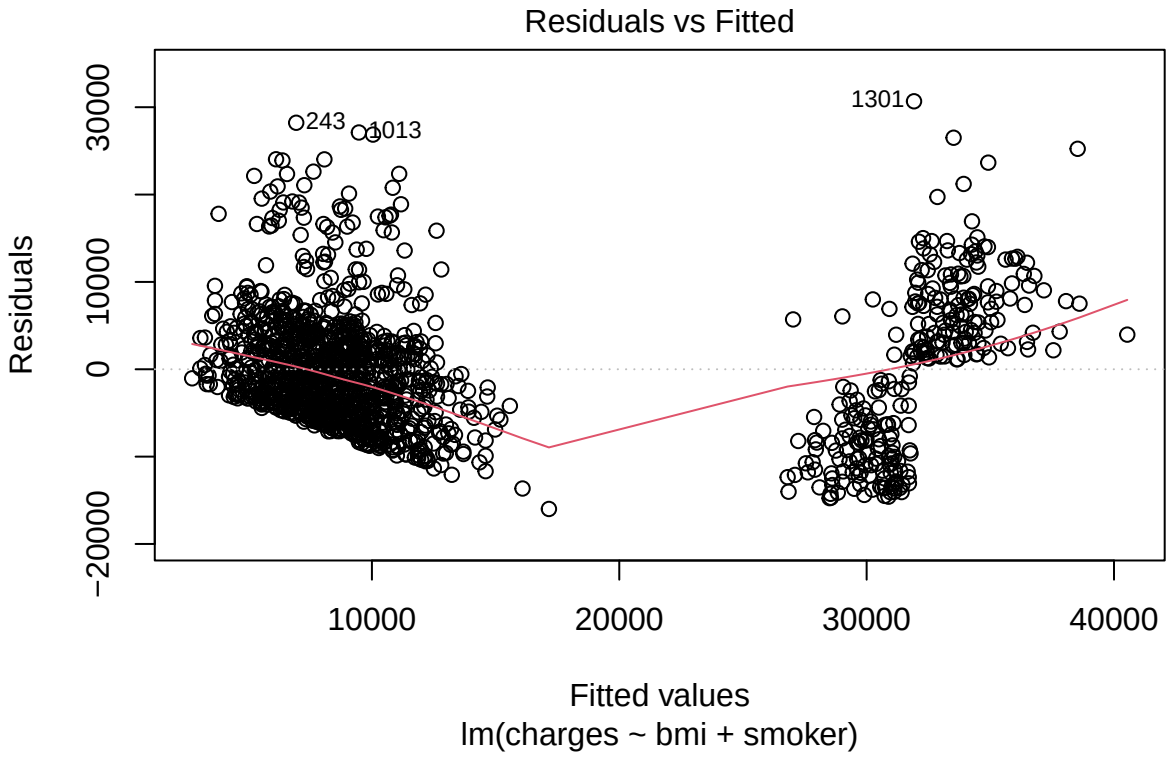
```
##
## Call:
## lm(formula = charges ~ bmi + smoker, data = hw5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15992.7  -4600.2   -802.4   3636.2  30677.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3459.10     998.28  -3.465 0.000547 ***
## bmi           388.02      31.79  12.207  < 2e-16 ***
## smokeryes   23593.98     480.18  49.136  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7088 on 1335 degrees of freedom
## Multiple R-squared:  0.6579, Adjusted R-squared:  0.6574
## F-statistic:  1284 on 2 and 1335 DF,  p-value: < 2.2e-16
```

b. The coefficients regarding smoker shows that those who smoke is more likely to have higher charges with more weight than bmi for $\beta_2 = 23593.98 > \beta_1$ and its low p-value suggesting statistical significance. In our context, a smoker is expected to pay \$23593.98 as medical bills if other variables were to be ignored.

c. The coefficients regarding bmi shows that increase in bmi is likely to show increase in charges for the coefficient $\beta_1 = 388.02$ is positive with its low p-value suggesting statistical significance. In our context, a person is expected to pay \$388.02 per bmi unit as medical bills if other variables were to be ignored.

d.

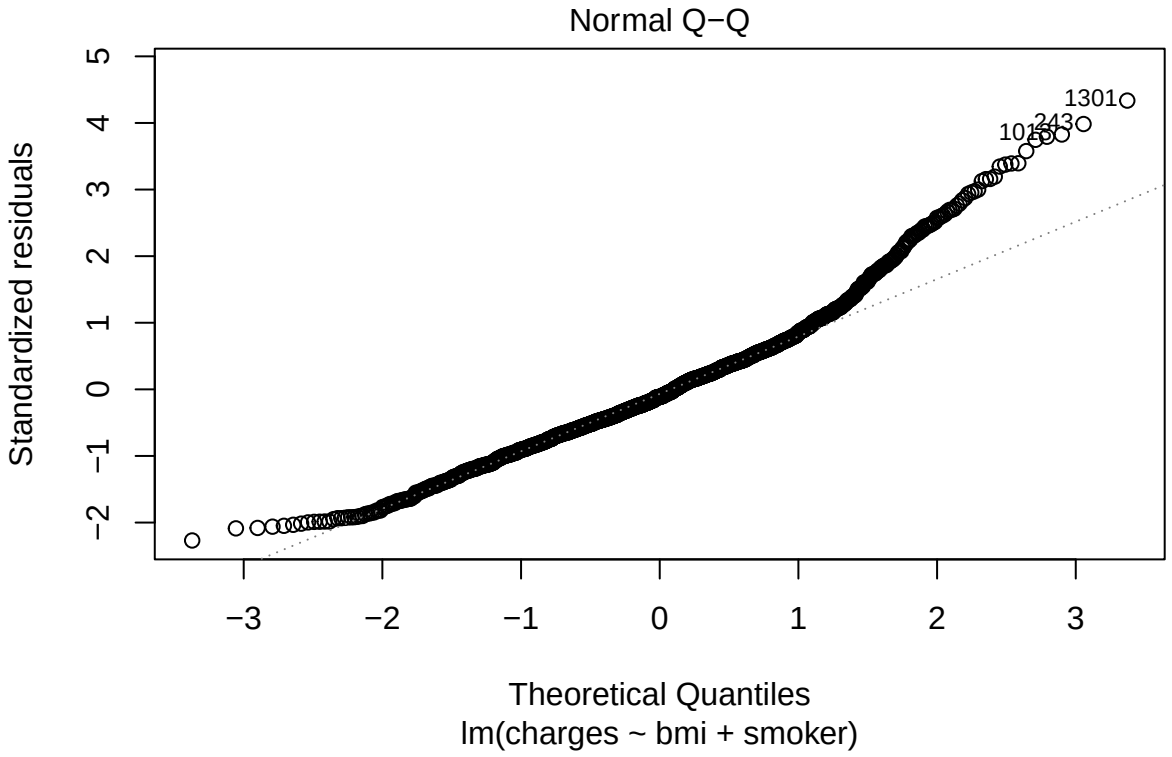
```
plot(lm_c_bs, which = 1)
```



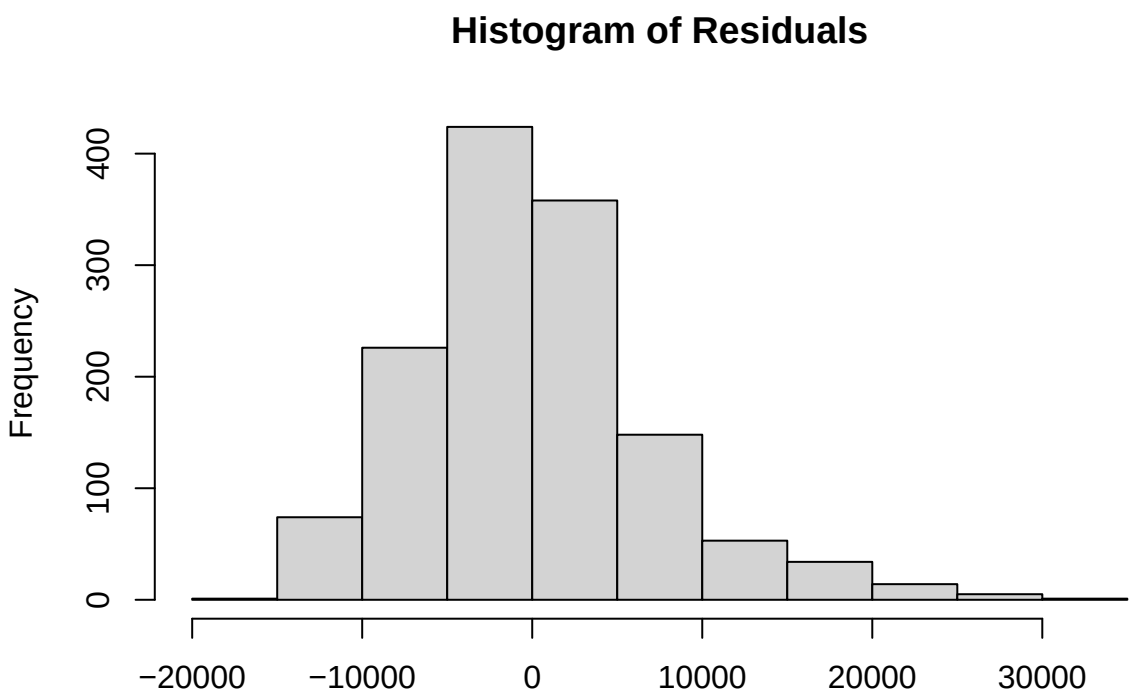
The plot shows some hints for violation of assumptions. The lower residuals relate to lower variance and the higher residuals relate to higher variance, hinting violation of heteroscedasticity. Furthermore, the red fitted line shows a dip hence also hinting nonlinearity.

e.

```
plot(lm_c_bs, which = 2) #Q-Q plot
```



```
hist(lm_c_bs$residuals, xlab = "", main = "Histogram of Residuals") #histogram of residuals
```



The Q-Q plot shows some points being off of the fitted line, especially towards standard residuals > 1, as well as the histogram of residual showing somewhat of a normal distribution but not adequately. Hence, the normality assumption also does not hold.

5.

```
lm_log <- lm(log(charges) ~ bmi + smoker, data = hw5)
```

a.

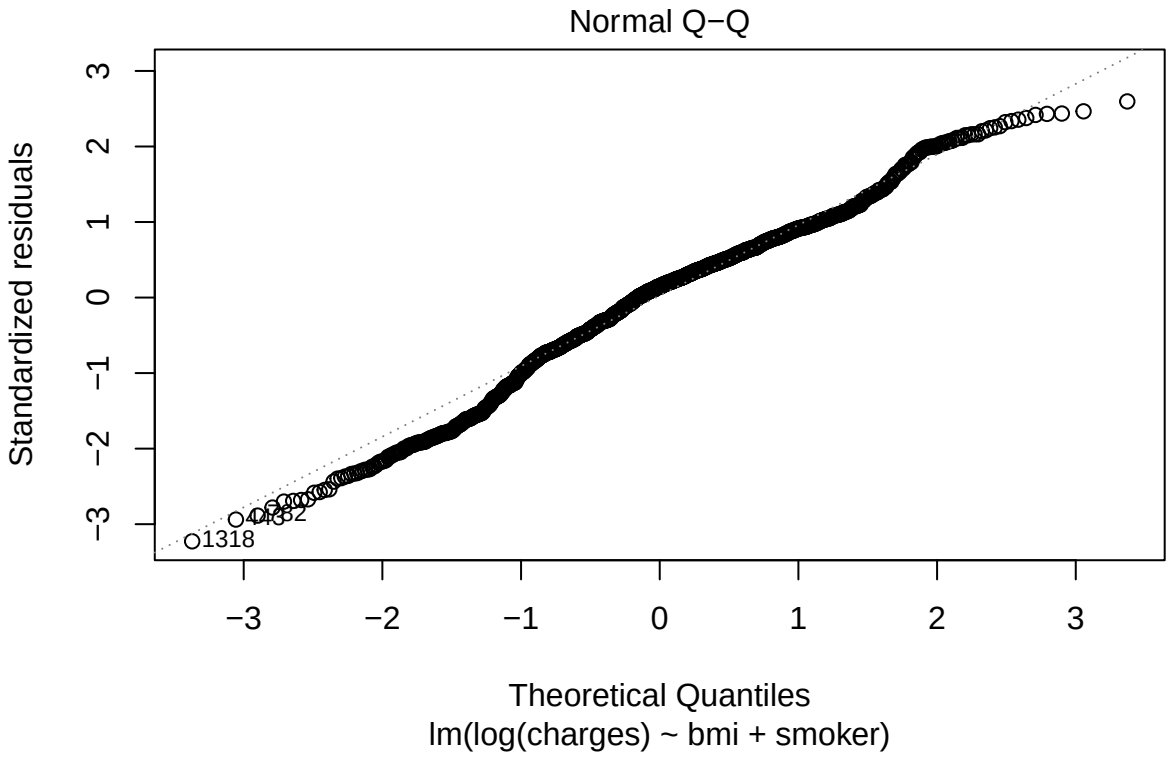
```
summary(lm_log)
```

```
##
## Call:
## lm(formula = log(charges) ~ bmi + smoker, data = hw5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.17030 -0.40754  0.09871  0.44368  1.75504
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.186576   0.095254  85.945  < 2e-16 ***
## bmi         0.019629   0.003033   6.472 1.36e-10 ***
## smokeryes   1.514765   0.045818  33.061  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6763 on 1335 degrees of freedom
## Multiple R-squared:  0.4598, Adjusted R-squared:  0.459
## F-statistic: 568.3 on 2 and 1335 DF,  p-value: < 2.2e-16
```

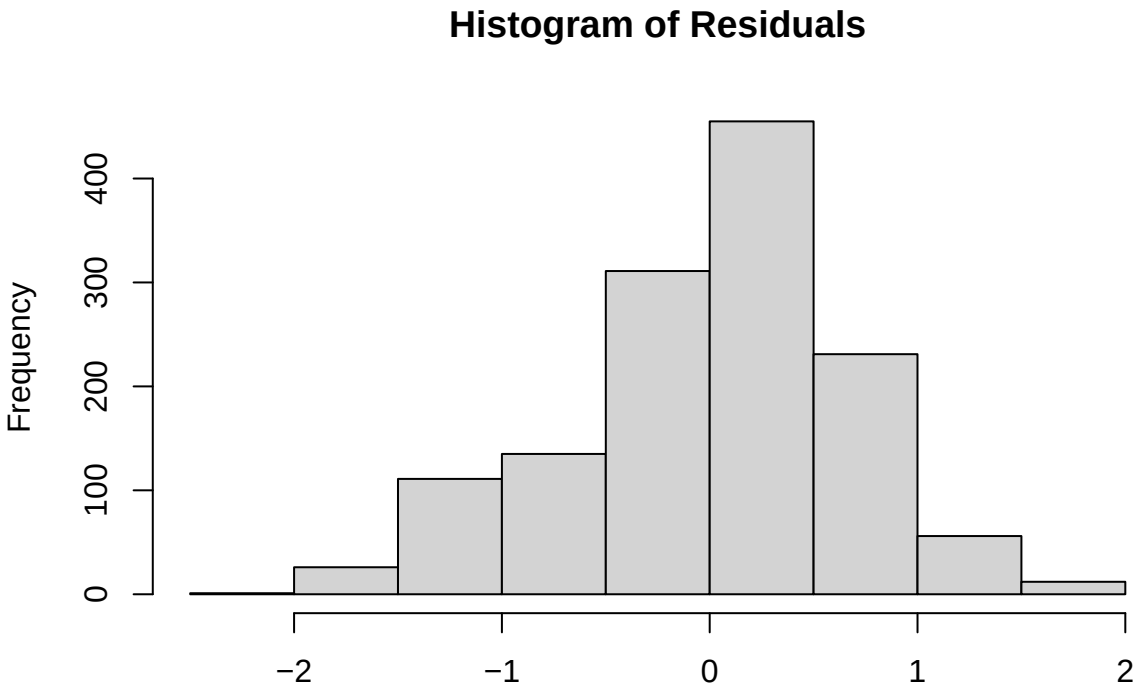
b. The coefficients regarding smoker shows that those who smoke is more likely to have higher charges with $\beta_2 = 1.514765$ and its low p-value suggesting statistical significance. Suppose y is the medical charges for a person and s indicates smoker for $s = 1$. In our context, $log(y) = s * 1.514765$ hence if a person is a smoker, then the log of medical bill is expected to increase by \$1.514765 if other variables were to be ignored.

c.

```
plot(lm_log, which = 2) #Q-Q plot
```



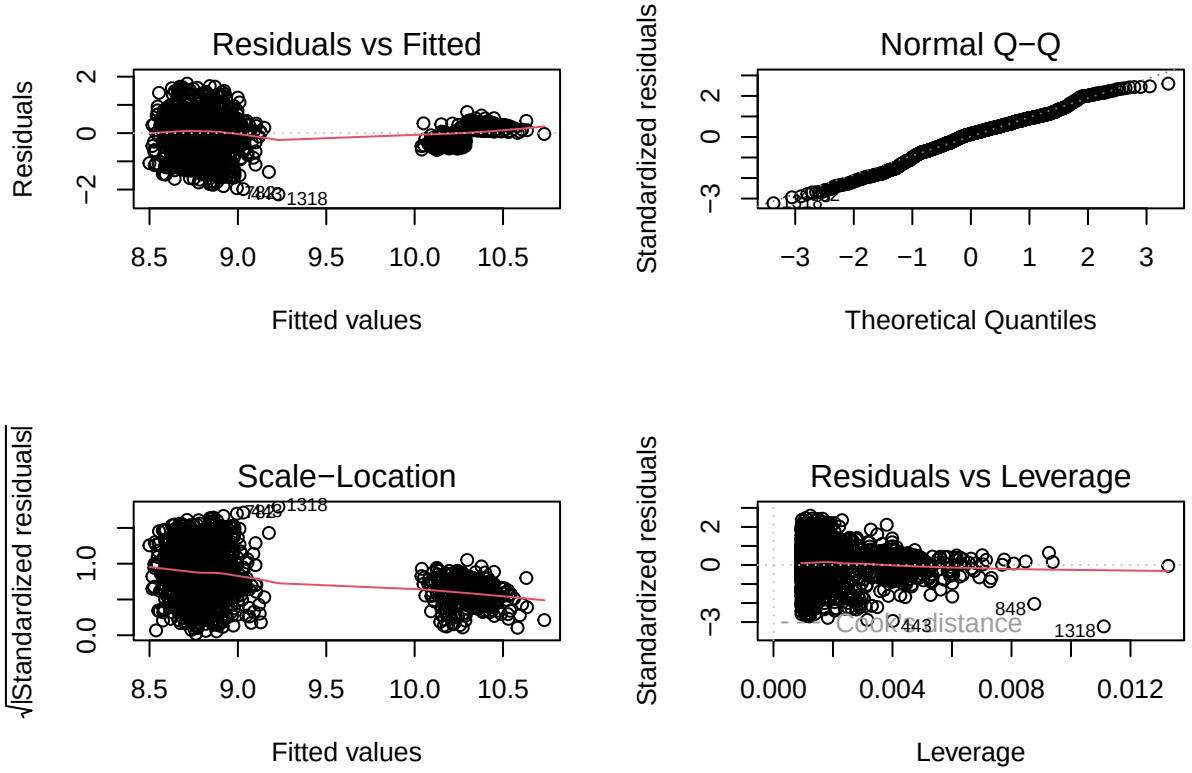

```
hist(lm_log$residuals, xlab = "", main = "Histogram of Residuals") #histogram of residuals
```



The points on QQ plot fits the fitted line relatively well with histogram of residuals showing a rough normal distribution. Hence, the normality assumption is not violated.

d.

```
par(mfrow = c(2, 2))
plot(lm_log)
```



The difference between model (4) and model (5) is that model (5) violates less assumptions. Unlike model

(4), the residual vs fitted plot of model (5) shows relatively even distribution of residual above and below the fitted line, as well as the fitted line being relatively straighter along the 0 residual than model (4). The point in the QQ plot nearing the presented fitted line as well as other plots showing linearity.

6.

a. Suppose our model fits $Y = \beta_0 + \beta_1 * x_1$. Let β_1 be the coefficient responding to bmi.

$$\mathcal{H}_0 : \beta_1 = 0$$

$$\mathcal{H}_1 : \beta_1 \neq 0$$

Our null hypothesis ($\mathcal{H}_0$) states that there is no relationship between charges and bmi.

b.

```
summary(lm_log)
```

```
##
## Call:
## lm(formula = log(charges) ~ bmi + smoker, data = hw5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.17030 -0.40754  0.09871  0.44368  1.75504
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.186576   0.095254  85.945  < 2e-16 ***
## bmi         0.019629   0.003033   6.472 1.36e-10 ***
## smokeryes   1.514765   0.045818  33.061  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6763 on 1335 degrees of freedom
## Multiple R-squared:  0.4598, Adjusted R-squared:  0.459
## F-statistic: 568.3 on 2 and 1335 DF,  p-value: < 2.2e-16
```

We see that the p-value for bmi is $1.36e^{-10}$ and we reject null hypothesis for p-value of bmi $1.36e^{-10} < 0.00001$ for $\alpha = 0.05$. Hence bmi is statistically significant in our model and has a significant effect on predicted medical charges. c.

```
confint(lm_log) #defaults at 95%
```

```
##                   2.5 %     97.5 %
## (Intercept) 7.99971307 8.37343981
## bmi         0.01367858 0.02557889
## smokeryes   1.42488225 1.60464784
```

Contribution

Our group homework process goes as the following:
1. Each member attempts to complete the homework
2. Compare and discuss the answers
3. Complete a finalized version to submit

All members have contributed about the equal amount to complete this homework.