project

Winnie Lu

2022-03-21

Exploratory Data

- Briefly examining the data shows that several dummy variables that may influence the insurance costs (Sex, smoker, and region). We also see the basic summary statistics for all the predictors in our data, and note that besides the mean being greater than the median, there are no missing values across the predictors. To reflect on our summary statistics, our distribution is incredibly right skewed, even though a majority of the annual insurance expenses range from $0 to $15,000.

```
insurance <- read.csv("insurance.csv")
head(insurance)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

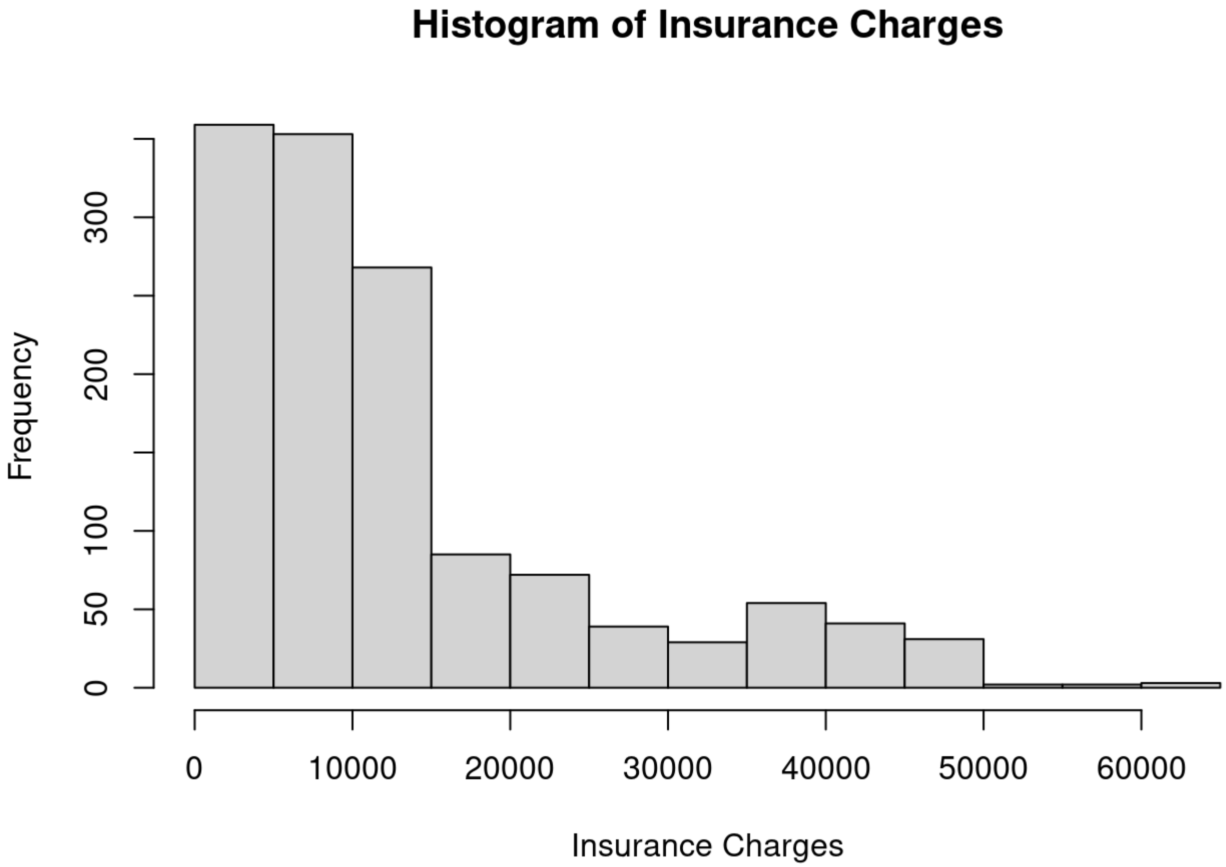
```
summary(insurance)
```

```
##       age            sex                 bmi           children    
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000  
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000  
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000  
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095  
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000  
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000  
##     smoker             region             charges     
##  Length:1338        Length:1338        Min.   : 1122  
##  Class :character   Class :character   1st Qu.: 4740  
##  Mode  :character   Mode  :character   Median : 9382  
##                                        Mean   :13270  
##                                        3rd Qu.:16640  
##                                        Max.   :63770
```

```
sum(is.na(insurance))
```

```
## [1] 0
```

```
hist(insurance$charges, main = "Histogram of Insurance Charges", sub = "Figure 1", xlab
 = "Insurance Charges", ylab="Frequency")
```



###

Figure 1

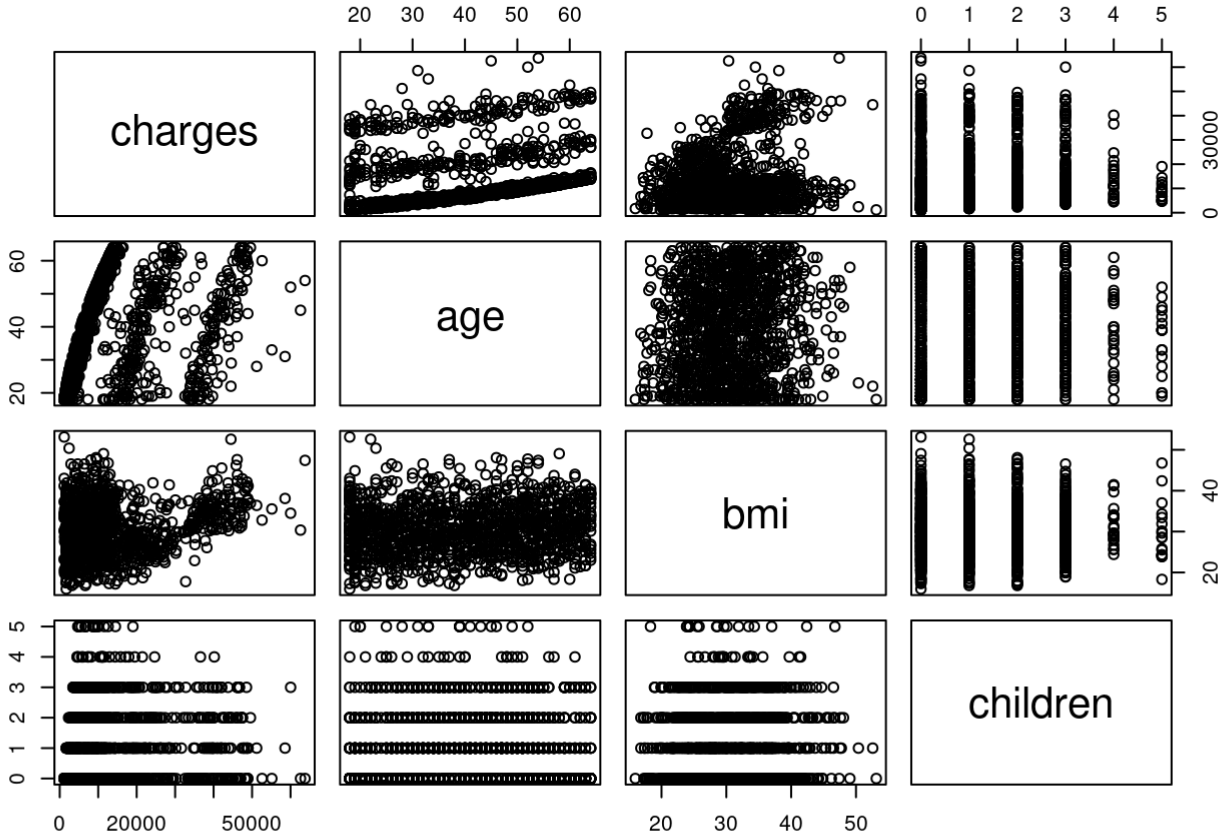
Visualization - We start by visualizing our data to identify any obvious pattern between the predictors and response, or among the predictors themselves. Since sex, smoker and region are categorical variables, we will examine them after. - Preliminary analysis in Fig 1A. show that age and bmi have a slight positive relationship with charges. As age and bmi increases, we see a general increase in charges. These are observations we would expect: generally, younger people are healthier than the elderly. A healthy individual has a BMI between 18.5 to 24.9, whereas a BMI over 29.9 is considered to be within the obese range, so it makes sense to see the pattern we see. (SOURCE). However, it is relevant to note that people with lower BMI still have high medical expenses. An interesting observation here is that with an increasing amount of children per household, we see that charges appear to decrease. This finding contradicts what we would expect from insurance companies; as the common notion appears to be that having more children would compromise the family into buying more premiums or policies for them. It is notable to even say that with more children, there would be a greater chance of children inheriting or developing a poorer health condition than the older siblings. We will explore further possible reasons behind this later.(source here). Our correlation matrix supports these observations. - There appears to be no significant relationship difference between the sex of the policy holder and charges. In contrast, smokers tend to pay more for their policies compared to non-smokers. Evidently, smokers are more likely to be associated with various health risks; therefore, paying more for certain premiums or paying more in terms of frequency in physician visits. Such factors include, but not limited to the umbrella of heart diseases: myocardial infarctions, atherosclerosis, or respiratory distress: emphysema, bronchitis, etc, and even dental costs! (Source) - Lastly, there does not seem to be any co linearity among the predictors.

```
pairs(charges ~ age + bmi + children, data=insurance)

round(cor(insurance[, -c(2, 5, 6)]), 4)
```

```
##             age    bmi children charges
## age      1.0000 0.1093   0.0425  0.2990
## bmi      0.1093 1.0000   0.0128  0.1983
## children 0.0425 0.0128   1.0000  0.0680
## charges  0.2990 0.1983   0.0680  1.0000
```

```
library(ggplot2)
```

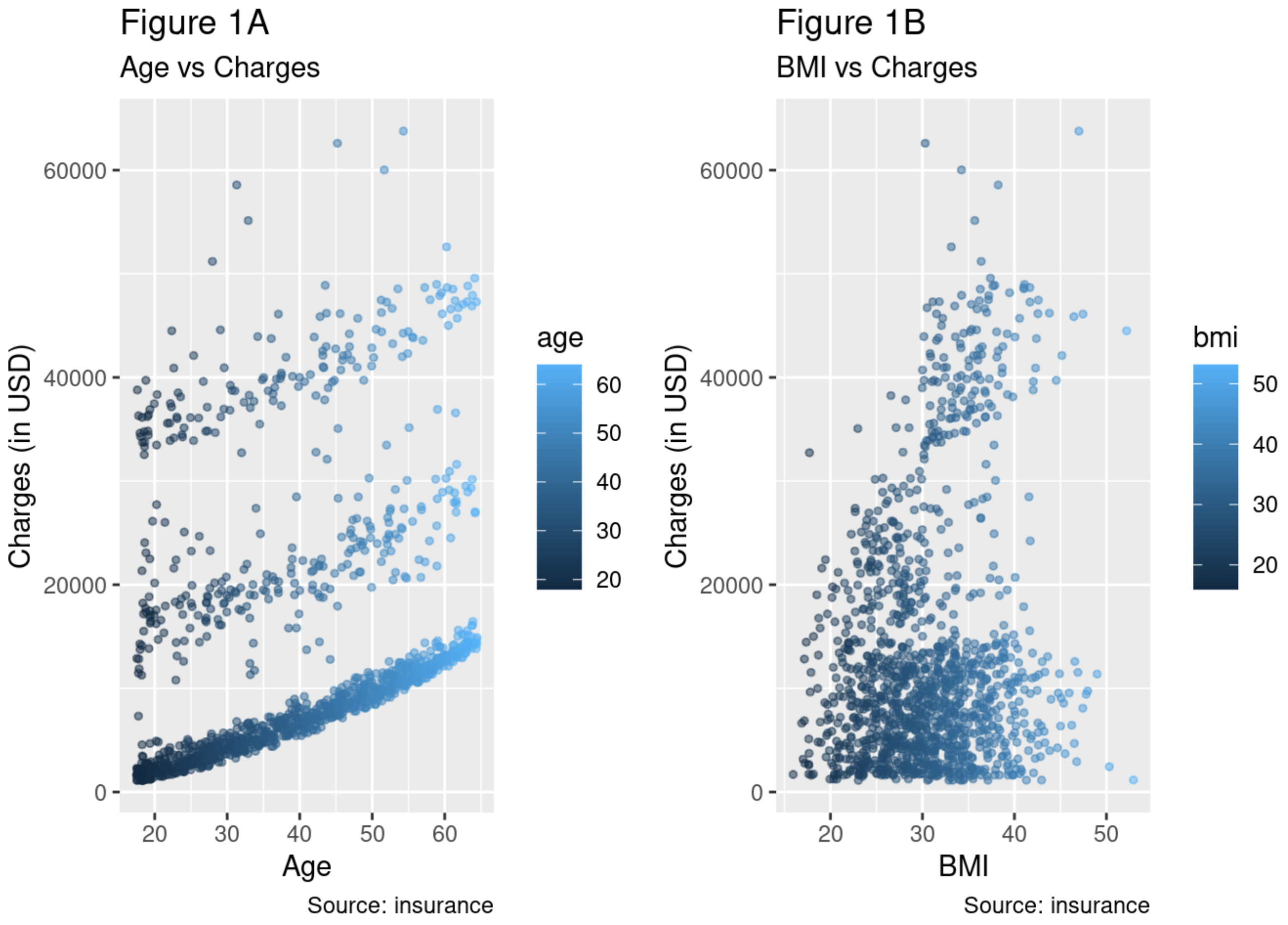


```r
library(gridExtra)

a <- ggplot(insurance, aes(age, charges)) +
  geom_jitter(aes(color = age), alpha = 0.5, size = 1, width = .5) +
  labs(subtitle="Age vs Charges",
       y="Charges (in USD)",
       x="Age",
       title="Figure 1A",
       caption = "Source: insurance")


b <- ggplot(insurance, aes(bmi, charges))+
    geom_jitter(aes(color = bmi), alpha = 0.5, size = 1, width = .5) +
    labs(subtitle="BMI vs Charges",
       y="Charges (in USD)",
       x="BMI",
       title="Figure 1B",
       caption = "Source: insurance")

grid.arrange(a, b, ncol = 2)
```



```
s <- ggplot(insurance, aes(sex, charges))+
    geom_jitter(aes(color = sex), alpha = 0.5, size = 1, width = .5) +
    labs(subtitle="Sex vs Charges",
       y="Charges (in USD)",
       x="Sex",
       title="Figure 1C",
       caption = "Source: insurance")

c <- ggplot(insurance, aes(children, charges))+
    geom_jitter(aes(color = children), alpha = 0.5, size = 1, width = .5) +
    labs(subtitle="Number of Children vs Charges",
       y="Charges (in USD)",
       x="Number of Children",
       title="Figure 1D",
       caption = "Source: insurance")


grid.arrange(s, c, ncol = 2)
```

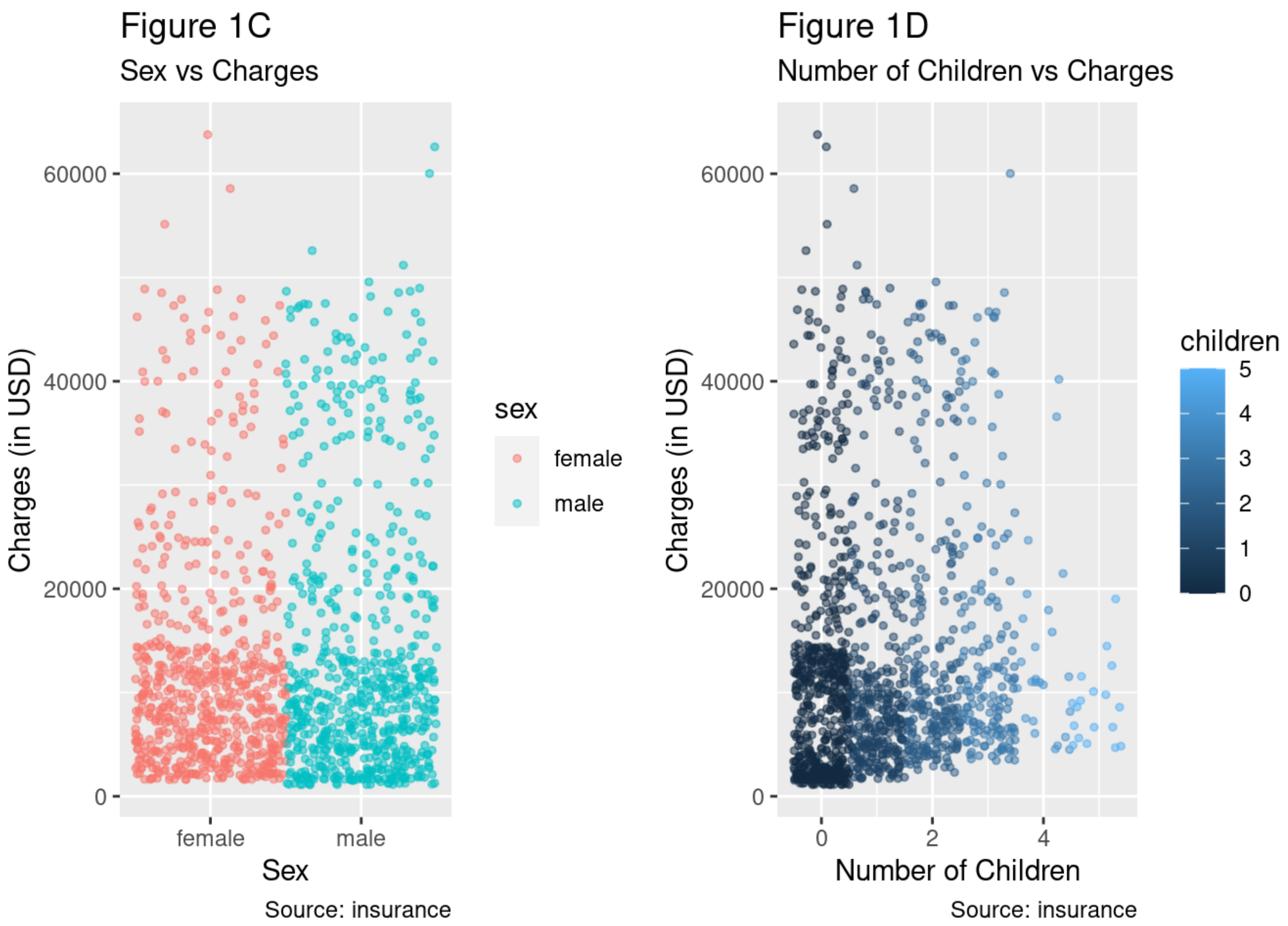


Figure 1C

Sex vs Charges

Source: insurance

Figure 1D

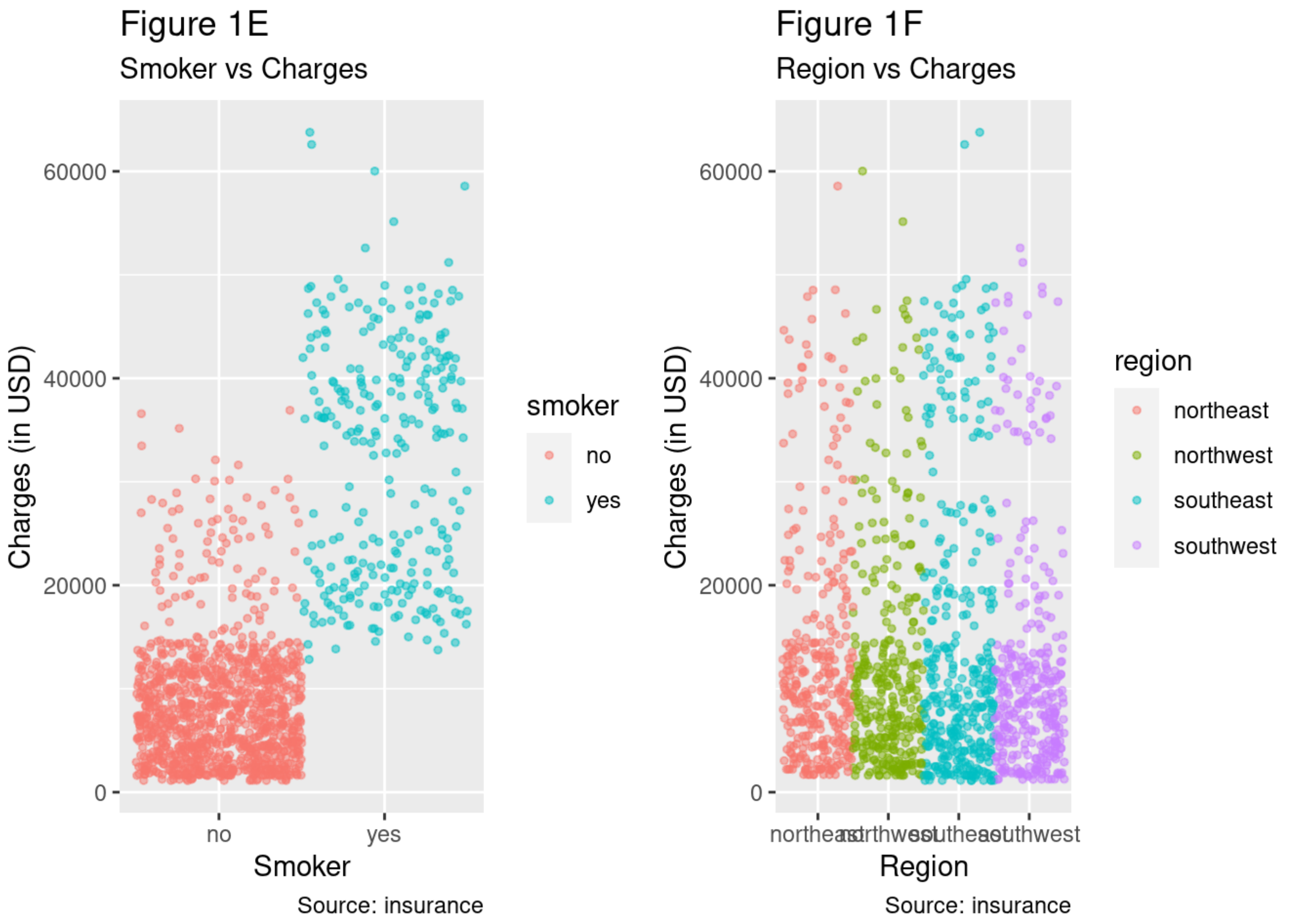
Number of Children vs Charges

Source: insurance

```
sm <-ggplot(insurance, aes(smoker, charges))+
    geom_jitter(aes(color = smoker), alpha = 0.5, size = 1, width = .5) +
    labs(subtitle="Smoker vs Charges",
       y="Charges (in USD)",
       x="Smoker",
       title="Figure 1E",
       caption = "Source: insurance")

r <-ggplot(insurance, aes(region, charges))+
    geom_jitter(aes(color = region), alpha = 0.5, size = 1, width = .5) +
    labs(subtitle="Region vs Charges",
       y="Charges (in USD)",
       x="Region",
       title="Figure 1F",
       caption = "Source: insurance")

grid.arrange(sm, r, ncol = 2)
```



First Model

- In our first model, the F-test (p-value: < 2.2e-16) tells us that at least one predictor is not equal to another. So, at least one predictor is significant to our model. By examining the individual t-tests, the predictor, sex does not appear to be statistically significant. However, we see that age and smokers appear to greatly influence the charges. Since there are dummy variables present, one category of the variable is used as a reference. As a result, the estimates are interpreted relative to the reference. Thus, males have $\$131.30$ less medical expenses each year relative to females and smokers cost an average of $\$23,848.50$ more than non-smokers per year. We also see that the coefficient for the reference group, northeast region, tends to have the highest average charges compared to the other 3 regions in our model. The $R^{2}_{adj} = 0.7494 $, which indicates that about 74.94% of the variability we see in insurance charges are explained by the predictors in our current model. Since this value is a reflection on our model, we can deem it not too bad. To conclude, we see high residuals, meaning high errors: over 50% of the errors fall within the first and third quartile: predictions were between $2848.10 over the true value and 41393.9 under the true value.

- Obviously, this model isn't the best, seeing as how it violates our LINE assumptions: we see a discernible pattern in figure 2A and strong deviation of points from the line in figure 2B; violating our equal variance and normality assumptions.

```
model1 <-lm(charges ~ age + sex + bmi + children + smoker + region, data=insurance)
summary(model1)
```

```
##
## 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
```

```
par(mfrow=c(1,3))
plot(model1, 1) + title( sub ="Figure 2A", col.sub = "blue")
```

```
## integer(0)
```

```
plot(predict(model1), rstandard(model1), xlab="Predicted", ylab="Std. Residuals") + title(main = "Std. Residuals vs Fitted", sub = "Figure 2B", col.sub="blue")
```

```
## integer(0)
```

```
plot(model1, 2) + title(sub ="Figure 2C", col.sub = "blue")
```

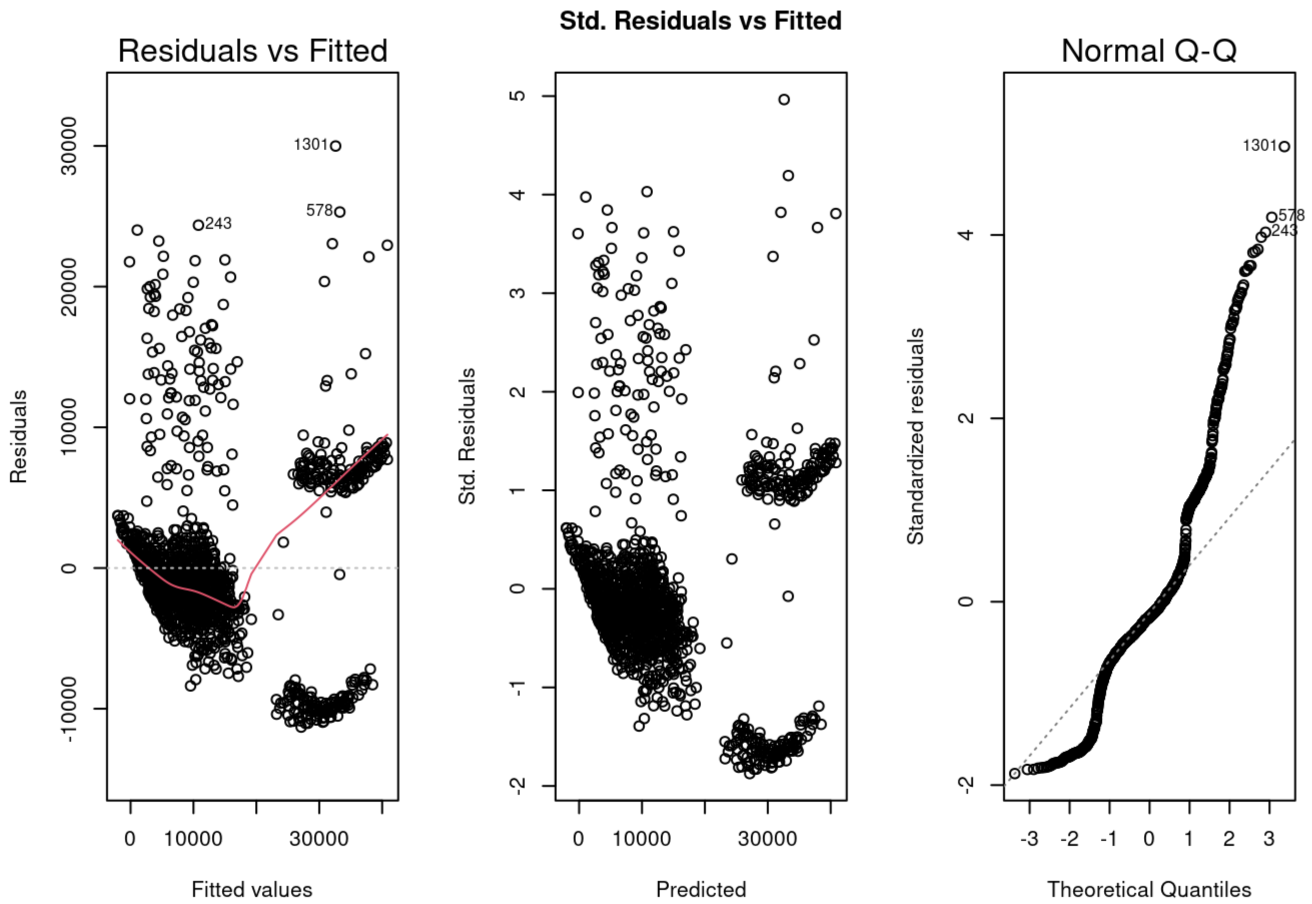


Figure 2A Figure 2B Figure 2C

```
## integer(0)
```

Second Model

- Although there are clear violations to our equal variance and normality assumptions, we will not be using the box-cox method to remedy this. Currently, our goal is to identify the major influence(s) to our response, so our focus will be on the effects of charges via predictors. For the second model, we've dropped the predictors Region and Sex, since they appeared to be statistically insignificant. Given our F-statistic, at least one of the our predictors is significant to our model, and our individual t-test show that all predictors in our second model are significant. Among our current set of predictors, smoking and obesity are the predominant factors influencing the insurance charges. According to the CDC, obese patients have a BMI of over 30.(source) We can include this into our model to show the effects of obesity on insurance charges. We can see that if a patient is a smoker, then the medical costs increase by 23819.41 dollars. Additionally, if a patient is obese, then they have increased medical cost of 2904.47 dollars. The $R^2_{adj} = 0.754$ shows an improvement compared to our first model, where $R^2_{adj} = 0.7494$. However, we can see that our linear regression model is not the best model, because of the violations of normality and heteroscedascity present (Fig3C). We require a different approach(es) for a better prediction. The clusters present in Figure 3A

```
insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)
model2 <-lm(charges ~ age  + bmi + bmi30 + children + smoker, data=insurance)
summary(model2)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + bmi30 + children + smoker,
##     data = insurance)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12530  -3503   -232   1494  28075
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7816.28    1233.73  -6.335 3.23e-10 ***
## age           258.03      11.78  21.909  < 2e-16 ***
## bmi           131.71      44.93   2.932 0.003428 **
## bmi30        2904.47     547.33   5.307 1.31e-07 ***
## children      473.87     136.41   3.474 0.000529 ***
## smokeryes   23819.41     407.10  58.511  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6007 on 1332 degrees of freedom
## Multiple R-squared:  0.7549, Adjusted R-squared:  0.754
## F-statistic: 820.4 on 5 and 1332 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(1,3))
plot(model2, 1) + title(sub = "Figure 3A", col.sub = "blue")
```

```
## integer(0)
```

```
plot(model2, 2) + title(sub = "Figure 3B", col.sub = "blue")
```

```
## integer(0)
```

```
plot(resid(model2) ~ charges, data=insurance) + title(main = "Residual vs Charges", sub=
"Figure3C", col.sub="blue")
```

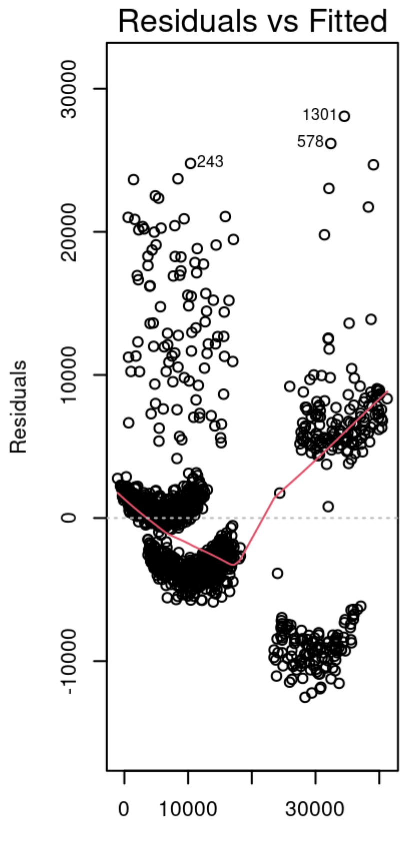


Figure 3A

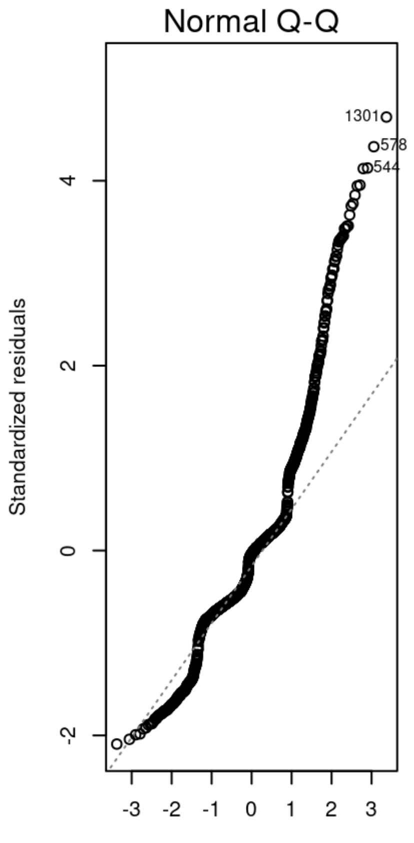


Figure 3B

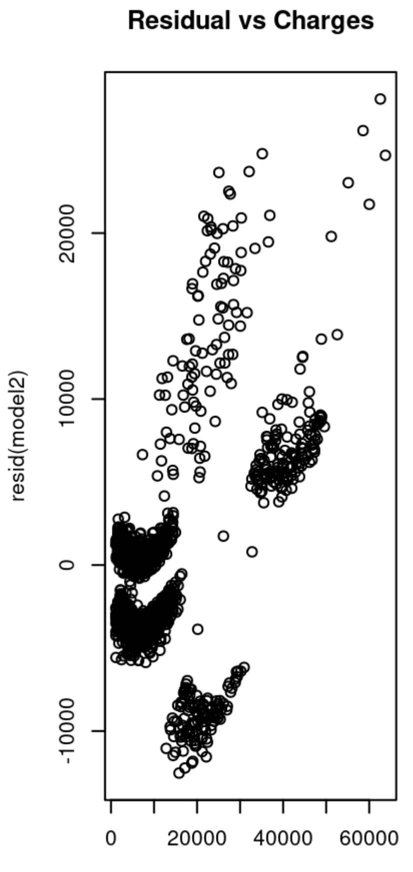


Figure3C

```
## integer(0)
```

Third Model

- Now, we're going to identify the predictor(s) that has the most influence over our response. To do so, we're going to examine the interaction between smoking and BMI. From the overall F-statistic (p-value = < 2.2e-16), at least one of our predictors is significant; this is confirmed by our individual t-test, where we see that all of the predictors, with the borderline exception of the obesity (BMI30), is significant. We're still going to include obesity into our model for comparative purposes: we want to see how obesity affects costs. Consequently, we can see that the interaction between obesity and smoking has a significant effect: smoking alone increases the costs by 13,412.40 dollars, while obese smokers increases costs by 19684.869 dollars. Although our assumptions are still violated, as shown from Figures 4A-C, it still shows a drastic improvement from our previous diagnostics (Figures 2A-C and Figures 3A-C). To support this final model as our "best" model, we contrast the adjusted R.squared values from the previous models to our current one. Our previous adjusted R.squared = 0.7494, and 0.754 from models 1 and 2, respectively, while our current adjusted R.squared = 0.8614, so approximately 86.14% of the variability is explained by the predictors in our third model.

```
insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)
model3 <-lm(charges ~ age  + bmi + bmi30 + children + smoker + smoker:bmi30, data=insura
nce)
summary(model3)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + bmi30 + children + smoker +
##     smoker:bmi30, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19074.0  -1865.8  -1254.9   -440.4  24385.3
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -5097.881    929.887  -5.482 5.02e-08 ***
## age               264.226      8.842  29.883  < 2e-16 ***
## bmi                97.698     33.738   2.896  0.00384 **
## bmi30            -809.791    426.763  -1.898  0.05798 .
## children          512.424    102.395   5.004 6.35e-07 ***
## smokeryes       13412.395    445.186  30.128  < 2e-16 ***
## bmi30:smokeryes 19684.869    612.394  32.144  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4509 on 1331 degrees of freedom
## Multiple R-squared:  0.862,  Adjusted R-squared:  0.8614
## F-statistic:  1386 on 6 and 1331 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(1,3))
plot(model3, 1) + title(sub = "Figure 4A", col.sub = "blue")
```

```
## integer(0)
```

```
plot(model3, 2) + title(sub = "Figure 4B", col.sub = "blue")
```

```
## integer(0)
```

```
plot(resid(model3) ~ charges, data=insurance) + title(main = "Residual vs Charges", sub=
"Figure4C", col.sub="blue")
```

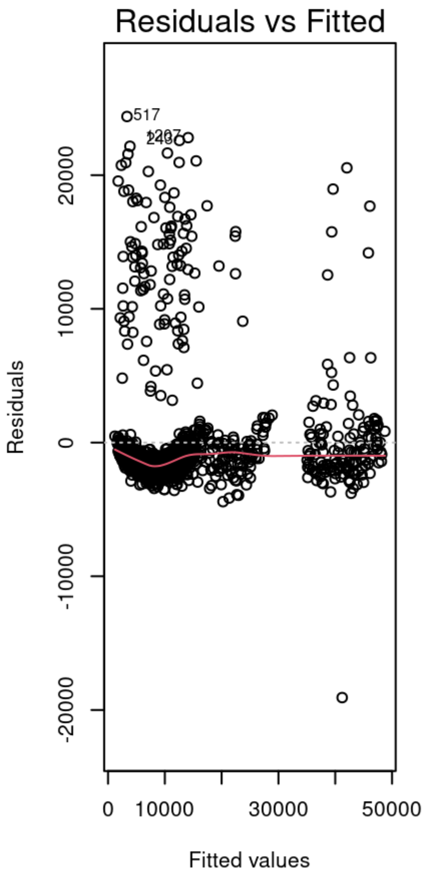


Figure 4A

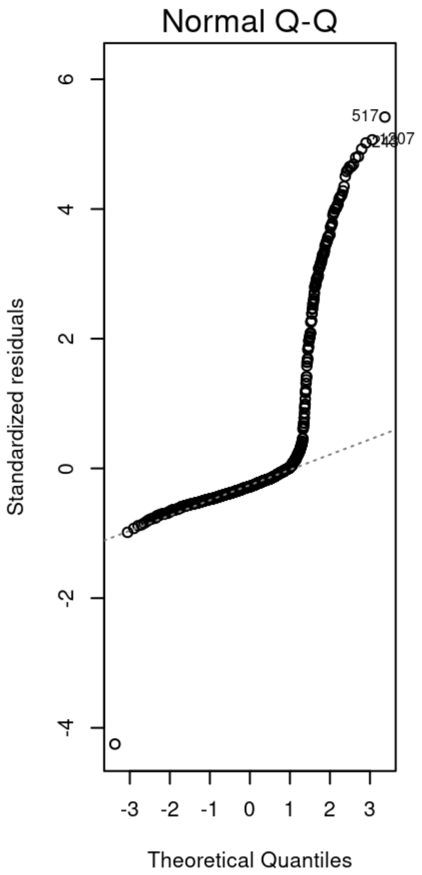


Figure 4B

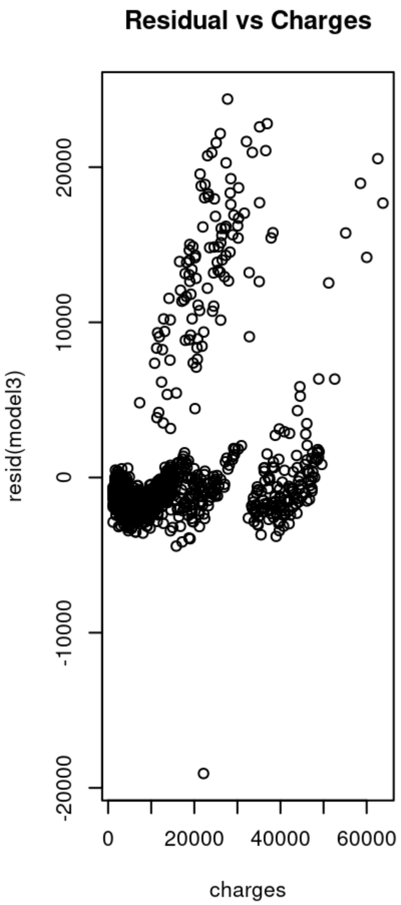


Figure4C

```
## integer(0)
```