What determines insurance coverage?

Ann Mudanye, Anuska Jain, Bushra Tasneen, Phuong Chau

12/21/2018

Abstract

The cost of healthcare insurance is higher in the United States compared to many other developed nations (Frakt et al, 2018). Healthcare costs are expected to increase by 5.5 % annually on average from 2017 to 2026 (Abutaleb, 2018). In this study, we are investigating whether there are any significant associations between certain demographic characteristics of an individual and the annual cost of healthcare charged to them. The hypothesis of our study aims to find whether characteristics such as whether a person is a smoker,the number of children they have,their sex, geographic region, BMI (Body Mass Index) or age are significantly associated with the annual healthcare insurance cost. Initially, we attempted to construct linear models. However, we could not make conclusions about which of the predictors are significant because the model violated conditions of normality and homoscedasticity. Therefore, we generated bootstrap distributions for each of the predictors that we had initially considered. Using the bootstrap distribution, we conclude that the characteristics which are significantly associated with the cost of healthcare insurance are: whether a person is a smoker, their age, their BMI, their region of residence and the number of children they have.

Introduction

This study examines whether or not factors such as age, sex, body mass Index (BMI), region, number of children and tobacco use of an insurance member have an effect on the insurance cover for the member.

Insurance costs are increasing in the United States. One news article reports that health insurance costs in the US have been increasing over time and in 2016 health insurance costs for a typical family of 4 was over $25,000. They also report that on average, health care costs for a family of 4 go up $100 every month on average (USA Today, 2018). Given that health care costs are increasing over time, various articles have been published online with tips for consumers on how to choose the best health insurance plan and how to reduce their insurance premium. One such article suggested that consumers should quit smoking to lower their insurance costs (Insurance.com, 2017). With insurance costs increasing and consumers looking to reduce their health care costs, we seek to use regression analysis in evaluating whether or not there is a link between the factors mentioned earlier (age, sex, BMI, region, number of children and tobacco use of an insurance member) and insurance coverage for the member.

We expect that women, individuals with a high BMI, older people, smokers, people living in the southeast region and individuals with children will incur higher insurance costs than their counterparts.There is also the likelihood of women getting pregnant in their lifetime and having to cater for the maternity insurance can be very costly to companies. Individuals with a higher BMI are prone to develop diseases such as diabetes and have complications such as joint problems. In the case of pregnancy, having a high BMI is normal but still results in high medical costs. As individuals get older, the immune system is bound to get weaker meaning more doctor visits. Smokers are likely to suffer from lung complications. If an individual has children, insurance companies have to cater for the extra costs of the respective children (Botkin). This leaves us with questions such as what value of any of the given factors will result in an increased insurance cost per individual. Overall, the goal of this study is to explore the question: What factors are most significant in determining insurance costs for an individual? Understanding that some of these factors such as sex cannot be controlled by an individual, people do have the ability to lose or gain weight depending on their BMI, quit smoking, or move to a different location. The information found through our exploration could help individuals be aware of what factors to control hence taking a step towards reducing their health insurance costs.

Data

We found and downloaded the data from Kaggle. The data was originally shared by a GitHub contributor in 2015 along with other data sets from the textbook called Machine Learning with R by Brett Lantz. The data includes information of 1338 insurance beneficiaries. Our observational units are these 1338 individual beneficiaries. This data set was created using demographic statistics from the U.S. Census Bureau. However, we do not have information regarding which census year was used for the data. This is a population data set containing randomly picked beneficiaries from the four major geographic regions in the U.S namely Northeast, Southeast, Southwest and Northwest. We will be generalizing our findings to the population of the U.S.

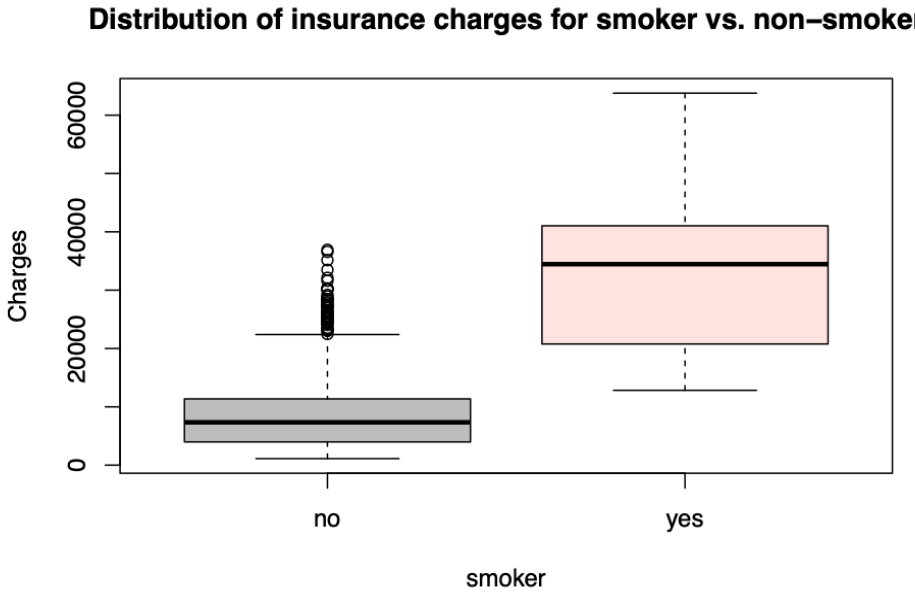
Our response variable is the annual individual medical costs billed by health insurance providers in dollars as recorded during a census year. This cost has not been adjusted for inflation. Insurance costs range from $112 to $63770 with the mean and median cost being $13270 and $9382 respectively.

Our predictor variables include the other variables reported in the data. They are:

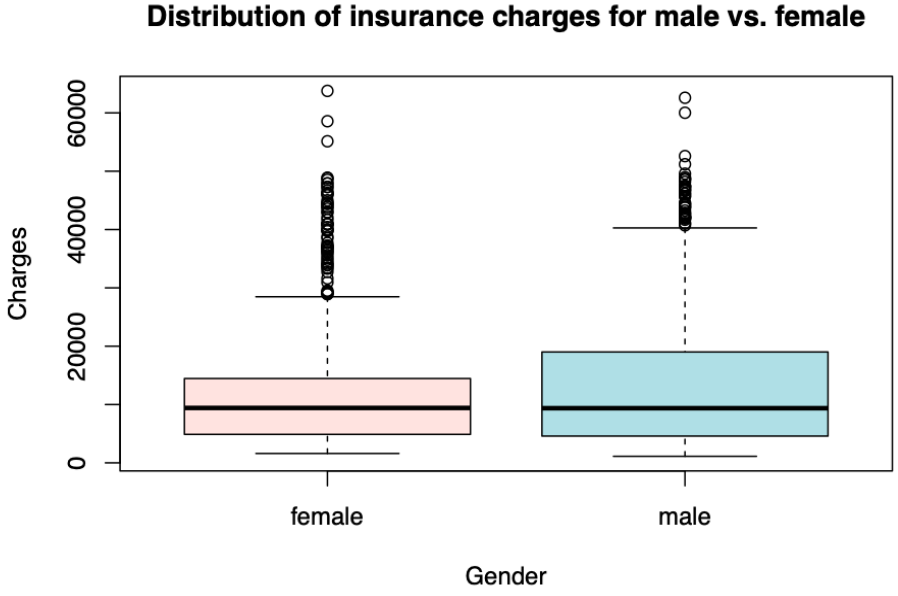
1. Age: Quantitative unit: The age of the individual whose medical charges we are calculating. Age is a continuous numerical variable. Age ranges from 18-64 years old with the mean and median age being 39.21 and 39 years old respectively.
2. Sex: Categorical variable: Sex of the individual which falls into one of the categories: Male or Female
3. BMI: Quantitative unit: BMI (Body Mass Index) is defined as the mass divided by the height squared. BMI is expressed in the unit kg/m2. BMI ranges from 15.96-53.13 kg/m2 with the mean and median BMI being 30.60 and 30.40 kg/m2 respectively.
4. Children: Quantitative unit: This variable tells us the number of children the individual has. The number of children ranged from 0-5 children with the mean and median number of children being 1.095 and 1 respectively.
5. Smoker: Categorical variable: Smoker is a categorical variable which indicates whether the individual is a smoker or not. The two categories under this variable are yes and no.
6. Region: Categorical variable : Region is a categorical variable that indicates the residential location of an individual. The four categories under this variable are Northeast, Northwest, Southeast and Southwest.

Summary Statistics

We first acquainted ourselves with the data before constructing any models. We created boxplots for every categorical predictor variable and scatterplots for the quantitative variables to see what the effect of a category was on insurance charges and whether or not there were differences between these categories.

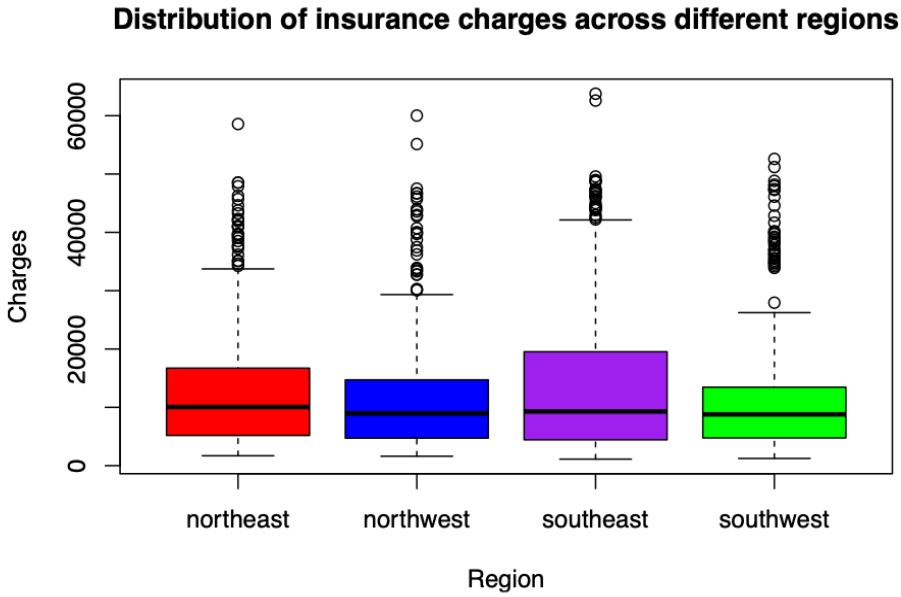


As can be seen from the boxplot, smokers pay higher insurance charges as compared to non-smokers. This is consistent with what we had thought as smokers are vulnerable to suffering from respiratory and lung diseases.

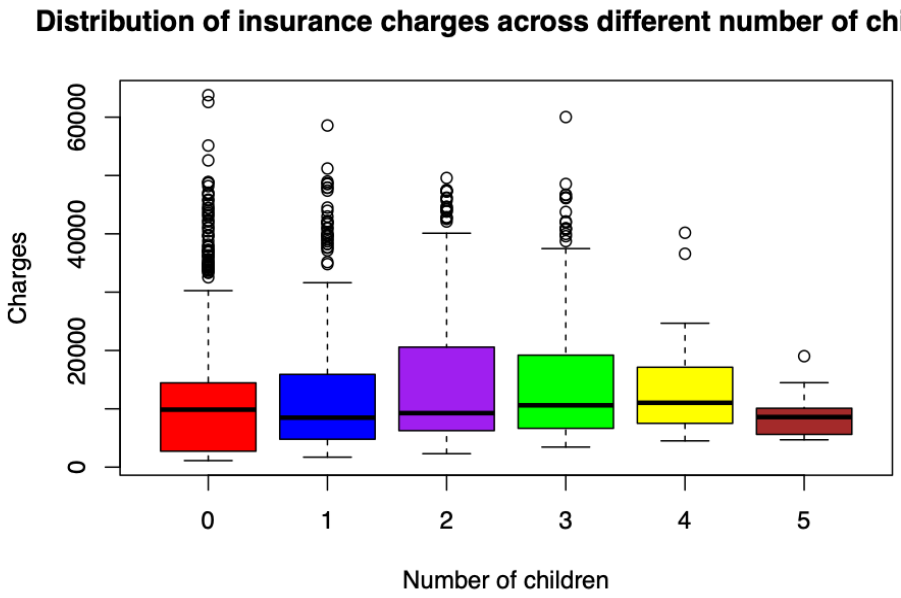


As can be seen from the boxplot, male and female beneficiaries pay similar insurance charges. This is contrary to what we expected to find. One potential reason why this may be the case is because Americans tend to

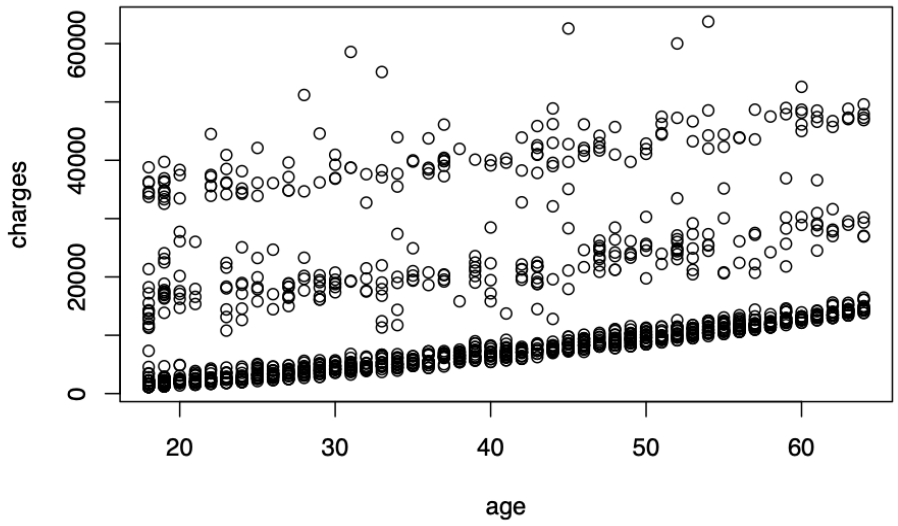
invest in family health plans and not individual health plans so it may be hard to determine differences in gender.



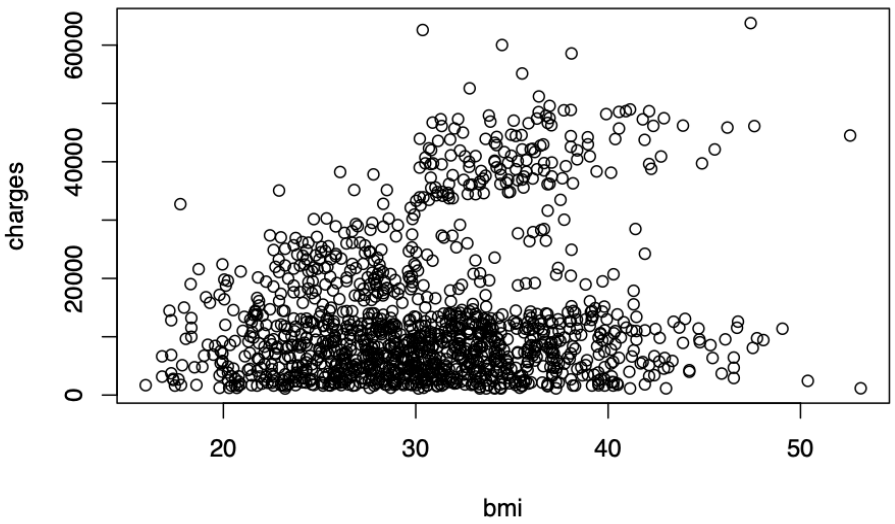
As can be seen from the boxplot, those living in the southeastern region pay higher insurance charges compared to other regions. States in this region include Florida, North Carolina, South Carolina among others. Health insurance may be more expensive in this region because of the vulnerability of this region to environmental disasters (flooding etc). People living in the Northeastern region have the next higher costs of insurance charges after the southeastern regions. States like New York, Massachusetts are included in this region. The reason for higher insurance charges in this region may be because of higher living costs.



We expected that the more children one has, the higher the insurance cost. As can be seen from the boxplot, insurance costs do not increase with an increase in the number of children. It was reported that in the year 2016, Americans on an average had 2.06 kids (Luscombe, 2018). This makes the demand for a plan for a family with 2 to 3 kids higher and therefore more expensive.



As can be seen from the scatterplot, there is a positive relationship between the insurance charges and age. This is consistent with what we had thought as older people are more vulnerable to suffering from diseases and health problems compared to younger people.



As can be seen from the scatterplot, there is a somewhat positive relationship between the Insurance charges and BMI. This is consistent with what we had thought as people with higher BMI are more vulnerable to suffering from health problems as compared to those with lower BMI.

Hypotheses

We are testing the significance of each predictor in the model.

Our hypotheses are given by:

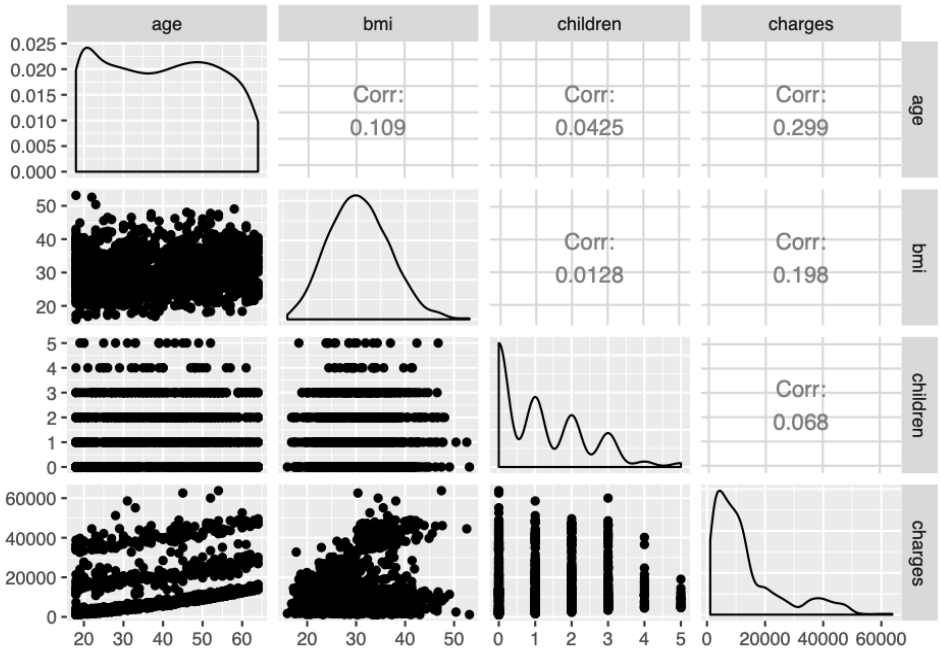
$H_0 : \beta_i = 0$ and $H_a : \beta_i \neq 0$

The model that we are testing is given below:

$\widehat{charges} = \hat{\beta}_1.smokerYes + \hat{\beta}_2.sexmale + \hat{\beta}_3.age + \hat{\beta}_4.bmi + \hat{\beta}_5.regionnorthwest + \hat{\beta}_6.regionsoutheast + \hat{\beta}_7.regionsouthwest + \hat{\beta}_8.children$

Methods

Multicollinearity




```
##              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
```

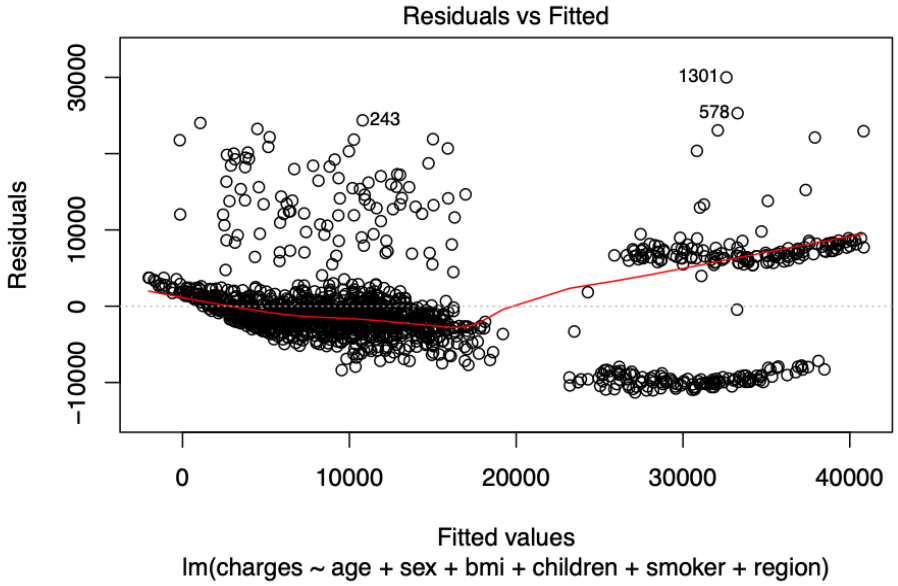
To ensure that there are no issues of multicollinearity in our model, we run diagnostic tests to find the correlation coefficients between the numerical variables and the VIF values for the predictor variables. From the output above, it can be seen that the correlation coefficients are very small and that all of the VIF values are less than 5. We can say that there is no cause to be concerned about multicollinearity when including all predictor variables in the model.

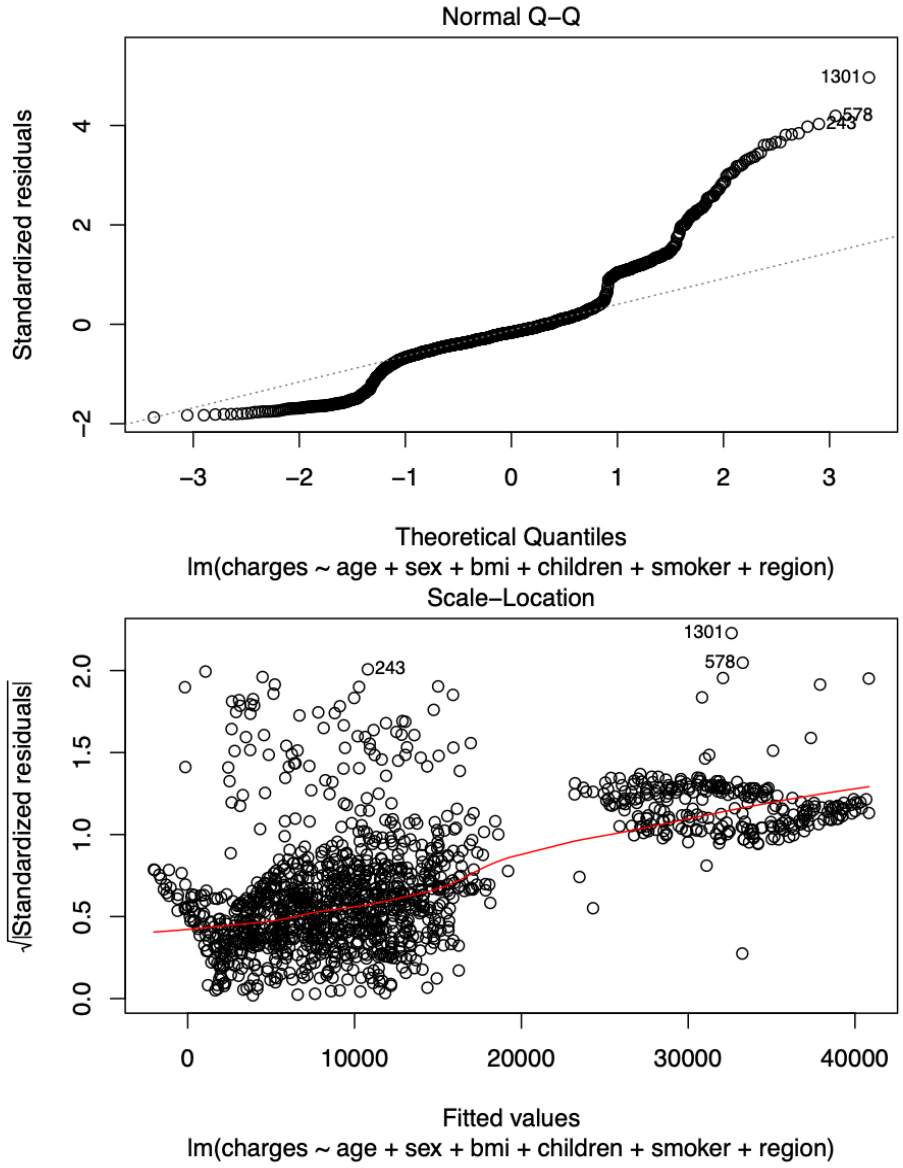
Inital Model

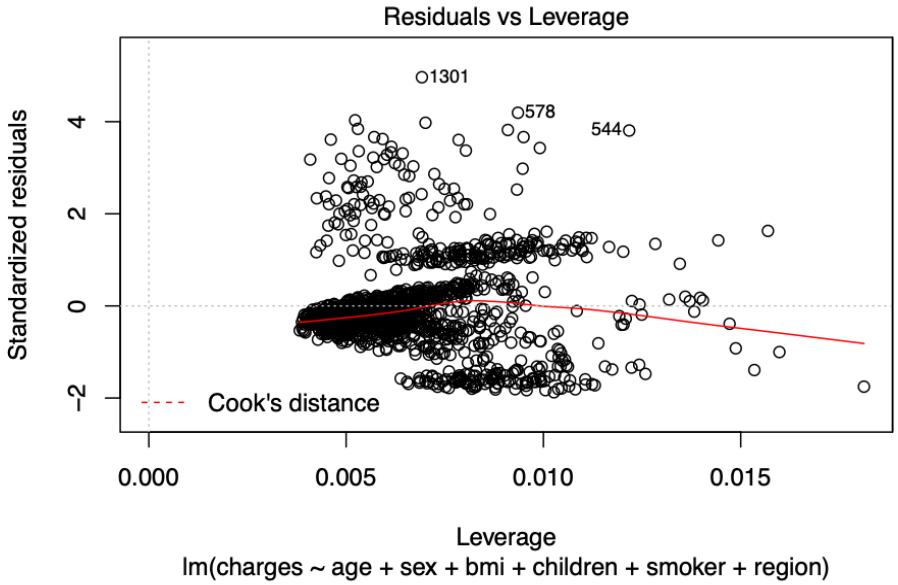
```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker +
##     region, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Our initial model is the full model (i.e. all predictor variables are included). As can be seen from the output, the variable sex and the variable for region northwest are insignificant in predicting the insurance charge as their p-values are greater than 0.05. We then went on to check whether or not the LINE conditions were satisfied and if we could move forward with the fitted model.



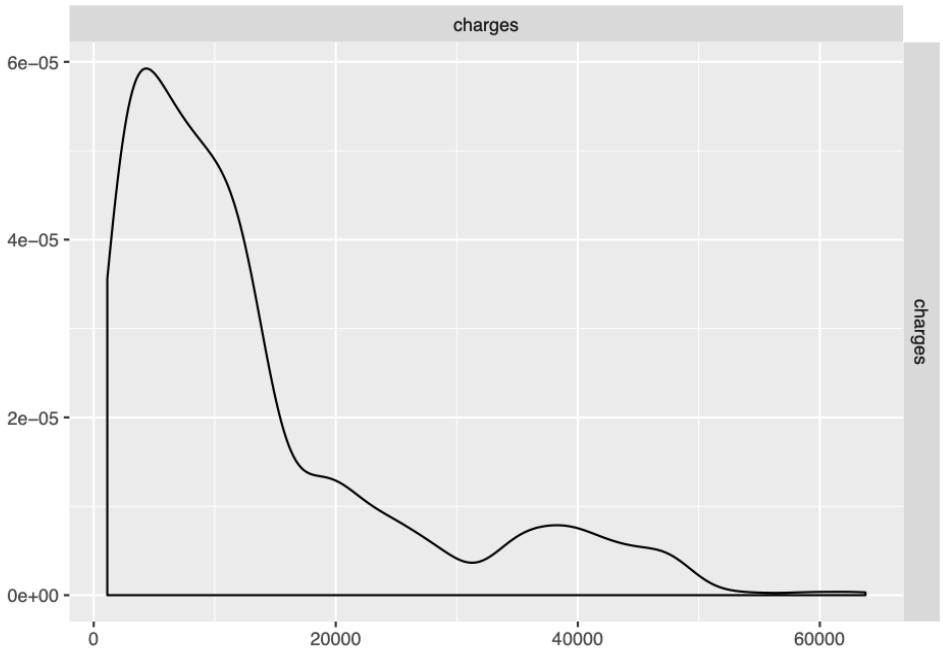

Normal Q–Q
Standardized residuals
1301
578
243
Theoretical Quantiles
lm(charges ~ age + sex + bmi + children + smoker + region)
Scale–Location
√|Standardized residuals|
1301
578
243
Fitted values
lm(charges ~ age + sex + bmi + children + smoker + region)



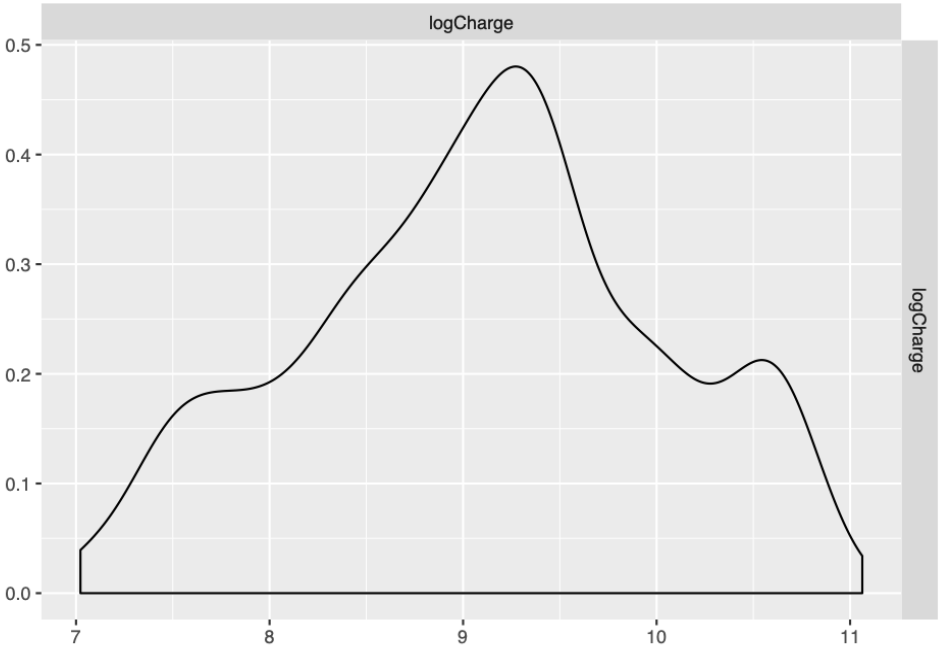
As can be seen from the output above, the LINE conditions for our initial model are not satisfied. The observations are not following a straight line in the Normal Q-Q plot violating the linearity and normality condition. The observations are also unevenly distributed in the RVF plot and not centered around 0 which violates the equal variance condition.We notice 3 residual clusters in the RVF plot. Independence is assumed in this situation, as we are not aware of how the sample was collected. However, as the linearity, normality and equal variance conditions are so grossly violated, we choose not to move forward with this model.

Transformation

Given that our initial model is not suitable for predicting insurance charges, we then transform the variables in our model.

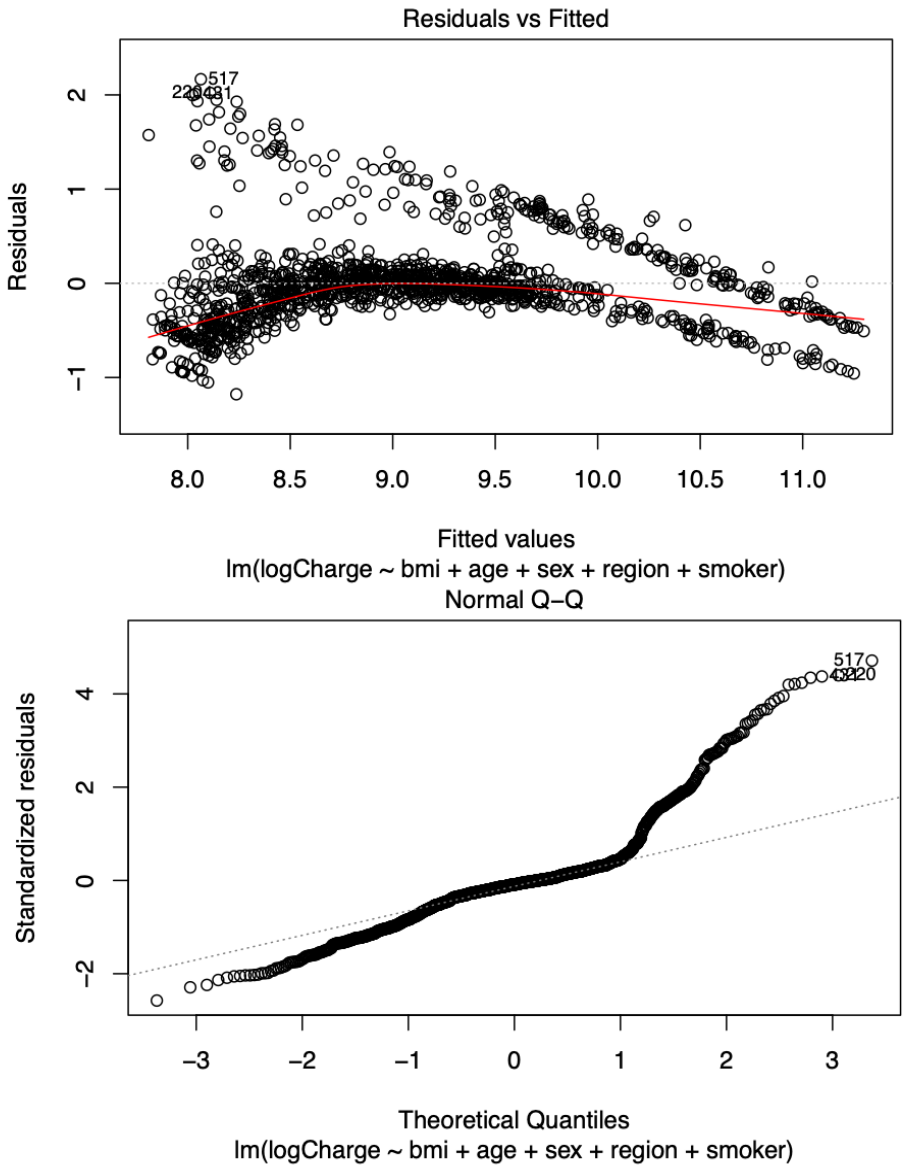


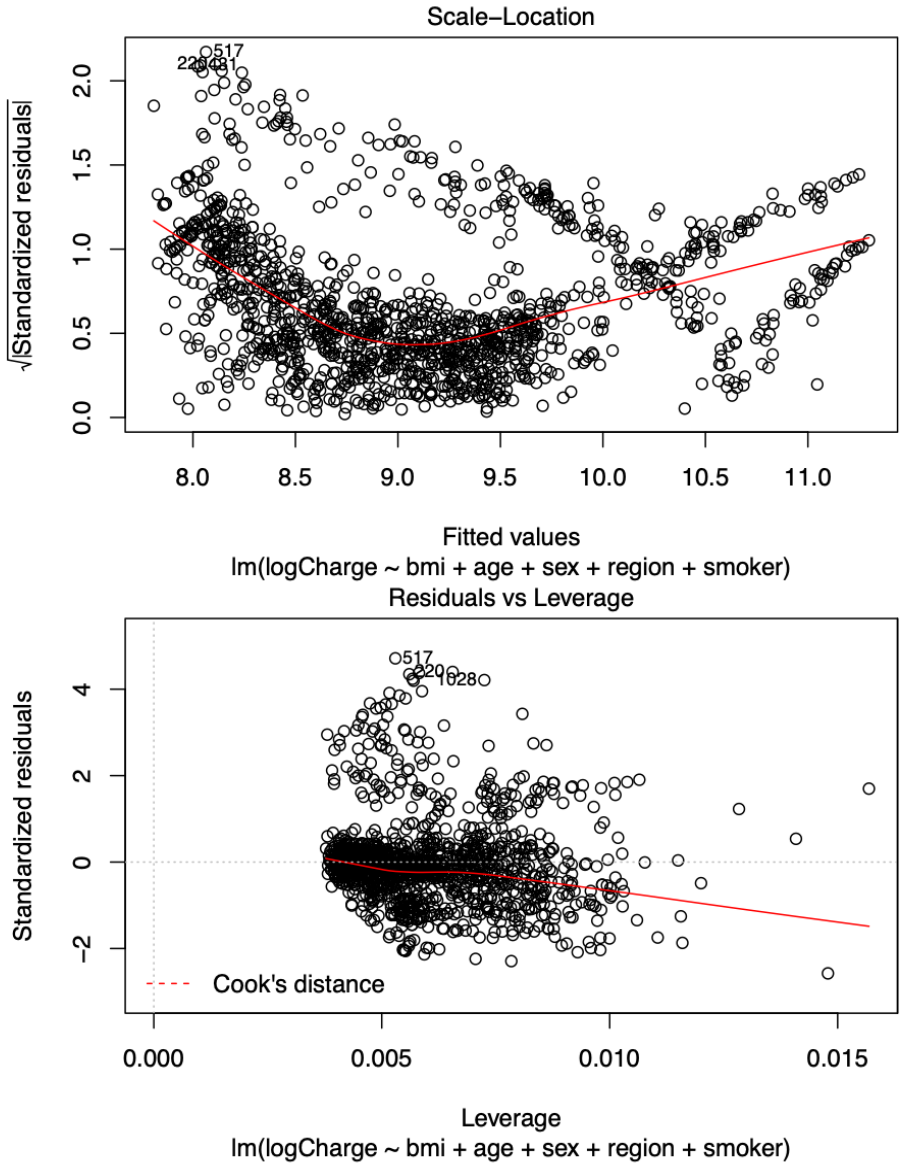
Looking back at the diagnostics for correlation, we notice that our response variable, insurance charges, Is right skewed. We log the variable to transform the distribution to a normal, bell shaped curve.



As can be seen from the output above, we observe our expected change in the distribution. Next, we fit a model with the variable *logCharge* (logged charges) instead of the original variable,*charges*.

```
##
## Call:
## lm(formula = logCharge ~ bmi + age + sex + region + smoker, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17783 -0.22056 -0.04044  0.10483  2.16636
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.1122965  0.0746151  95.320  < 2e-16 ***
## bmi              0.0136460  0.0021738   6.278 4.64e-10 ***
## age              0.0349422  0.0009037  38.665  < 2e-16 ***
## sexmale         -0.0711870  0.0253045  -2.813  0.00498 **
## regionnorthwest -0.0533212  0.0361872  -1.473  0.14086
## regionsoutheast -0.1580922  0.0363867  -4.345 1.50e-05 ***
## regionsouthwest -0.1196106  0.0363164  -3.294  0.00102 **
## smokeryes        1.5574014  0.0314034  49.593  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4608 on 1330 degrees of freedom
## Multiple R-squared:  0.7502, Adjusted R-squared:  0.7489
## F-statistic: 570.6 on 7 and 1330 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
Fitted values
lm(logCharge ~ bmi + age + sex + region + smoker)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(logCharge ~ bmi + age + sex + region + smoker)



As can be seen from the summary output above, the variable sex is significant under the new model using logged charges (with a p-value less than 0.05). We then check to see if the LINE conditions are satisfied under this new model.

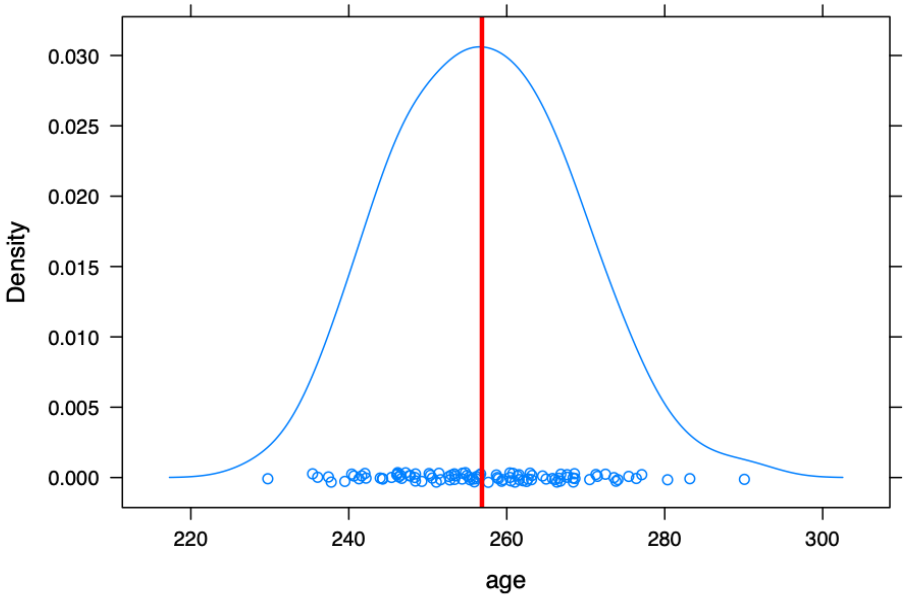
The LINE conditions for our new model are also not satisfied. The observations are not following a straight line in the normal Q-Q plot violating the linearity and normality condition. The observations are unevenly distributed in the RVF plot and not centered around the 0 line which violates the equal variance condition.

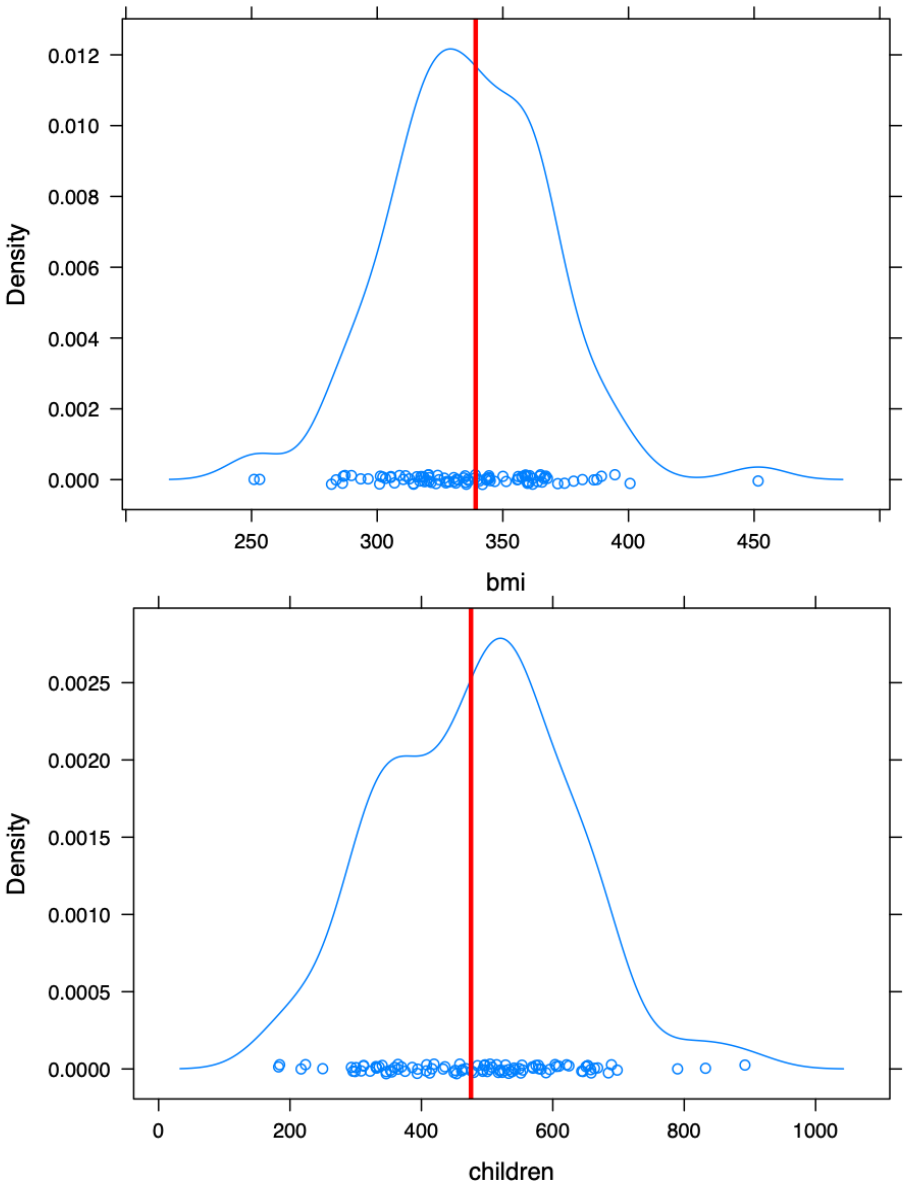
With the new model, we do not come across distinct residual clusters as we had with the initial model. Independence is, again, assumed in this situation, as we are not aware of how the sample was collected. However, as the linearity, normality and equal variance conditions are so grossly violated, we choose to disregard this model as well.

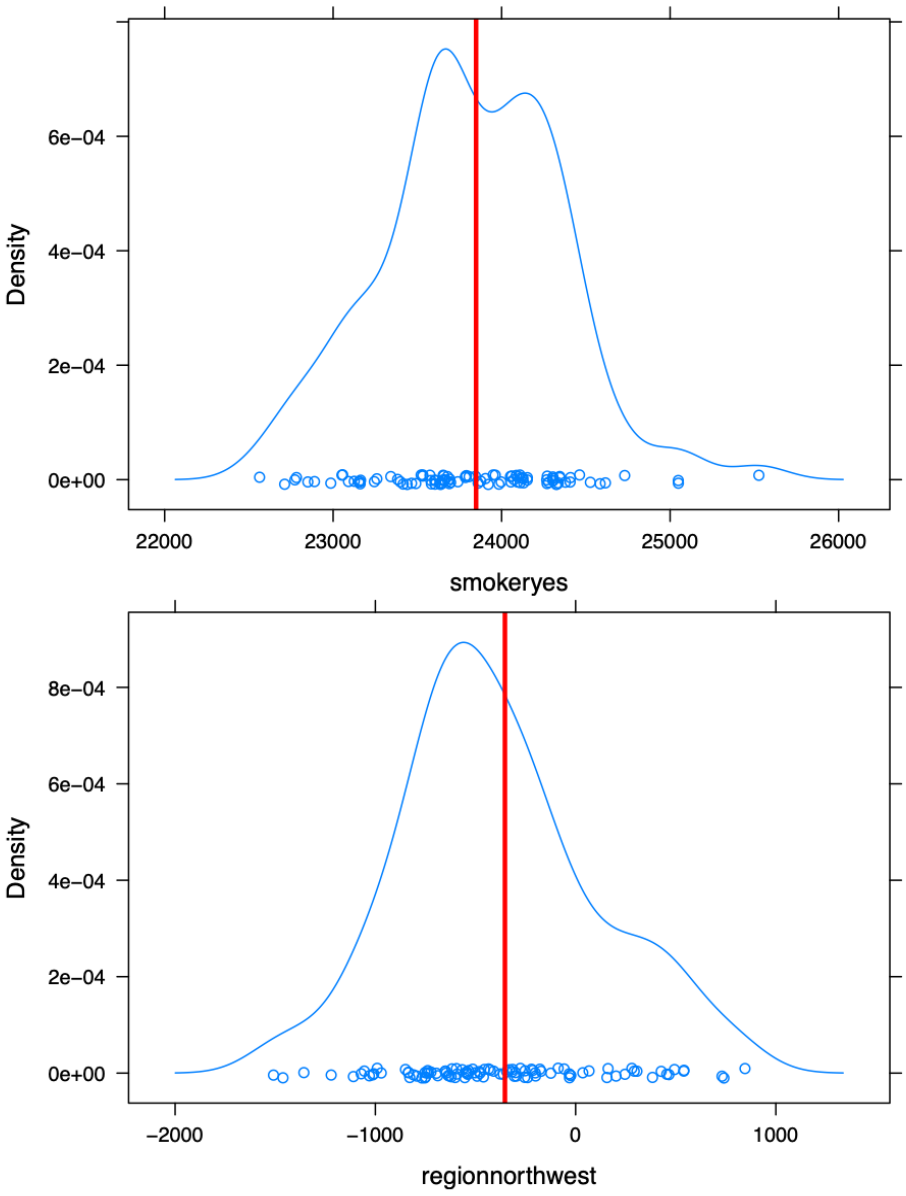
It is interesting to note that in the RVF plots of both generated models, is that there is something driving this data into separate clusters (3 for the initial model and 2 for the new model). We thought that the may be driven by the variables sex or smoker. We go ahead and test different models by transforming our children variable by taking the log of (children+1), but this transformation also fails to satisfy the LINE conditions. We choose to use the bootstrap method. This will be discussed in the results section of this paper.

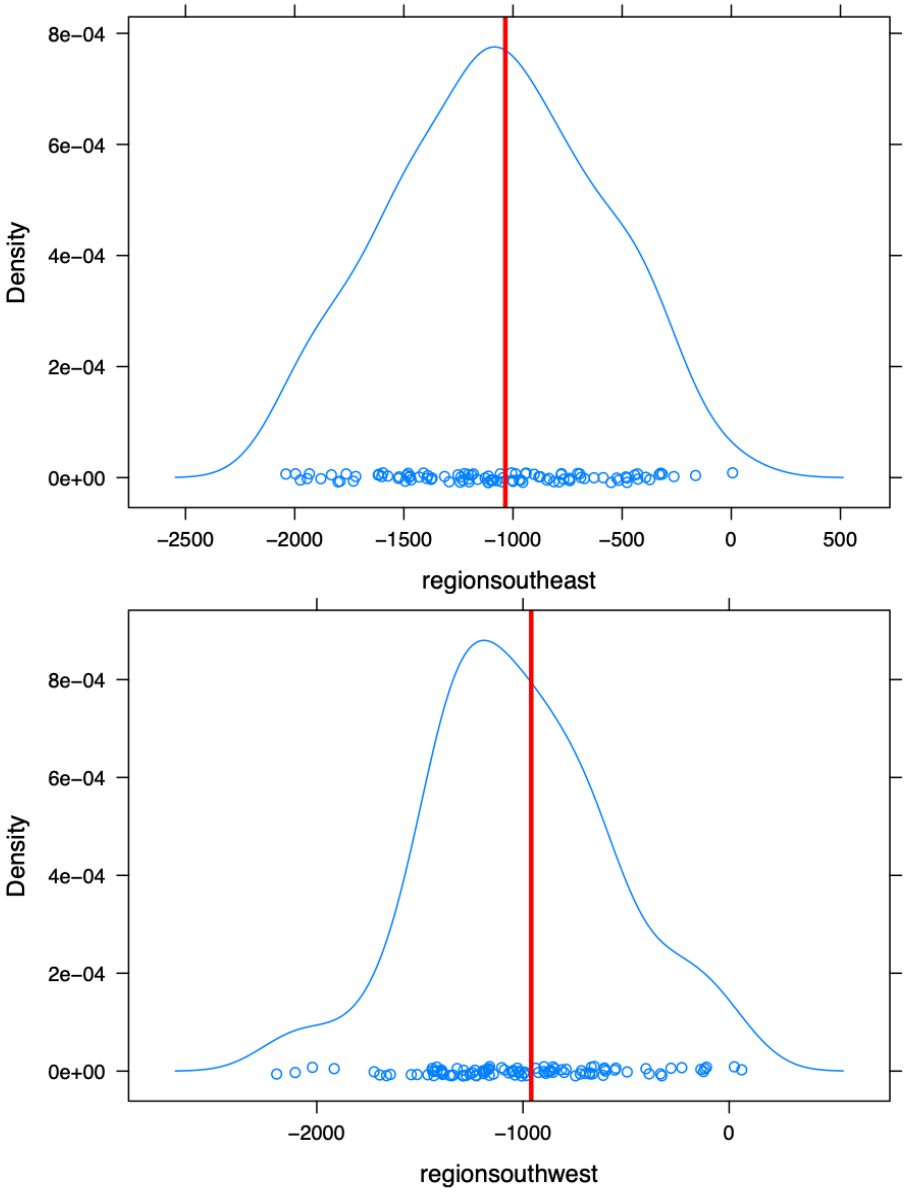
Results

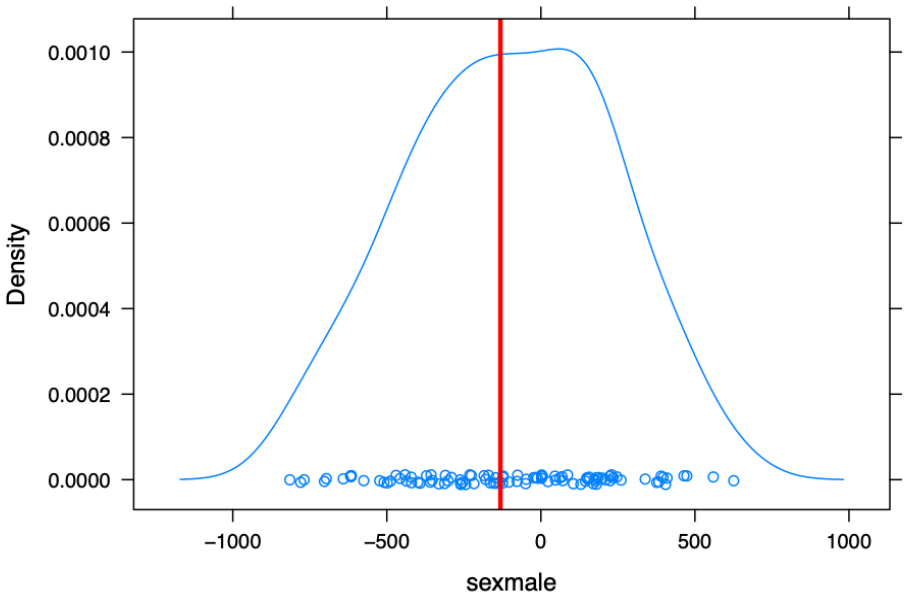
Since the normality condition necessary for inference of multiple linear regression is not met, we perform bootstrap modeling.




Density
0.012
0.010
0.008
0.006
0.004
0.002
0.000
250
300
350
400
450
bmi
Density
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
200
400
600
800
1000
children


Density
6e−04
4e−04
2e−04
0e+00
22000
23000
24000
25000
26000
smokeryes
Density
8e−04
6e−04
4e−04
2e−04
0e+00
−2000
−1000
0
1000
regionnorthwest


Density
8e−04
6e−04
4e−04
2e−04
0e+00
−2500
−2000
−1500
−1000
−500
0
500
regionsoutheast
Density
8e−04
6e−04
4e−04
2e−04
0e+00
−2000
−1000
0
regionsouthwest



The plots above are the density plots of the bootstrap statistics of smokeryes, age, bmi, children, regionsouthwest, regionnorthwest, regionsoutheast, and sexmale. Regionnorthest and sexfemale are reference categories for region and sex respectively. Because these density plots show that the bootstrap distribution for each of the variables is normal, the 95% confidence intervals are based on the standard deviation of the bootstrap statistics.

```
z_star <- qnorm(c(0.025, 0.975)) # 95% CI

slope <- coef(model)["smokeryes"]
slope + z_star * sd(~smokeryes, data=insurance_bootstrap)
```

```
## [1] 22798.05 24899.02
```

```
slope1 <- coef(model)["sexmale"]
slope1 + z_star * sd(~sexmale, data=insurance_bootstrap)
```

```
## [1] -780.5870  517.9583
```

```
slope2 <- coef(model)["age"]
slope2 + z_star * sd(~age, data=insurance_bootstrap)
```

```
## [1] 234.1587 279.5540
```

```
slope3 <- coef(model)["bmi"]
slope3 + z_star * sd(~bmi, data=insurance_bootstrap)
```

```
## [1] 277.2010 401.1859
```

```
slope4 <- coef(model)["regionnorthwest"]
slope4 + z_star * sd(~regionnorthwest, data=insurance_bootstrap)
```

```
## [1] -1318.3896   612.4618
```

```r
slope5 <- coef(model)["regionsoutheast"]
slope5 + z_star * sd(~regionsoutheast, data=insurance_bootstrap)
```

```
## [1] -1958.3222  -111.7219
```

```r
slope6 <- coef(model)["regionsouthwest"]
slope6 + z_star * sd(~regionsouthwest, data=insurance_bootstrap)
```

```
## [1] -1859.23921   -60.86277
```

```r
slope7 <- coef(model)["children"]
slope7 + z_star * sd(~children, data=insurance_bootstrap)
```

```
## [1] 202.3221 748.6789
```

None of the generated bootstrap confidence intervals include 0 except sexmale and regionnorthwest . We reject the null hypothesis of all other predictors except for the null hypothesis of regionnorthwest and sexmale. That being said, we have enough statistical evidence allowing us to drop the sexmale and regionnorthwest variables as they are not significantly associated with insurance charges. However, we keep region as one of the explanatory variables in our model since the other region categories are significant predictors based on our 95% confidence interval generated from bootstrap. We want to keep the wholeness of the predictor,region, in our model. Therefore, we find that age, bmi, children, smoker, and region are statistically significant predictors in our final model.

Final Model

The final model is represented by the following equation:

$\widehat{charges} = -11990.27 + 23836.30 * smoker + 256.97 * age + 338.66 * bmi - 352.18 * regionnorthwest - 1034.36 * regionsoutheast - 959.37 * regionsouthwest + 474.57 * children$

The summary output of our final model is shown below:

```
##
## Call:
## lm(formula = charges ~ bmi + age + region + smoker, data = insurance)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11905  -3041  -1000   1542  29419
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11601.56     976.20 -11.884   <2e-16 ***
## bmi                340.01      28.67  11.858   <2e-16 ***
## age                258.64      11.93  21.680   <2e-16 ***
## regionnorthwest   -303.52     477.85  -0.635   0.5254
## regionsoutheast  -1038.63     480.49  -2.162   0.0308 *
## regionsouthwest   -915.94     479.56  -1.910   0.0564 .
## smokeryes        23852.48     413.51  57.683   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6085 on 1331 degrees of freedom
## Multiple R-squared:  0.7487, Adjusted R-squared:  0.7475
## F-statistic: 660.8 on 6 and 1331 DF,  p-value: < 2.2e-16
```

The summary table above showed that except for regionnorthwest, all of other variables have a p-value smaller than the significance level of 0.05, which means that our data provides enough evidence that these variables are statistically significant. The R^2 value is 0.7509, which means that our model explains 75.09% of the variability in annual individual medical insurance.

Discussion and Conclusion

The aim of this project is to determine which factors would be best at determining an individual's health insurance cover fee. Our findings show that age, BMI, region,number of children and whether an individual smoked or not are significantly associated with insurance charges. Our model can be interpreted as follows : holding all other variables constant, for every extra child an individual has,one would expect the insurance charges to be $474.57 higher. Keeping all other variables constant, for every 1 year increase in age,one would expect the insurance charges to be $256.97 higher. A smoker is expected to have an insurance charge that is $23,836.30 higher than a non-smoker with the condition that all other variables are held constant. Holding all other variables constant, an individual from the northwest region, southeast region, and southwest region of the US is expected to have an insurance charge $352.18 less,$1034.36 less and $959.37 less than that of an individual in the northeast region respectively. As the coefficients depict, Individuals in the southeast region would have to pay the lowest insurance charge.

Our final model may be generalized to the entire US population.Furthermore, our model could be improved with the addition of other predictors that are known to also affect insurance premium charge such as income and ethnicity.

References

Frakt, A., Carroll, A. (2018, January 2). Why the U.S. Spends So Much More Than Other Nations on Health Care. Retrieved from: https://www.nytimes.com/2018/01/02/upshot/us-health-care-expensive-country-comparison.html

Abutabel, Y. (2018, February 14). U.S. healthcare spending to climb 5.3 percent in 2018: agency. Retrieved from: https://www.reuters.com/article/us-usa-healthcare-spending/u-s-healthcare-spending-to-climb-5-3-percent-in-2018-age

Botkin, K. (n.d.). 10 Factors That Affect Your Health Insurance Premium Costs. Retrieved from: https://www.moneycrashers.com/factors-health-insurance-premium-costs/

Boulton, G. (2018, June 07). You'll be shocked at how much health insurance costs for a family of four.Retrieved from: https://www.usatoday.com/story/money/business/2018/06/06/health-care-costs-price-family-four/676046002/

Income.com (2017, October 4). Tips for trimming your health insurance premiums. Retrieved from: https://www.insure.com/health-insurance/savings.html

Luscombe, B. (2018, January 19). More American Women Are Having Children. Retrieved from: http://time.com/5107704/more-women-mothers/

Choi, M. (2018, February). Medical Cost Personal Datasets: Insurance Forecast by using Linear Regression. Retrieved from: https://www.kaggle.com/mirichoi0218/insurance