STA141A - Project: Factors of Insurance Cost

12/10/2021

Names	Contribution	Email
Jialin Chen	Setting up model, data clean up	jilichen@ucdavis.edu
Pile He	Visualization and Statistical Analysis, rmd editor	abhe@ucdavis.edu
Jiefei Li	Statistical Analysis, Function, Conclusion, Interpretation	jfeli@ucdavis.edu
Christie Ngo	Setting up model, data clean up, introduction, rmd editor, question of interest	ccngo@ucdavis.edu
Wesley Tat	Setting up model, data clean up, conclusion, rmd editor	wrtat@ucdavis.edu

A. Introduction

In the United States, the cost of healthcare has risen steadily over the years. The nation spends trillions of dollars on covering medical bills. Because the cost of healthcare in the U.S. is notoriously high, this topic is presently a crucial matter in politics. In 2020, an estimated 28 million Americans did not have healthcare. Revealing any trends on the state of healthcare will provide insights on necessary actions to take especially how much certain demographics are paying for medical costs.

B. Data Background and Questions of Interest

Data background: This dataset comes from Brett Lantz's dataset from his textbook, Machine Learning with R. There are 7 variables for 1,388 observations all pertaining to the primary beneficiary of the health insurance; this includes 4 continuous variables and 3 categorical ones. Age covers the age of the primary beneficiary, sex is the sex of insurance contractor, BMI refers to the body mass index, children is the count of dependents on health insurance, smoker is the status of whether or not one smokes, region is the residental area of the insurance contractor, and charges is the cost (USD) billed by the health insurance.

Questions of interest: We are interested in the factors that have the most influence on medical costs. More specifically, whether or not BMI, age, and smoker status affects insurance. We also want to know which region has the highest insurance costs and whether or not there is a relationship between BMI and being a smoker.

1. Which region has the highest insurance charge overall?
2. Does BMI, age, and smoker status have significant effects on medical costs?
3. Is there a relationship between BMI and being a smoker?
4. Which of these factors have the greatest influence on medical costs?

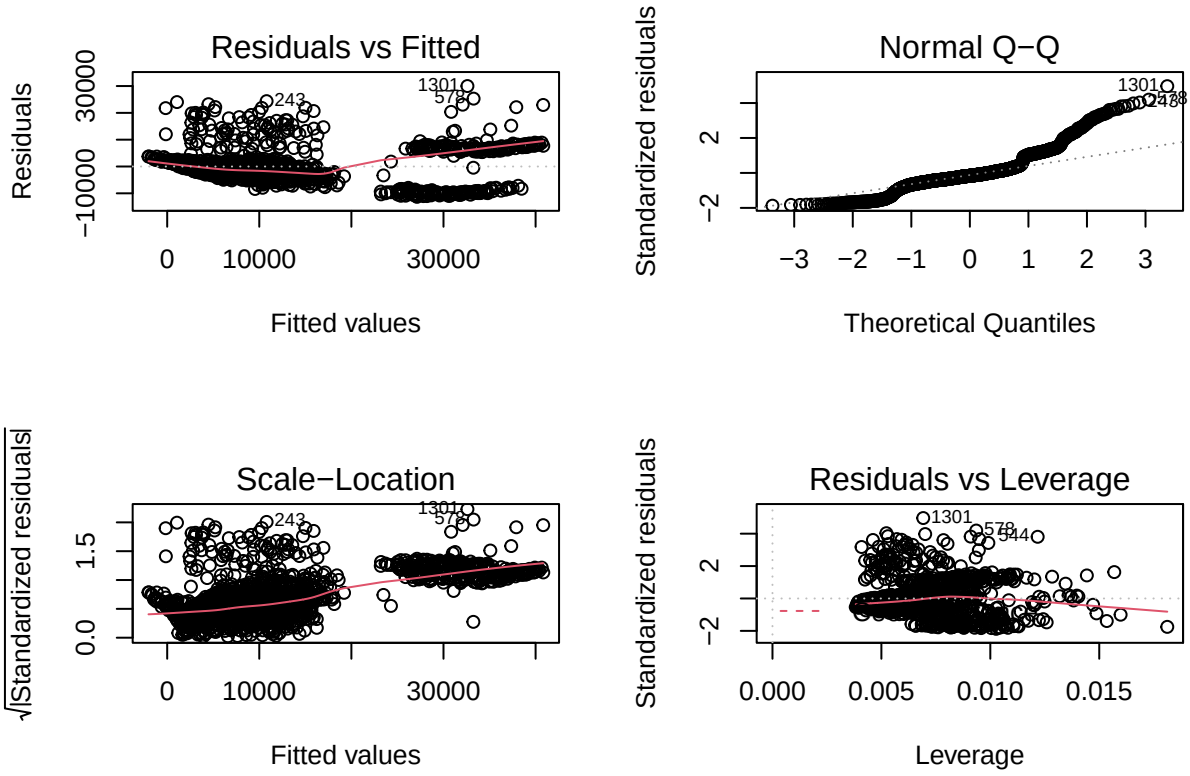
C. Methodology

We will analyze the insurance data with linear regression. We also will use transformation based on residual plots if needed to correctly assess our data. We also plan to subset the data if necessary. Additionally, we will use t-test to find the most important predictors. We will use boxplots and F-test for analysis.

D. Set up for the Model

Analysis of Full Model:

(i) Assumption of Full Model



First, looking at the residuals vs fitted graph, we can see that the data is heteroskedasticity as there is no random equal variability and instead there are patches and patterns of how the residuals are plotted. The data are in clusters as well which can indicate no random variability and heteroskedasticity. Additionally, the residuals vs fitted tells us that it is nonlinear as there is a pattern scattered around the residual vs fitted. We can also see there are potential outliers with point 1301, 578 and 243, with some leverages as the extreme X.

When we look at the normal qq-plot, the assumption of normality does not hold because most of the points are not surrounding the regression line. Instead it is non normal shape, and spread out from the regression line. It also indicates our three outliers which are points #243, #578, #1301. Since there are heavy tails, it shows that our data does not come from a Normal distribution.

When we test which points are outliers, we get 28 values in total which is very high amount of outliers meaning either our data does not work with linear regression or they are genuine outliers. Out of our 28 outliers, we have 5 points that are both outliers and leverages.

(ii) Summary of Full Model

For our full model, we created a multiple linear regression with our charges variable as the y variable. Additionally, the rest of the variables became predictor variables such as age, sex, region, children, bmi,

and smoker. The full model has a R^2 of 74.94% which is not strong indicator or a weak indicator. It is a moderate R^2 which indicates how the predictor variables can explain 74.94% of the changes in the response variable (charges). We can also discuss how the data is heteroskedasticity, does not come from a normal distribution, and is non linear. There are some outliers and leverage points. We can also discuss the how the residual standard error is very high in this model as it is 6062 which indicates how terrible a regression model would fit the dataset and how much the response will deviate from the true regression line.

When we do the F-Test, with the null hypothesis being all betas are equal to 0 and alternative hypothesis being at least one of the betas does not equal 0. We reject the null hypothesis when p < alpha, and we fail to reject the null hypothesis when p>alpha. Since our p value is 2.2e-16, and we can make our p = 0.001, p<alpha and we reject the null hypothesis. This indicates how we have at least one of the predictor values that are statistically significant. However, we must have further analysis on each individual predictor to determine whether they are statistically significant in terms of the data. We will also transform the data to fix the heteroskedasicity, nonlinearity, and violation of assumption of normality problem.

(iii) Model Decision

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-11939	987.8	-12.09	5.579e-32
age	256.9	11.9	21.59	7.783e-89
sexmale	-131.3	332.9	-0.3944	0.6933
bmi	339.2	28.6	11.86	6.498e-31
children	475.5	137.8	3.451	0.000577
smokeryes	23849	413.2	57.72	0
regionnorthwest	-353	476.3	-0.7411	0.4588
regionsoutheast	-1035	478.7	-2.162	0.03078
regionsouthwest	-960.1	477.9	-2.009	0.04476

We will run an individual hypothesis test on each variable to determine whether or not they are statistically significant. If they are not statistically significant, we will remove the variable as it is not necessary and does not impact the observation much.

The hypothesis test for each variable will be:

- Null hypothesis: beta# = 0
- Alternative hypothesis: beta# != 0

We reject the null hypothesis when p < alpha; we fail to reject the null hypothesis.

All predictors except for sex and regionnorthwest were less than alpha thus we were able to reject null hypothesis and say they were statistically significant. However, since sex was greater than alpha, we failed to reject the null hypothesis and it was statistically insignificant. Thus, we removed sex. We need to keep regionnorthwest because the regions are mutually exclusive. The region still makes a difference because it is a categorical variable. If the value does not add much to insurance cost that means that it does not matter if you are from regionnorthwest as the insurance there is relatively the same comparatively to other regions of the US.

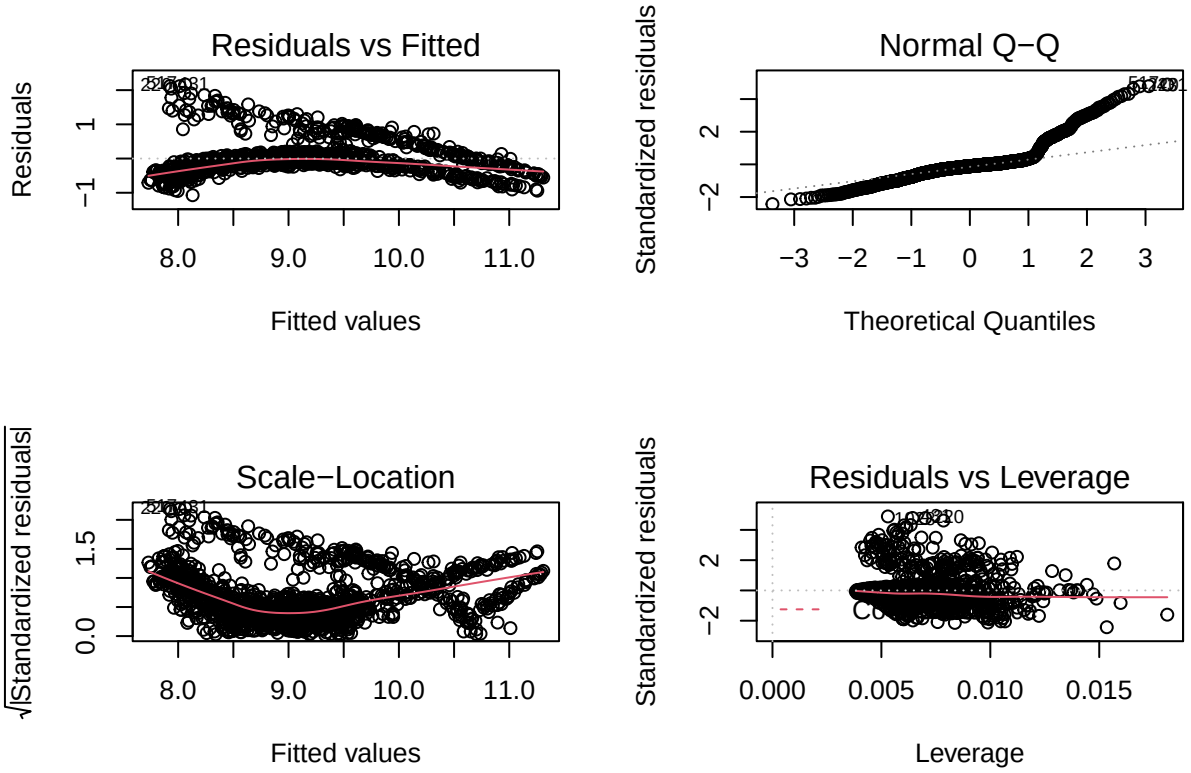
Reduced model: charges~age+bmi+children+region+smoker

We dropped sex as a predictor, as the gender is not statistically significant based on the hypothesis test shown above. We will transform the data to see if we can improve the model and the plots.

2. Transformations and Diagnostics

After plotting our fitted multiple linear regression, we saw that there was heteroskedasticity, nonlinearity, and violation of assumption of normality problem. Also, there is some outliers that needs to be fixed to allow us to use a multiple linear regression model.

We will transform the data by doing log transformation, square transformation, and inverse transformation. Whichever, performs the best by giving us the lowest residual standard error and high R^2 value.



When we transformed the data with log, the residual standard error become very small and the R^2 increase slightly. However, When looking at the residuals vs fitted, there is obvious patterns indicating heteroskedasticity. Also it is nonlinearity because there is relationship among the residual vs fitted which is not linear. When looking at the normal QQ plot we see that it is heavily tailed so assumption of normality is violated. Additionally, when we look at scale-location plot, we see how the data is not fixed from the original and there is many outliers and leverages point shown from residuals vs leverage as there are many extreme points.

When we transformed the data with sqrt, the residual standard error become somewhat smaller than the original as it is 22.38 and the R^2 increase slightly from the original summary before the reduced model. However, When looking at the residuals vs fitted, there is obvious patterns indicating heteroskedasticity with the three clusters, also nonlinearity since there is a relationship between residual and fitted shown through the clusters. When looking at the normal QQ plot we see that it is heavily tailed so assumption of normality is violated. Additionally, when we look at scale-location plot, we see how the data is not fixed from the original and there are many outliers and leverages point shown from residuals vs leverage as there are many extreme points.

When we transformed the data with inverse, the residual standard error become very small compared to the original as it is 0.0001087 and the R^2 significantly decreased from the original summary before the reduced model. However, When looking at the residuals vs fitted, there is obvious patterns indicating

heteroskedasticity with the relationship and curve between the residual vs fitted values and that shows how it is nonlinearity. When looking at the normal QQ plot we see that it is heavily tailed so assumption of normality is violated. Additionally, when we look at scale-location plot, we see that and there is many outliers. Additionally, there are many leverages point shown from residuals vs leverage as there are many extreme points.

After attempting to transform the data, it does not fix the original problems we have stated C) i) when we looked upon the plots at each transformation. However, it was able to fix the very high residual standard error when we did log transformation and increased our R^2 value slightly. While the other transformation such as square root and inverse did not help fix the plots and only helped with our R^2 and residual standard error as shown above.

We believe to fix this problem, we will attempt to subset the data between non smoker and smokers as that is a factor that could potentially increase insurance or decrease insurance for a person or not.

Smoker Subset

(i) Assumption of Full Model

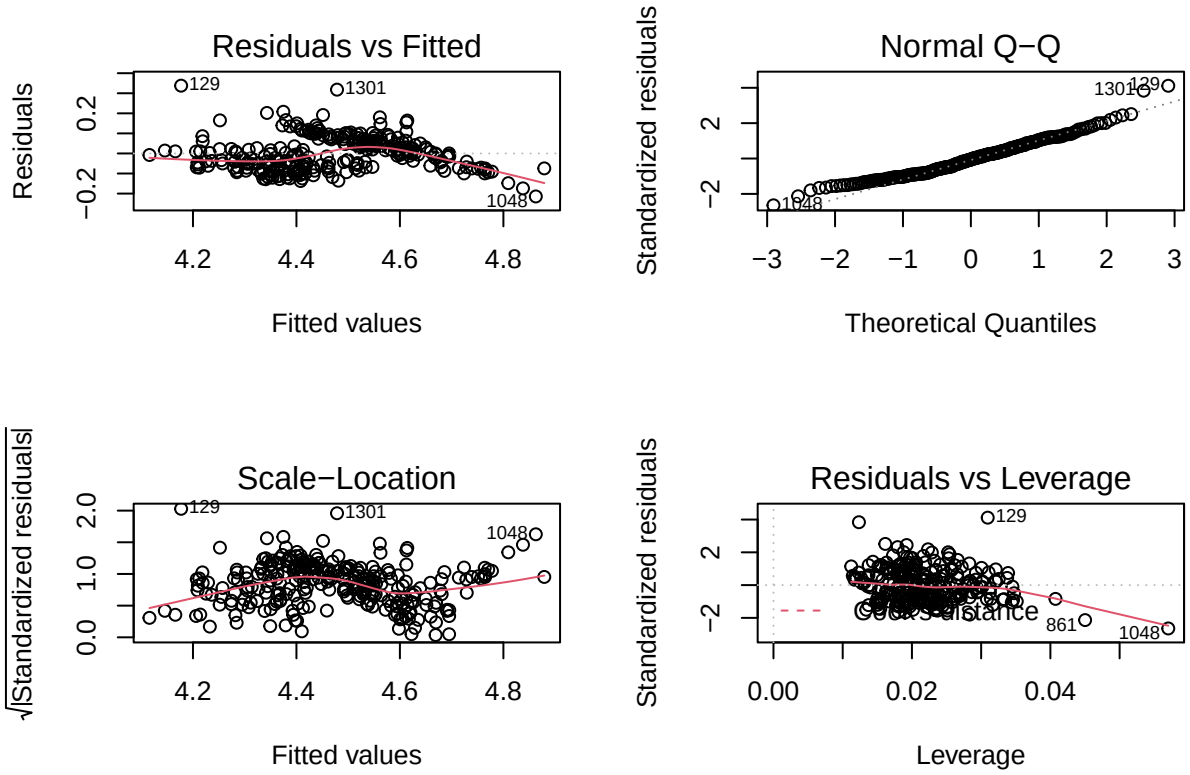
In the full model of the smoker subset, the R-squared is 0.7576, which means there is approximately 75.76% of the observed variation of the insurance charges can be explained by the measurement of BMI, age, children, and regions. The significant predicted variables are age, BMI, and region southeast. The Residual standard error is high in this model (5745), indicating that this regression model doesn't fit this dataset. However, there is a relationship between the predicted variables and response variable (Charges in medical insurance) because a small p-value ($< 2.2e\text{-}16$) is obtained in the F-test of overall significance. Therefore, we will analyze those significant predicted variables in-depth for more accurate predictions.

We drop the predicted variable children because it's not statistically significant (P value > 0.05). The assumptions of the normality, linearity, and homoscedasticity are violated. Therefore, we will use transformation methods to more closely meet the statistical assumptions and reduce the residual errors.

(ii) Transformations and Decisions

The inverse transformation has the lowest R-squared and the residuals error is 7.629e-06. The Normal Q-Q plot indicates that it fix the assumptions of normality, but not the linearity and homoscedasticity in residual plot.

The square-root transformation has the highest R-squared and the residuals error is 16.22. The residuals and normal Q-Q plots indicate that they hold the assumptions of normality, linearity and homoscedasticity.



0.7291 <0.7588<0.7627, A higher r-squared indicates a better fit for the model. Although square-root transformation has the highest the r-squared, the r-squared in log transformation is only 0.005 smaller and residual error (0.08342) is much smaller than square-root transformation.Therefore, we decided to use log transformation for the significant predicted variables.

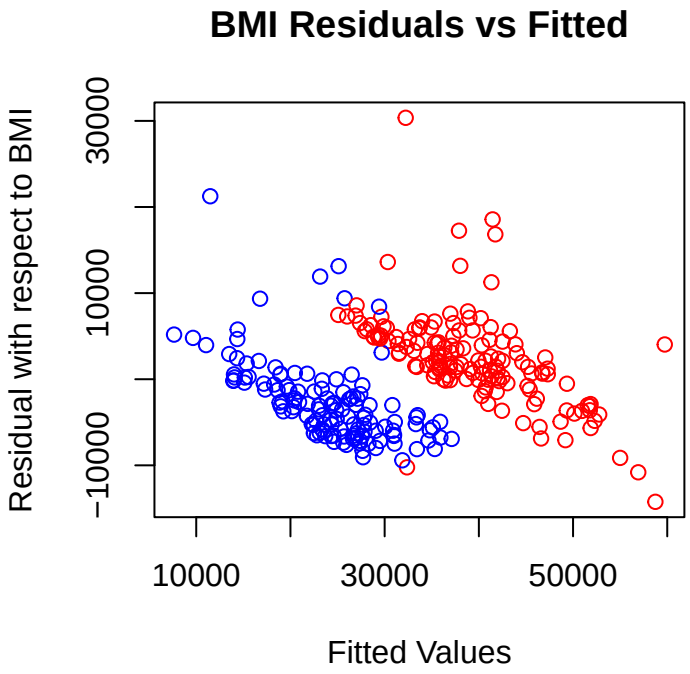
The R^2 for square transformation is 0.7588, which is slightly greater than the full model. There is approximately 75.88% of the observed variation of the insurance charges can be explained by the measurement of BMI, age and regions.

The residuals vs Fitted plot, roughly holds the assumption of linearity and homoscedasticity because most of the observed dots are randomly around the 0, and some dots in the middle are scattered above the residual line. Observations 129th, 1301th and 1048 are appear to be the outliers.

The Normal Q-Q plot, the assumption of normality holds because the majority of the observed dots are aligned with the line. However, there are two potential outliers located on upper tail (129th, 1301th) and one on the lower (1048th).

For the leverage point plot, (5+1)/274=0.02189, there are nearly half of the observations above the horizontal line 0.02.

There are 2 upper outliers (observations 129th 4.115 and 1301th 4.667), and a total of 120 high leverages points.



There are two groups in the smoker subset, the blue observations represent the patients who are smokers with a BMI less than 30, and red observations indicated that smokers with a BMI equal or greater than 30. The patterns of the residual errors appear on both groups are consistent, as the fitted values increased the residuals decreased.

The reason why we did not further analyze BMI in the smoker subset is because there were fewer points to draw conclusions from; this may lead to overrepresented or underrepresented conclusions. The heteroscedasticity is more prevalent due to the inconsistency in the datapoints and linear regression is not a suitable model to make predictions in that case. Therefore, we need more more observations and in-depth statistical analysis for this smoker subset.

Nonsmoker subset

When performing linear regression, we find that BMI is no longer a significant predictor because the p-value is 0.465 and we will drop this variable. There is a lower R^2 value; the model only explains 40.78% of the variance in charges. To remedy this, we will choose the best transformation.

The quantile-quantile plot curves towards the end and the residual plots are not random. This model explains 68.12% of variance in charges.

This quantile-quantile plot shoes the most curvature and the residuals still show a pattern. This model only explains 56.90% of the variance in charges.

This quantile-quantile plot now shoes curvature on both tails and the residuals show a distinct curved pattern.

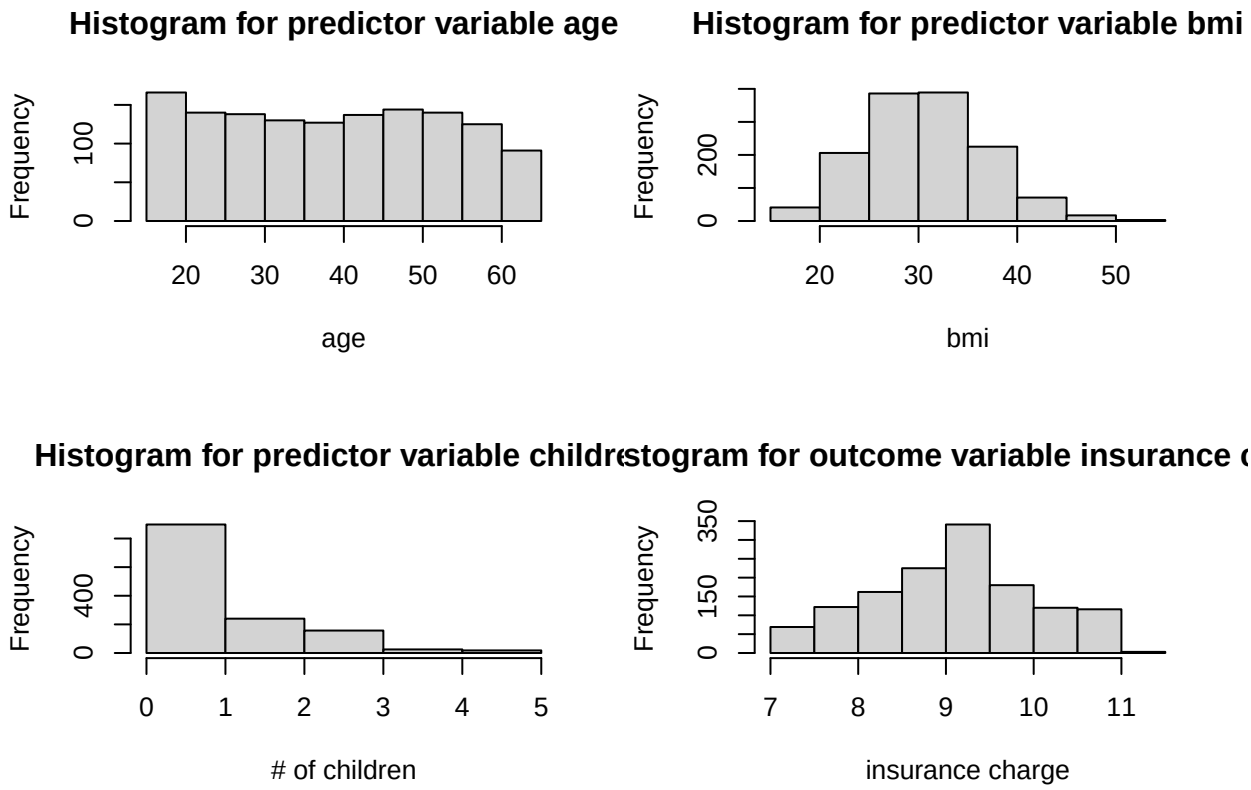
We chose the log transformation with the highest R^2 value. The quantile-quantile curves towards the end. The residuals also show a pattern. We have not completely resolved our violations but we determined this was sufficient to move forward. Our R^2 value has decreased to 68.12% but the residual standard error has decreased drastically to 0.4214.

```
plot(summary(nonsmoker_fit2)$residuals~fitted(nonsmoker_fit2),col=ifelse(nonsmoker$bmi>=30,"red","blue"))
```

BMI varied greatly within nonsmokers; the points are scattered without distinct groups. Subsetting by BMI does not reveal any significant information about insurance charges.

E. Visualizations and statistical analysis

(i) Variable summary



The predictor variable age does not have much variation. All ages seem to have the same frequency. The predictor variable BMI has a bell shape curve where the median lies at around BMI score of 30. The predictor variable Children is skewing to the right with most of the participants having 0 to 1 child. The outcome variable insurance charge is also skewing to the right and most of the data is below $15000. The correlation coefficient between age and BMI is 0.11 and the correlation coefficient between age and number of children is 0.04. Both are very weak. The regions and sex categorical variables are distributed evenly. But for the smoker variable, there are 1064 non-smokers and 274 smokers.

(ii) F test of Smoking and Non-smoking Standardized Regression

Since the F-statistic p-values are small ($< 2.2e\text{-}16$) for both two models (smoker and non-smoker), we would reject the null hypotheses and conclude that these predictor variables and log(insurance) have significant relationship for both smoking and non-smoking population.

(iii) T Test of Smoking Standardized Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.456	0.06712	126	7.718e-240
age	0.009088	0.0008443	10.76	1.113e-22
bmi	0.04928	0.001933	25.49	1.691e-73
children	0.006841	0.01014	0.6744	0.5006
regionnorthwest	-0.01196	0.03459	-0.3458	0.7298
regionsoutheast	-0.05557	0.03218	-1.727	0.08533
regionsouthwest	-0.005959	0.03491	-0.1707	0.8646

This model has an R^2 value equal to 0.7592, indicating a good fit for the data.

From the table above, it shows that only age and bmi have p values less than the 0.05 level of significance. Thus, for smoking population, only age and BMI have significant impact on insurance rate.

Both BMI and age have a positive impact on the insurance rate indicated by their positive coefficients. When BMI increases by 1, the log(insurance) will increase by 0.04928, meaning the insurance will be multiplied by 1.050514. When age increases by 1, the log(insurance) will increase by 0.009088, meaning the insurance will be multiplied by 1.009129.

(iv) T Test of Smoking Standardized Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.079	0.07585	93.34	0
age	0.04157	0.0009217	45.1	1.437e-248
bmi	0.001301	0.002238	0.5812	0.5612
children	0.1283	0.01056	12.15	6.874e-32
regionnorthwest	-0.07491	0.03662	-2.046	0.04102
regionsoutheast	-0.1703	0.03759	-4.531	6.545e-06
regionsouthwest	-0.1816	0.03671	-4.947	8.779e-07

This model has an R squared value equal to 0.685, indicating a moderate fit for the data.

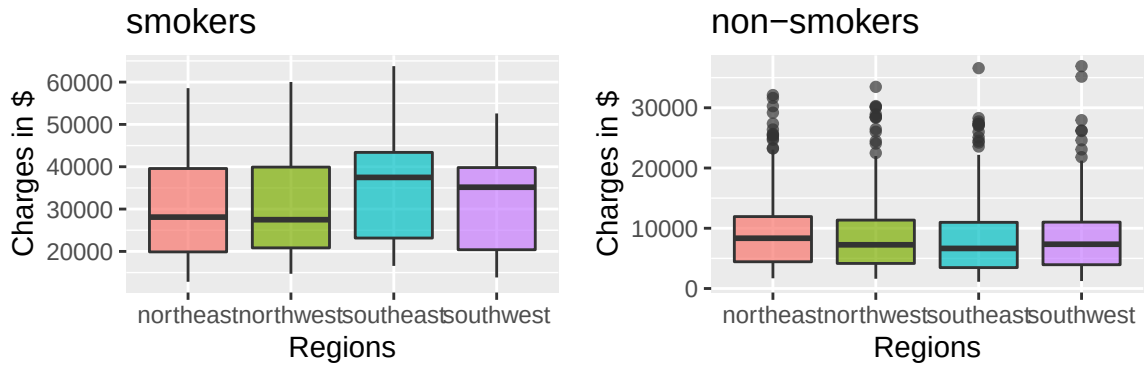
From the table above, it shows that age, children, and region have p values less than the 0.05 level of significance. Thus, for non-smoking population, age, children, and region have significant impact on insurance rate.

Age and children have a positive impact on the insurance rate, indicated by their positive coefficients. When the age increases by 1, the log(insurance) will increase by 0.0415749, meaning the insurance will be multiplied by 1.042451. When the number of children increases by 1, the log(insurance) will increase by 0.1282896, meaning the insurance will be multiplied by 1.136882.

However, if the region falls into northwest, southeast, and southwest, it will have an negtaive impact on the insurance comparing to the default region, northeast. For region equal to northwest, the log(insurance) will decrease by 0.0749107, meaning the insurance will be multiplied by 0.9278263 comparing to the northeast. For region equal to southeast, the log(insurance) will decrease by 0.1702948, meaning the insurance will be multiplied by 0.8434161 comparing to the northeast. For region equal to southwest, the log(insurance) will decrease by 0.1815788, meaning the insurance will be multiplied by 0.8339525 comparing to the northeast.

F. Interpretation and Reporting

Which region has the highest insurance charge overall?



By only looks at the boxplot for insurance charges vs regions boxplot, all regions have relatively similar median of insurance, with northeast having slightly higher medium insurance rate and southeast with slightly higher range. However, when we separate the smokers with non-smokers by subsetting the entire data set into two data sets, the boxplot of smokers data set shows that the median of southeast and southwest regions are significantly higher than the medians of northeast and northwest. For example, the median insurance charges for southeast smokers is 37000, while the median insurance charges for northwest smokers is only 27000. We could observe a huge difference in the insurance charges between southern regions and northern regions. And this effect does not hold true for the non-smokers. As the third plot shows that the medians of insurance cost of all four regions are roughly the same and all around the amount of 7500. It might because insurance companies that are popular in different regions have different criteria for calculating the cost of insurance based on the smoking status. It also might because in our data set there are only 274 smokers vs. 1064 non-smokers. The trend might be caused by the insufficient amount of data for smokers.

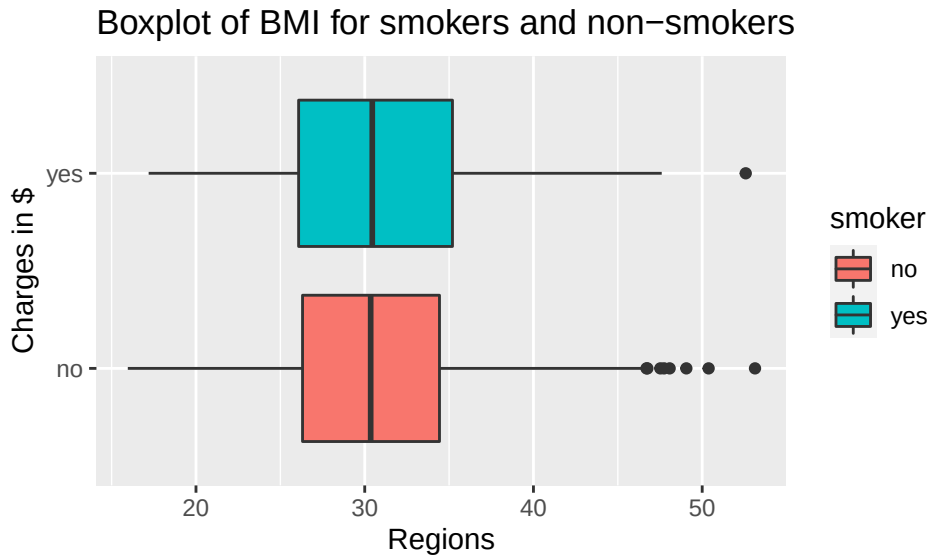
Does BMI, age, children, and smoker status have significant effects on medical costs?

Being a smoker or not is characterized as a very significant factor that affects the normalization of our dataset, so we separate the whole dataset into smoker and non-smoker subgroups.

In the smoker groups, BMI and age both have a positive impact on medical costs, meaning increase in BMI and age will lead to increase in medical costs.

In the non-smoker groups, age and number of children have a positive impact on medical costs, meaning increase in age and number of children will lead to increase in medical costs. Additionally, northwest, southeast, and southwest tend to have a lower medical costs comparing to northeast. The medical cost is ranke by northeast $>$ northwest $>$ southeast $>$ southwest.

Is there a relationship between BMI and being a smoker?



The boxplot of smokers vs. non-smokers shows that overall there is no significant difference between the BMI scores of smokers and non-smokers. The inter quartile range for smokers are slightly higher than the IQR for non-smokers. This means that even the amount of non-smokers is almost four times of the amount of smokers in our data set(1064 vs. 274), the variation of BMI of smokers is slightly higher than the variation of BMI for non-smokers. So this could be an evidence for how smoking could affect BMI by increasing the variation of BMI.

Which of these factors have the greatest influence on medical costs?

By calculating the absolute value of the estimates for each variable in smoker model and non-smoker model, we conclude that for smoker model, being in the region southeast has the largest absolute value of the estimated coefficient which is 0.0555. And for the non-smoker model, the most significant factor is region southwest with an absolute value of estimate of 0.1816. Yet it is worth to notice that the variable region southeast and all other categorical regions factors could only be either 0 or 1. So even they have relatively large estimated coefficients, the net change they bring to the insurance charge is still going to be small since there are continuous variables such as age and BMI that could change drastically as opposed to only one unit change for region variables. For example, the factor BMI in the smoker case has an estimated coefficient of 0.0493, which is slightly smaller than the estimate of the region southeast. But BMI varies way more aggressive than region. As shown in the visualization part, BMI has a bell curved shape and most of participants has a BMI from 20 to 45. So as a result, the factor BMI will contribute way more for the insurance charge than region factor. Therefore, in order to decide which factor is indeed the most significant, we need more in depth statistical analysis method and choose a reasonable criteria in the future.

G. Conclusion

Originally our dataset had multiple predictors, such as the sex of a person, the BMI of a person, the age of a person, the number of children the person has, the region of the country they live in, and whether they were a smoker or nonsmoker were considered for their contribution to the cost of insurance. We eliminated sex as it was not statistically significant. Then, we attempted to transform the data to possibly transform our raw data to be more homoscedasticity, more linear, and eliminate the leverages and outliers. However, assumption of homoscedasticity and assumption of linearity were still violated. Additionally, much of the outliers and leverage points remained. After assessing the data, we determined that we should subset the model based on smoking status based on our reduced model. After transforming the data, the coefficient of determination improved slightly, and the residual standard error dropped dramatically. Splitting the data to smokers and nonsmokers assumption of homoscedasticity and assumption of linearity were not violated.

In order to find out which region has the most insurance charges, we performed boxplots to both the model with smokers and non-smokers. The boxplot of non-smokers showed that all four regions have almost the same median and range for insurance charges. But the smoker model shows that the region southwest and southeast have a significantly higher median of insurance charges. To answer the question that asks if there is a relationship between smoking status and BMI, we created a boxplot to see if there is a difference between smokers and non-smokers. And as a result, the boxplot for smokers is almost the same as the boxplot for non-smokers only with a slight difference in the IQR of the smokers. This indicates that whether smoking or not has less to no effect on the BMI of a person based on our data.

In addition, to find out which predictor variable is the most significant for each model, we compared the absolute value of estimates. And as a result, the most significant variable for smoker model is region southeast, and for non-smoker model is region southwest. And it is worth mentioning that since region variables could only be 0 and 1, the net change the region variables bring to the dependent variable insurance charge is not that significant. So in the future we should consider using more in-depth analysis and more comprehensive criteria for deciding which variable is the most significant for a model. The United States compared to many first world countries like Germany, Switzerland, and Canada, do not have universal nationwide health insurance [3]. The purpose of health insurance is to protect oneself against the high price of getting medical attention. Health insurance is purchased from many private organizations that charge exorbitant prices for their insurance. The reason for the high price tag on insurance is because of the growing for-profit healthcare facilities [4]. Due to the cost of insurance, many people are uninsured as they cannot afford health insurance in the United States. That leaves about 16% of the United States uninsured, which is around 50 million citizens who are currently uninsured [3].

In this project, we are analyzing which factors are the most significant when determining charges for insurance as it is a largely debated issue in the US as many push for universal health insurance. After thorough statistical analysis, we determined which region of the US, based on if you were smoker or nonsmoker, were the most significant factors as they had the largest absolute values compared to the other predictors. Living in the southeast has the highest insurance charges. We saw significant changes whether or not if you were smoker and nonsmoker for insurance cost. If we look at our question of interest "Which region has the ighest insurance charge overall", we divided it into smoker and nonsmoker and we can see how smokers have significant more cost than nonsmokers. This can be contributed to how smokers are susceptible to lung cancer while nonsmokers are not [5]. However, we concluded that BMI varies more drastically than region but the data was not normally distributed. Therefore, we need a more in depth statistical analysis method to determine how BMI could be more determining than region.

H. How much does your insurance cost?

```
age = as.integer(readline(prompt="Enter age: "))
```

```
## Enter age:
```

```
bmi = as.integer(readline(prompt="Enter BMI: "))
```

```
## Enter BMI:
```

```
child = as.integer(readline(prompt="Enter number of children: "))
```

```
## Enter number of children:
```

```
smoker = as.integer(readline(prompt="Are you a smoker? 1 for Yes and 0 for No "))
```

```
## Are you a smoker? 1 for Yes and 0 for No
```

```
region = readline(prompt="Where are you from? Enter one: southeast, southwest, northeast, northwest ")
```

```
## Where are you from? Enter one: southeast, southwest, northeast, northwest
```

```
sw = 0
se = 0
nw = 0
if(region == "southwest"){
  sw = 1
} else if(region == "southeast"){
  se = 1
}else if(region == "northwest"){
  nw = 1
}

insurance_calculator = function(age, bmi, child, smoker, sw, se, nw){
  if(smoker == 1){
    charge = 8.456 + 0.009088*age + 0.04928*bmi + 0.006841*child - 0.012*nw - 0.05557*se - 0.006*sw
  } else if (smoker == 0){
    charge = 7.079 + 0.04157*age + 0.001301*bmi + 0.1283*child - 0.075*nw - 0.1703*se - 0.1816*sw
  }
  cat("The estimated insurance charge for you is ", exp(charge))
}
## We put age = 21, bmi = 20, children = 1, smoker = yes, region = notheast as eample
insurance_calculator(age=21, bmi=20, child=1, smoker=1, 0, 0, 0)
```

```
## The estimated insurance charge for you is  15356.42
```

References

1. Bureau, US Census. "Health Insurance Coverage in the United States: 2020." Census.gov, 18 Oct. 2021, https://www.census.gov/library/publications/2021/demo/p60-274.html.
2. "National Health Accounts Historical", https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical.
3. Ridic, Goran, et al. "Comparisons of Health Care Systems in the United States, Germany and Canada." Materia Socio-Medica, AVICENA, D.o.o., Sarajevo, 2012, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3633404/.
4. Health Care for Profit or People? https://www.scu.edu/mcae/publications/iie/v1n4/healthy.html.
5. Tindle, Hilary A, et al. "Lifetime Smoking History and Risk of Lung Cancer: Results from the Framingham Heart Study." Journal of the National Cancer Institute, Oxford University Press, 1 Nov. 2018, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6235683/.

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(MASS)
library(class)
library(pander)

getwd()
med_in = read.csv("insurance.csv")
med_lm = lm(charges ~ age+sex+bmi+children+smoker+region, data=med_in) #charges ~ . plots charges as Y
#is the rest of the predictor variable
sum.med = summary(med_lm) # dropping all regions because they do not matter
#as they do not pass the hypothesis test
#(sum.med$residuals~fitted(med_lm), col=ifelse(med_in$smoker=="no","blue","red"))
#plot(sum.clean$residuals~fitted(clean_lm),
#col = ifelse(clean_med$bmi >= 30, "red", "blue"))
par(mfrow = c(2, 2))
plot(med_lm)
outliers = rstandard(med_lm)[rstandard(med_lm) < -3 | rstandard(med_lm) > 3]
indices = as.numeric(names(outliers))
leverages = hatvalues(med_lm)
select = leverages[indices]
pander(select[leverages[indices] > ((12)/1388)])
#head(hatvalues(med_lm))
#length(hatvalues(med_lm)[hatvalues(med_lm) == outliers])
sum.med = summary(med_lm)
pander(sum.med$coefficients)
#lm_nonsmoker1 = lm(charges~ age+bmi+children+region+smoker, data =med_in)
#summary(lm_nonsmoker1)

ins_loglm = lm(log(charges)~age+bmi+children+smoker+sex+region,data = med_in)
par(mfrow = c(2, 2))
plot(ins_loglm)
ins_sqrtlm = lm(sqrt(charges)~age+bmi+children+region+smoker, data = med_in)
```

```r
summary(ins_sqrtlm)
par(mfrow = c(2, 2))
plot(ins_sqrtlm)
ins_inverselm = lm((1/charges)~age+bmi+children+region+smoker, data = med_in)
summary(ins_inverselm)
par(mfrow = c(2, 2))
plot(ins_inverselm)
smoker= med_in[(med_in$smoker == "yes"),]
head(smoker)
smoker.fit <-lm((charges)~age+bmi+children+region,smoker)
par(mfrow = c(2, 2))
plot(smoker.fit)
sum.smoker <-summary(smoker.fit)
#sum.smoker
smoker.fit<-lm((charges)~age+bmi+region,smoker)
sum_smoker.fit<-summary(lm((charges)~age+bmi+region,smoker))
par(mfrow=c(2,2))
plot(smoker.fit)
smoker_inv.fit <-lm(1/(charges)~age+bmi+region,smoker)
sum_smoker_inv.fit<-summary(smoker_inv.fit) #0.7291
par(mfrow = c(2, 2))
plot(smoker_inv.fit)
smoker_sqrt.fit <-lm(sqrt(charges)~age+bmi+region,smoker)
sum_smoker_sqrt.fit <-summary(smoker_sqrt.fit) #0.7627
par(mfrow = c(2, 2))
plot(smoker_sqrt.fit)
smoker_log.fit <-lm(log10(charges)~age+bmi+region,smoker)
sum_smoker_log.fit<-summary(smoker_log.fit) #0.7588
par(mfrow = c(2, 2))
plot(smoker_log.fit)
#Outliers
smoker_outliers <- rstandard(smoker_log.fit)[rstandard(smoker_log.fit) < -3 |
                            rstandard(smoker_log.fit) > 3]
#leverage point
p <- 5
n <- length(rownames(smoker)) #274
n
high.leverage = which(hatvalues(smoker_log.fit)>(p+1)/n)
high.leverage
length(high.leverage)
plot(sum_smoker.fit$residuals~fitted(smoker.fit),col=ifelse(smoker$bmi>=30,"red","blue"), main = "BMI R
nonsmoker = med_in[(med_in$smoker == "no"),]
nonsmoker_fit = lm(charges~bmi+children+age+region, data=nonsmoker)
outliers_ns = rstandard(nonsmoker_fit)[rstandard(nonsmoker_fit) < -3 | rstandard(nonsmoker_fit) > 3]
indices_ns = as.numeric(names(outliers_ns))
leverages_ns = hatvalues(nonsmoker_fit)
select_ns = leverages_ns[indices_ns]
high_leverage = select_ns[leverages_ns[indices_ns] > ((4+1)/1064)]

nonsmoker_clean = nonsmoker[!seq_len(nrow(nonsmoker)) %in% na.omit(as.numeric(names(high_leverage))),]
nonsmoker_fit2 = lm(charges~bmi+children+age+region, data=nonsmoker_clean)
par(mfrow = c(2,2))
plot(nonsmoker_fit2)
```

```r
nonsmoker_log = lm(log(charges)~children+age+region, data=nonsmoker_clean) # R^2 0.6812

nonsmoker_squareroot = lm(sqrt(charges)~children+age+region, data=nonsmoker_clean) # R^2 0.569

nonsmoker_inverse = lm((1/charges)~children+age+region, data=nonsmoker_clean)# R^2 0.6596

# choose LOG transformation
par(mfrow=c(2,2))
plot(nonsmoker_log, main = "Log Transformation")
par(mfrow=c(2,2))
plot(nonsmoker_squareroot, main = "Square Root Transformation")
par(mfrow=c(2,2))
plot(nonsmoker_inverse, main = "Inverse Transformation")
plot(summary(nonsmoker_fit2)$residuals~fitted(nonsmoker_fit2),col=ifelse(nonsmoker$bmi>=30,"red","blue")
library(pander)
insurance_nonsmoker = med_in[(med_in$smoker == "no"),]
insurance_smoker = med_in[!(med_in$smoker == "no"),]
insurance_smoker$region = as.factor(insurance_smoker$region)
insurance_nonsmoker$region = as.factor(insurance_nonsmoker$region)
lm_nonsmoker = lm(log(charges)~ age+bmi+children+region, data = insurance_nonsmoker)
lm_smoker = lm(log(charges)~ age+bmi+children+region, data = insurance_smoker )
par(mfrow=c(2,2))
hist(med_in$age, xlab = "age", main = "Histogram for predictor variable age")
hist(med_in$bmi, xlab = "bmi", main = "Histogram for predictor variable bmi")
hist(med_in$children, xlab = "# of children", main = "Histogram for predictor variable children", breaks
hist(log(med_in$charges), xlab = "insurance charge", main = "Histogram for outcome variable insurance cl
pander(summary(lm_smoker)$coef)
pander(summary(lm_nonsmoker)$coef)
library(ggplot2)
ggplot(insurance_smoker, aes(x=as.factor(region), y=charges, fill=region))+
    geom_boxplot(alpha=0.7) +
    theme(legend.position="none")+
    ggtitle("smokers") +
    xlab("Regions") + ylab("Charges in $")
ggplot(insurance_nonsmoker, aes(x=as.factor(region), y=charges, fill=region))+
    geom_boxplot(alpha=0.7) +
    theme(legend.position="none")+
    ggtitle("non-smokers") +
    xlab("Regions") + ylab("Charges in $")
ggplot(data = med_in, aes(x= bmi, y = smoker, fill = smoker)) +
  geom_boxplot()+
  ggtitle("Boxplot of BMI for smokers and non-smokers") +
    xlab("Regions") + ylab("Charges in $")
age = as.integer(readline(prompt="Enter age: "))
bmi = as.integer(readline(prompt="Enter BMI: "))
child = as.integer(readline(prompt="Enter number of children: "))
smoker = as.integer(readline(prompt="Are you a smoker? 1 for Yes and 0 for No "))
region = readline(prompt="Where are you from? Enter one: southeast, southwest, northeast, northwest ")
sw = 0
se = 0
nw = 0
if(region == "southwest"){
  sw = 1
```

```
} else if(region == "southeast"){
  se = 1
}else if(region == "northwest"){
  nw = 1
}

insurance_calculator = function(age, bmi, child, smoker, sw, se, nw){
  if(smoker == 1){
    charge = 8.456 + 0.009088*age + 0.04928*bmi + 0.006841*child - 0.012*nw - 0.05557*se - 0.006*sw
  } else if (smoker == 0){
    charge = 7.079 + 0.04157*age + 0.001301*bmi + 0.1283*child - 0.075*nw - 0.1703*se - 0.1816*sw
  }
  cat("The estimated insurance charge for you is ", exp(charge))
}
## We put age = 21, bmi = 20, children = 1, smoker = yes, region = notheast as eample
insurance_calculator(age=21, bmi=20, child=1, smoker=1, 0, 0, 0)
```