Factors Effecting Medical Costs in the United States

Anna Iuferova (30166345) David Mejia (30020246), Jon Peters (30158434), Joe Yong (30165975)

10/13/2021

Motivation

All over the world, health expenditures having been rising at a significant rate over the past decade. In the United States specifically, health care expenditures rose from $1.8 trillion in the year 2000 to over $3.8 trillion in 2019, adjusting for inflation (Kamal et al 2020). With the global covid-19 pandemic, many health care procedures have, and continue to be, postponed in order to give priority to healthcare to those affected by this virus. PWC predicts that that within the year 2022, health care expenditures will increase 6.5% following the ease of covid restrictions that will allow for these procedures to being taking place (PricewaterhouseCoopers 2021). With these continual increase in costs, it has become more of an issue for citizens in affected countries to take care of themselves and their families in in order to avoid the high cost of treatment on top of the increasing insurance premiums they already pay. The goal of this analysis is to understand what factors contribute most to medical costs in an effort to help families save on medical costs in the future. The following questions will guide this analysis centered around this idea:

1) Different regions could be effected by varying medical costs mainly due to the health challenges that regions faces, as well a cultural attitudes towards food, body images, and health in general. Is the individuals specific BMI rate (healthy, overweight, obese) dependent across different regions?
2) Cost of medical tests can fluctuate through different regions of the country for a variety of factors such as the availability and cost of transportation to these regions. Does the region affect the mean medical costs? Specifically, we will test the difference in mean variance between northwest and southeast regions and see how costs differ.
3) Body Mass Index (BMI) has been a key indicator in overall health as an indicator of mass to height of an individual. Although the usability of this index has slowly declined over the years, does a higher BMI, signaling the risk of obesity, lead to higher medical costs for these individuals? Does BMI have a linear relationship with medical costs? If so, how much of a factor is BMI in explaining medical costs?
4) It has been proven time and again that smoking leads to various health risks such as bronchitis and cancer, but overall does the cost of medical treatment increase for those smoking versus non-smoking? We will construct 95% confidence intervals and compare and contrast the true mean medical costs for smokers versus non-smokers. Understanding with 95% certainty how medical costs differ between smoking groups will better help infer greater risks of smoking, which in turn leads to better medical care.

The main parameters we are interested in studying specifically are the mean medical charges for those in the southeast region and the northwest ($\mu_{southeast}$ and $\mu_{southeast}$) as well as the mean medical charges for those that smoke compared to those that do not smoke (μ_{smoker} and $\mu_{non-smoker}$). Likewise, the change in medical cost for every unit increase in BMI is going to be a valuable parameter to evaluate as it will help infer the necessity for a healthy lifestyle in order to avoid heavy medical costs.

By studying these parameters, the overall purpose of this analysis is to identify factors that ultimately effect overall medical costs. By understanding what drives medical costs, families can better equip themselves to live a healthier lifestyle and understand what may be causing their medical costs to fluctuate in an effort to change lifestyle and save money of medical expenses as prices continue to rise.

Data Collection

Initially we had identified a variety of topics that we would be interested in doing our project on, ranging from Student performance, to medical cost to transit delays in Calgary. We evaluated multiple data sets to see if they would be suitable for this project. Ultimately we selected on the Medical Cost Personal Data set as it was the best data set on which to do inference testing. Due to the difficulty of finding and medical data due to confidentiality, this data set is simulated to match demographic statistics from the US Census Bureau, and is not refelctive of any real-life individual. Data has been provided by Miri Choi via Kaggle.com and is available for use under the Database Contents License (DbCL).

This data set contains date on insurance and medical information from people in the United States. The data set contains medical and non-medical information on 1338 individuals each with seven separate categories of data, including age, sex, BMI, number of children, smoker, region and charges. Further description of each category below:

- Age: Age of patient
- Sex: Sex of patient (Male or Female)
- BMI: Body mass index, objective index of body weight (kg / m ˆ 2) using the ratio of height to weight.
- Children: Number of dependents for each patient
- Smoker: Does the individual smoke (yes or no)
- Region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- Charges: Individual medical costs billed by health insurance

The file came in a CSV format. Due to the clean format of the data set, there was no necessary cleaning, nor was there any difficulty in procuring and utilizing this data set for this analysis.

Analysis

BMI and Region and a Dempgraphic Highlight to Medical Costs

Crucial to our understanding of the variable factor that could explain medical charges is understanding the demographic information within this sample set. Two large factors that could effect medical costs are that of the individuals body mass index (BMI) and the region of the country they are located. BMI is a indicator of the individuals mass compared to their height. After a individuals BMI is recorded, they are repented within the following groups:

Group	BMI
Underweight	<18.5
Normal weight	<24.9
Overweight	25–29.9
Obesity	30>

Due to a lack of an a sufficient sample, those considered underweight will be grouped together with those that are underweight within this analysis. Likewise, region of the country could adversely affect medical costs as transportation costs, population size, and need for care drive costs within the given market. To this end, these two factors could largely weigh the overall need and cost of care and thus the following test will test for dependency between these two factors. The following hypothesis will be tested:

$$H_0 : BMI \text{ is independent of region.}$$

$$H_a : BMI \text{ is not independent of region.}$$

The conclusion to this test will let us know if there are any regions that are adversely affected by health struggles (higher BMI), which in turn could be lead to higher medical costs, as explored further in this analysis.

The following test for independence utilizes a chi-squared distribution to identify the relationship between the two factors. This relationship will be tested as a significance level of $\alpha = .05$.

```
testFit <- med %>% mutate(region = as.factor(region))
testFit$bmiRange <- ifelse(testFit$bmi < 25, 'Under/Normal', '')
testFit$bmiRange <- ifelse(testFit$bmi >= 25 & testFit$bmi < 30, 'Over', testFit$bmiRange)
testFit$bmiRange <- ifelse(testFit$bmi >= 30, 'Obese', testFit$bmiRange)

(contTable <- table(testFit$region, testFit$bmiRange))
```

```
##
##              Obese Over Under/Normal
##   northeast    143   98           83
##   northwest    148  107           70
##   southeast    243   80           41
##   southwest    173  101           51
```

Joining our sample set into the following table, we are able to visualize the current layout of BMI indicators as well as region. we find everything to be relatively similar to one another, with the exception of the southeast, which see's a significant larger number of obese individuals than those in other regions. This already is enough to signal that there is a dependent relationship between our two factors, but the test will performed in order to definitively conclude this hypothesis.

```
(chiTest <- chisq.test(contTable))
```

```
##
##  Pearson's Chi-squared test
##
## data:  contTable
## X-squared = 52.268, df = 6, p-value = 1.647e-09
```

With the reported p-value (1.6471139×10^{-9}) being less than the significance level of $\alpha = .05$, we will reject the null hypothesis. It is thus concluded that BMI is dependent on region. As a result, the following contingency table is constructed in order to see which regions and BMI rating contribute most to the resulting chi-squared statistic. We find that the largest contributors are those within the southeast region under all health ratings, as identified above. Not only was obesity overly dependent on the southeast region, but we likewise find that those who where overweight and those normal or underweight also had a large share of those dependent on their region. Also with a large contribution of are those with a normal or underweight score in the northeast region as well.

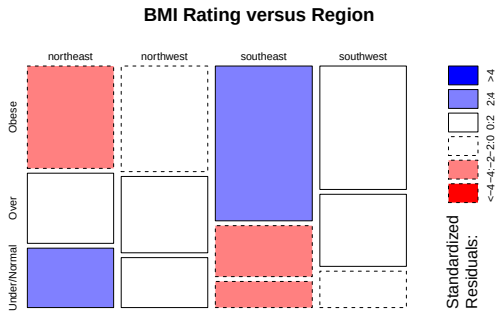
```
expectedMat <- as.array(margin.table(contTable,1)) %*%
  t(as.array(margin.table(contTable,2))) /
  margin.table(contTable)

(expectedMat - as.array(contTable ))^2 / expectedMat
```

```
##
##                     Obese         Over Under/Normal
##   northeast  4.645635725  0.219460730  9.445797193
##   northwest  3.279109563  1.869840508  1.848924917
##   southeast 13.344524510  5.956770957  9.872373839
##   southwest  0.009389184  0.559166933  1.217063064
```

Knowing these regions had a large contribution to the overall chi-squared test statistic, the following residuals were also reported (which is simply the square root of the contribution matrix), but the key is identifying how regions differ from their expected outcome, whether they fell above our below their expected values. The following mosaic plot is used as a visualization of the pearson residuals from our contingency table.

```
mosaicplot(contTable, shade = TRUE, main = "BMI Rating versus Region")
```



As noted above, the southeast across the board saw large differences what is expected, but overweight and normal/underweight individuals actually fell below the expected amount within that region. Likewise, the northeast, with a large contribution to the normal/underweight group found significant individuals than what was expected in that group, while finding significantly less in the obese group, signaling that they have healthier BMI's than those within other regions.

Altogether, we find that BMI is dependent of region. The consequences of a significantly higher BMI in the southeast leads us to conclude that the population is significantly less health then their counterparts. With this crucial information in mind, we will test how medical costs differ between varying regions, using the southeast as a key indicator, which in turn may give us some insight into the overall health of a population and the medical costs they incur.

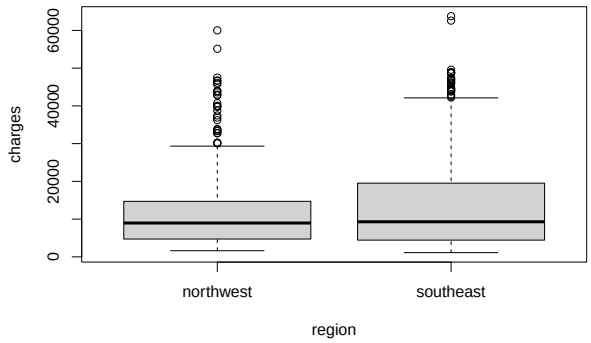
Difference in Region and Medical Costs

In order to learn if the region influences the mean of medical cost, we need to perform the t-test based on two different data samples related to Northwests and Southeasts. In the code below, we load the data, and choose only that data which belongs to two these particular regions.

```
options(scipen=3)
Data_SE <- med[med$region == "southeast",]
mean_SE = mean(Data_SE$charges)
var_SE = var(Data_SE$charges)
sd_SE = sd(Data_SE$charges)


Data_NW <- med[med$region == "northwest",]
mean_NW = mean(Data_NW$charges)
var_NW = var(Data_NW$charges)
sd_NW = sd(Data_NW$charges)

boxplot(charges~region, data = med[((med$region == "northwest") | (med$region == "southeast")),])
```



The above box plot us used to visualize the distribution of these two region. It is noted that both groups have relatively constant medians, but the variance seems to be a lot larger within the southeast. To get an idea of how the variance changes between the two group, the following descriptive statistics are noted below.

Region	Mean	Variance	Std. Dev.
Southeast	14735.4114376	1.951916×10^8	13971.098589
Northwest	12417.575374	1.2259532×10^8	11072.2769276

Due to the increase in variance for the southeast, a test for equal variance will be conducted, as this could affect the type of test that is conducted for difference in mean of medical charges between regions. The variance test uses the following hypothesis. Test for variance are dependent on a normal distribution for both groups. Normality will be assumed in this case as althought there are singificant outliers according to our boxplot, the sample size is large enough to pull the distribution to somwaht normal.

$$H_0 : \sigma^2_{southeast}/\sigma^2_{northwest} = 1$$
$$H_a : \sigma^2_{southeast}/\sigma^2_{northwest} \neq 1$$

```
var.test(Data_SE$charges, Data_NW$charges, ratio=1)
```

```
##
##  F test to compare two variances
##
## data:  Data_SE$charges and Data_NW$charges
## F = 1.5922, num df = 363, denom df = 324, p-value = 0.00002042
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.286915 1.967263
## sample estimates:
## ratio of variances
##           1.592162
```

with a resulting p-value close to zero, at a significance level of $\alpha = .05$, it is concluded that there is a significant difference in variance between these two regions. Thus, in performing T-test calculation, we will use the parameter var.equal = FALSE to utilize a Welch's two-sample unequal variance t-test, in order to account for these differences.

In order to learn if the mean in medical cost between two regions are equal the below hypothesis test is used. With changes in variance and a relatively equal mean, the following test will test if mean medical charges are

in fact significantly different, or if the change in variance keeps the relative mean somewhat constant between regions.

$$H_0 : \mu_{southeast} \leq \mu_{northwest}$$
$$H_a : \mu_{southeast} > \mu_{northwest}$$

The significance level for this test is also set to $\alpha = 0.05$

```
t.test(Data_NW$charges,Data_SE$charges, var.equal = FALSE, alternative = "two.sided",conf.level = 0.95)
```

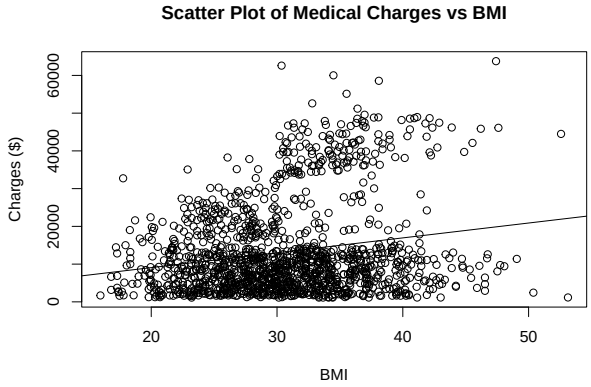
```
##
##  Welch Two Sample t-test
##
## data:  Data_NW$charges and Data_SE$charges
## t = -2.4252, df = 677.64, p-value = 0.01556
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4194.4223  -441.2498
## sample estimates:
## mean of x mean of y
##  12417.58  14735.41
```

So, with a P-value = 0.01556, which is less than $\alpha = 0.05$, it can concluded that the evidence in the data is statistically significant at level $\alpha = 0.05$, and we reject H_0. After completing test hypothesis, we can conclude that there is a difference in mean medical costs between northwest and southeast regions at level $\alpha = 0.05$. Although there are no indicators as to why these regions experience differences in average medical costs, from the previous analysis, it can be inferred that overall health standards in this region could be at least one factor driving these charges up. With this in mind, BMI will now be analyzed to identify its relationship to medical costs.

BMI's Linear Reltationship to Medical Costs

From the previous two analyses, it was found that an individuals BMI is dependent on the region of the the country that they live. However, it was also found that region does not adversely affect medical costs, and that there is no significant different between two sample regions that showed an "average" population and a relatively overweight population. To understand if BMI likewise has any effect of medical costs, and to identify any sort of relationship between these two factors, a simple linear regression model for BMI vs Medical Costs will be constructed. This model will be able determine if there is a linear relationship between BMI and medical costs. Using the R^2 value of the model we will be able to gauge how much of a factor BMI is in explaining variability in medical costs as well. First, a scatter plot between BMI and medical charges will be constructed to visualze the current relationship.

```
options(scipen = 0)
m_bmi_cost <- lm(med$charges~med$bmi)
plot(med$charges~med$bmi, main = "Scatter Plot of Medical Charges vs BMI",
     xlab = "BMI",
     ylab = "Charges ($)")
abline(m_bmi_cost)
```



The above scatter plot of Medical Charges vs BMI doesn't show a strong linear relationship. However, as there exists a linear relatinship, the following model will be used:

$$Charges = \alpha + \beta * BMI + e$$
$$e \sim N(0, \sigma^2)$$

```
summary(m_bmi_cost)
```

```
##
## Call:
## lm(formula = med$charges ~ med$bmi)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20956  -8118  -3757   4722  49442
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1192.94    1664.80   0.717    0.474
## med$bmi       393.87      53.25   7.397 2.46e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11870 on 1336 degrees of freedom
## Multiple R-squared:  0.03934,    Adjusted R-squared:  0.03862
## F-statistic: 54.71 on 1 and 1336 DF,  p-value: 2.459e-13
```

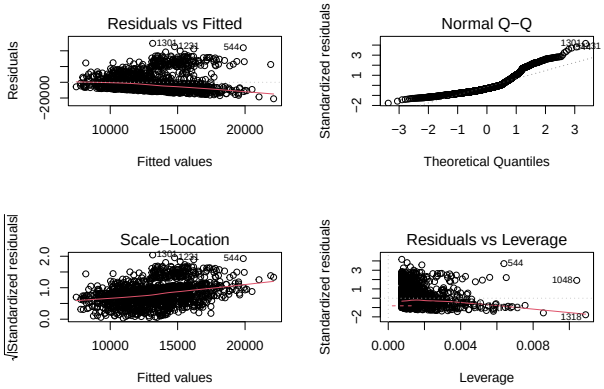
Our simple linear regression model will be a single line of best fit that shows the linear relationship between BMI and Medical costs. This line is found using ordinary least squares method and determining a line that results in the smallest sum of the square differences between Medical costs in the data set. fitting the data set to this model, the intercept is reported as of 1192.9 and has slope of 393.9, resulting in the model below.

$$\widehat{Charges} = 1192.9 + 393.9 * BMI$$

The intercept is the expected cost of medical treatment is an individual has a BMI of 0. In this example this has no meaning as it is not possible for someone to have a zero BMI. Likewise, the $\widehat{\beta_{BMI}}$ coefficient is reported as \$393.9, meaning that for every increase in 1 in BMI score, the increase in an individuals medical costs will be \$393.9, on average. The resulting t-test tests if there is a chance that the resulting coefficient for BMI above is equal to zero, or, in other words, there is no relationship between BMI and medical costs. With

a t-value of 7.397 and p-value of 2.46e-13. At a significance level of $\alpha = .05$, we conclude that there is indeed a relationship between BMI and medical charges. The R^2 value of 0.03934 shows that there is a very weak relationship between BMI and Medical Charges.

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



In order to infer any relationship between our two variables, however, the assumptions of linearity, independence, normality, and equal variance, all need to be met in order to trust that the data is fit properly to the given model.

- The linearity assumption does look like it has been met as there does appear to be a straight line in both the Residuals vs Fitted graph.
- The Scale-Location plot do not look very evenly distributed and and the line is upward sloping. The problems in this plot indicate that the equal variance assumption is not satisfied.
- The normality assumption doesn't look good either. We expect to see a straight line pattern, in the Normal Q-Q plot. There is quite a bit of problems with the number of curves on the Normal Q-Q plot.
- Independent is met as there are these patients are randomly sampled, and are independent of one another.
- A plot of Residuals vs Leverage does not show many influential cases or any major problems with influential cases.

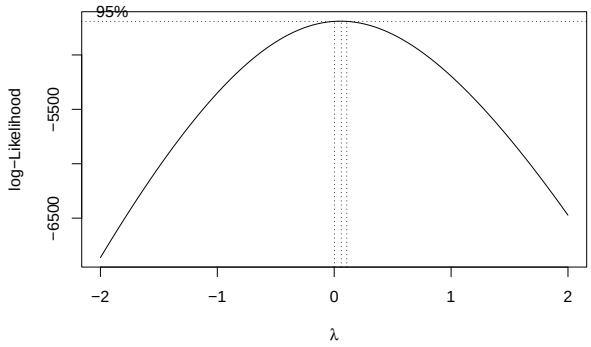
Given the failure to meet multiple assumption, the model as it is currently stated is not appropiate for use of inference or predicitoon. A Box-Cox transformation will be used to see if it will help fix/improve the linear regression model.

```
library(MASS)
```

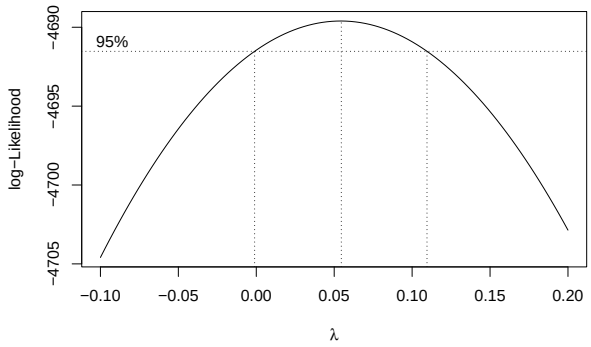
```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
boxcox(m_bmi_cost)
```




```
boxcox(m_bmi_cost, lambda = seq(-.1, 0.2,0.05))
```



From Box Cox the estimated $\lambda = 0.05$ we will assume a $\lambda = 0$ since it is so close to 0. As well the 0 values still falls with in the 95% interval for λ. The resulting $\lambda = 0$ value is a log transformation across the y-value, or medical costs. Thus, the updated model is :

$$\widehat{Charges} = exp^{(8.485+0.020*BMI)}$$

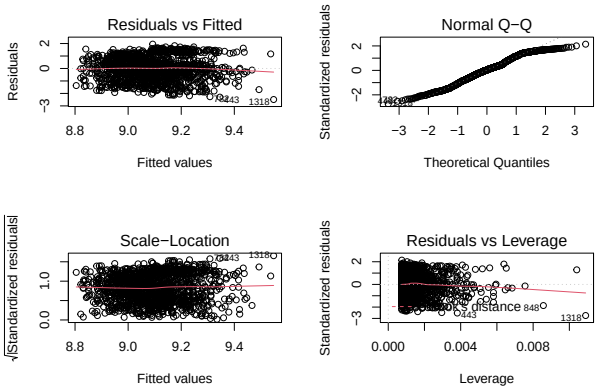
The assumptions are checked again to conclude that the assumptions are appropiatley met.

```
med$logcharges = log(med$charges)
m_bmi_logcosts = lm(med$logcharges~med$bmi)
summary(m_bmi_logcosts)
```

```
##
## Call:
## lm(formula = med$logcharges ~ med$bmi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.48894 -0.63536  0.03136  0.68007  1.95182
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.485243   0.127833  66.378  < 2e-16 ***
## med$bmi     0.020005   0.004089   4.892 1.12e-06 ***
## ---
```

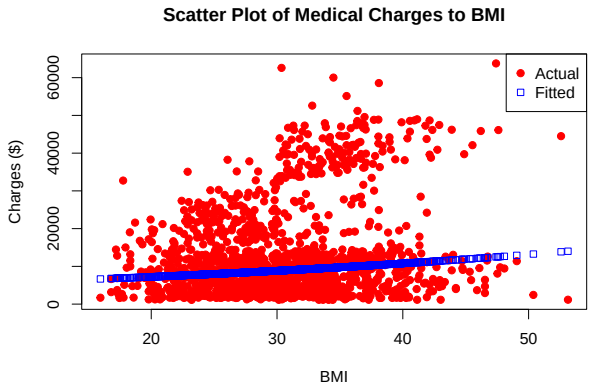
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9117 on 1336 degrees of freedom
## Multiple R-squared:  0.0176, Adjusted R-squared:  0.01687
## F-statistic: 23.94 on 1 and 1336 DF,  p-value: 1.117e-06
```

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



The Residual vs Fitted and Scale-Location plots look much more evenly distributed after the Box-Cox transformation versus the original, thus equal variance is met. Likewise the Normal Q-Q plot also shows much improvement after the Box-Cox transformation as the line is a lot straighter except for the lower and upper tails, thus normality is met. All assumptions are not relatively met, and the current model is appropriate for inference.

```
plot(med$charges~med$bmi, data = med, pch = 19, col = "red",
     main = "Scatter Plot of Medical Charges to BMI",
     xlab = "BMI",
     ylab = "Charges ($)")
points(med$bmi, exp(8.485 +0.020 * med$bmi), pch = 22, col = "blue")
legend("topright", legend = c("Actual", "Fitted"), col=c("red", "blue"),
       pch = c(19,22))
```



The resulting fit from this model is visualized above in an effort to show the relationship between BMI and medical charges as defined by our model.

Even though the model with the Box-Cox transformation does provide a better fit, the R^2 value is calculated as it represents the percentage of variability of y explained by x, which in this case would be the percentage of

variability of medical cost explained by BMI. The low R^2 value of 0.0176 indicates that BMI has a very weak relationship with medical charges, and thus BMI does not do a very good job at explaining medical costs.

As BMI is dependent on region, it is found that neither variable individually contributes much to overall medical charges. These factors are indirect factors to an individuals overall health and it is thus concluded that they do not have much effect on the medical costs an individual will incur. The only factor that could adversely effect health is whether the individual is indicated to be a smoker or not, which is the last factor that is explored in this analysis in relation to medical costs.

Smoking as a Significant Health Factor

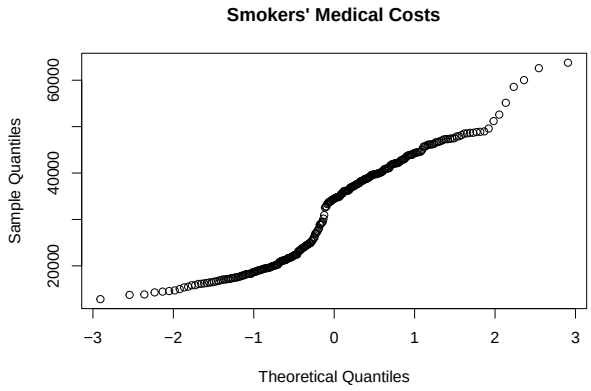
Since smoking correlates with an increased risk for certain illnesses, it is conceivable that cost of medical treatment is higher for smoking individuals versus non-smoking individuals. To verify this, 95% confidence intervals will be constructed to obtain possible ranges of the mean costs for the smoker population and the non-smoker population.

The first step in computing the confidence intervals is separating the 'charges' data into smokers and non-smokers:

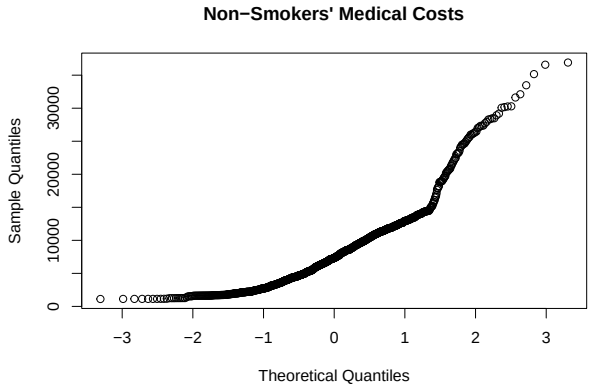
```
smoke = med[med$smoker == "yes",]$charges
nonsmoke = med[med$smoker == "no",]$charges
```

Now normality in the data can be verified by constructing Normal Quantile-Quantile plots:

```
qqnorm(smoke, main = "Smokers' Medical Costs")
```




```
qqnorm(nonsmoke, main = "Non-Smokers' Medical Costs")
```



Considering the large sample sizes for the two above cases, it would expected to see straight line on the

Quantile-Quantile plots to confidently assume that the samples are normal. While normality cannot be assumed, due to the curvature in the plot, a large sample interval estimator can be used to calculate the 95% confidence intervals for the population means μ_{smoke} and $\mu_{nonsmoke}$. The interval estimator in use is the following at level $\alpha = 0.05$:

$$(\bar{X} - Z_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + Z_{\alpha/2}\frac{S}{\sqrt{n}})$$

Point estimates ($\bar{X}$) must be calculated, as well as sample standard deviations (S):

```
options(scipen=3)

xsmoke = mean(smoke)
xnonsm = mean(nonsmoke)
sdsmoke = sd(smoke)
sdnonsm = sd(nonsmoke)
nsmoke = length(smoke)
nnonsm = length(nonsmoke)
Z_025 = qnorm(0.975, 0, 1)
```

Now, the 95% confidence interval for μ_{smoke} can be computed as follows: (xsmoke - Z_025 * sdsmoke / sqrt(nsmoke) , xsmoke + Z_025 * sdsmoke / sqrt(nsmoke)) =

$(32050.2318315 - 1.959964 * 11541.5471756/\sqrt{274}$, $32050.2318315 + 1.959964 * 11541.5471756/\sqrt{274})$

which gives: (30683.6462299 , 33416.8174332)

Next, the 95% confidence interval for $\mu_{nonsmoke}$ can be computed as follows:

(xnonsm - Z_025 * sdnonsm / sqrt(nnonsm) , xnonsm + Z_025 * sdnonsm / sqrt(nnonsm)) =

$(8434.2682979 - 1.959964 * 5993.7818192/\sqrt{1064}$, $8434.2682979 + 1.959964 * 5993.7818192/\sqrt{1064})$

which gives: (8074.12262 , 8794.4139757)

Thus, the 95% confidence intervals for μ_{smoke} and $\mu_{nonsmoke}$ are respectively (30683.65 , 33416.82) and (8074.12 , 8794.41). Since these two confidence intervals do not overlap in any part of their ranges, then at the 95% confidence level it appears that the average medical costs for smoking individuals will be higher than the average medical costs for non-smoking individuals.

A boot strap simulation will replicate samples from these two groups and calculate a bootstrapped confidence interval that will allow for a closer conclusion of the true mean for smokers and non-smokers.

```
library(boot)
```

```
## Warning: package 'boot' was built under R version 4.1.1
```

```
set.seed(2021)

bmean <- function(data,i){
  d = data[i,]
  return(mean(d))
}
```

The 95% percentile bootstrap interval for medical costs of smokers can be calculated as follows:

```
bootsmoke = boot(data = data.frame(smoke), statistic = bmean, R = 2000,
                 sim = "ordinary", stype = "i")
boot.ci(boot.out = bootsmoke, conf = 0.95, type = "perc")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bootsmoke, conf = 0.95, type = "perc")
##
## Intervals :
## Level     Percentile
## 95%   (30648, 33406 )
## Calculations and Intervals on Original Scale
```

The 95% percentile bootstrap interval for medical costs of non-smokers can be calculated as follows:

```
bootnonsm = boot(data = data.frame(nonsmoke), statistic = bmean, R = 2000,
                 sim = "ordinary", stype = "i")
boot.ci(boot.out = bootnonsm, conf = 0.95, type = "perc")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bootnonsm, conf = 0.95, type = "perc")
##
## Intervals :
## Level     Percentile
## 95%   (8076, 8793 )
## Calculations and Intervals on Original Scale
```

Given a bootstrap sample, it is concluded that with 95% certainty, the true mean for medical treatment for smokers and non-smokers is between (30,648, 33,406) and (8,076, 8,793) receptively. The 95% bootstrap percentile intervals for μ_{smoke} and $\mu_{nonsmoke}$ validate the previously calculated confidence intervals. This confirms the notion that at the 95% level smokers have higher medical costs than non-smokers by a significant factor.

Conclusion

From this analysis, it has been concluded that the most significant factor affecting medical cost for individuals across in the United Stats has been smoking. A test of independence concluded that the expected BMI is dependent on the region that the sample is take. It is expected that the southwest region to have statistically significant higher portion of its population with a higher BMI than the other regions. As well, we expect that the Northeast region will have a higher portion of its population to be in the normal or ideal categories of BMI. With test a mean difference between regions, it was found that the southeast region does not experience a significant difference in true mean compared to that of the northwest region. The southeast region specifically saw higher then expected counts of obesity and overweight individuals while the northwest region fell closest to what what expected from our test of independence. For this reason, these two region, NW and SE, were selected for a t-sample t-test on the difference in mean medical costs in these regions. It was found that although there was unequal variance between these two group, there was still significant differences in medical costs between these two regions. Based on what was observed, the southeast region has a higher sample mean, which could be as a result of the higher density of overweight and obese individuals in that region, although these can not be concluded based on the current data provided.

Likewise, based on the liner regression model that was constructed and tested, it was conclude that BMI is likewise not a significant factor in explaining medical costs. BMI has a weak linear relationship with medical costs and a Box-Cox transformation was needed in order to justify the necessary assumptions in order to infer from this linear model. The model had a final R^2 value of 0.0176, which shows that BMI and medical costs have a very weak relationship, or that BMI does not explain medical costs effectively. As one of the direct

indicators of overall health, a confidence intervals for the medical charges for those who smoke compared to those who did not smoke was used. It was found that with 95% confidence the true mean for medical charges for both μ_{smoke} and $\mu_{nonsmoke}$ are respectively (30,683.65 , 33,416.82) and (8,074.12 , 8,794.41). Thus, those who smoke will experience up to 4 times more medical related costs then those who do not smoke, inferring the the adverse health risks associated with smoking.

Overall, with the rise in medical costs, it ultimately does not matter where you live or the current composition of your body. There are plenty of other factors that could go into overall health and the individuals need for medical services. There are plenty of diseases and illnesses that can occur outside the overall realm of health that could occur in almost anyone. However, it was found the biggest contributor to individual medical costs is the presence of smoking within this analysis. Hence, the greatest way for individuals to lower their medical costs is avoid smoking and lead a healthier lifestyle.

References

1) Kamal et al (2020) *How has U.S. spending on healthcare changed over time?*[Online]. Available at: https://www.healthsystemtracker.org/chart-collection/u-s-spending-healthcare-changed-time/#item-usspendingovertime_3 (Accessed: 12 Oct 2021)

2) PricewaterhouseCoopers (PWC) (2021) *Medical cost trend: Behind the numbers 2022*[Online]. Available at: https://www.pwc.com/us/en/industries/health-industries/library/behind-the-numbers.html (Accessed: 12 Oct 2021)