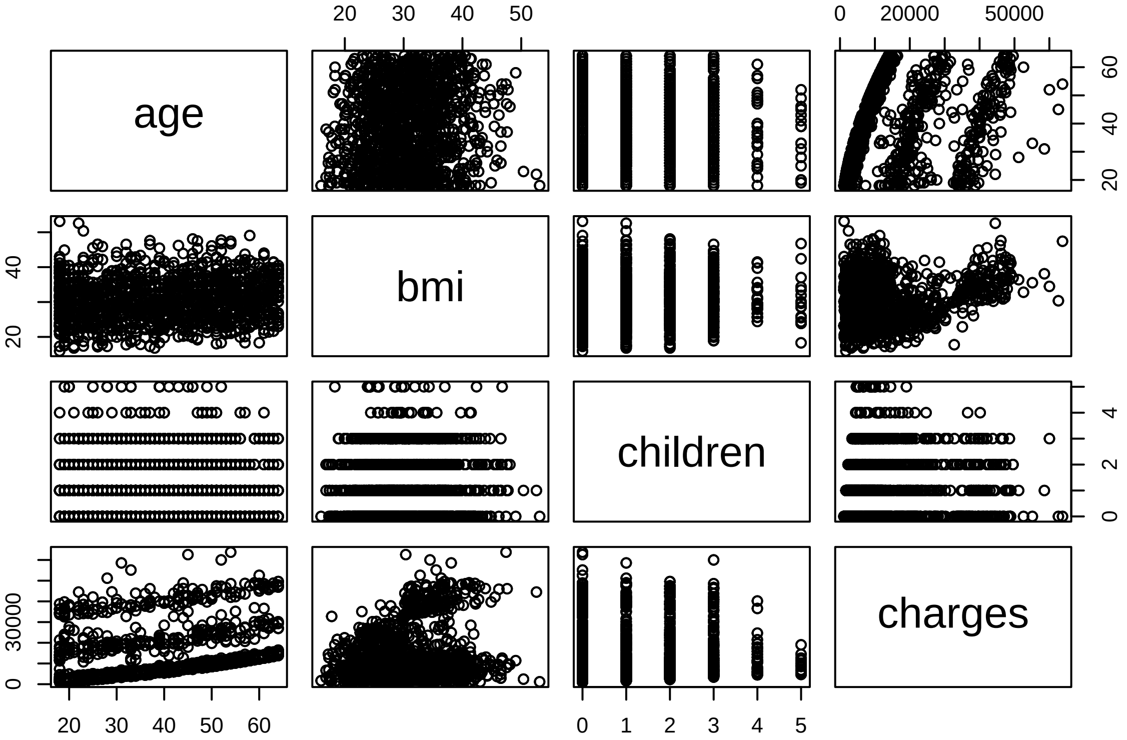
Assignment 4

Import Data (hidden)

Exploratory Data Analysis

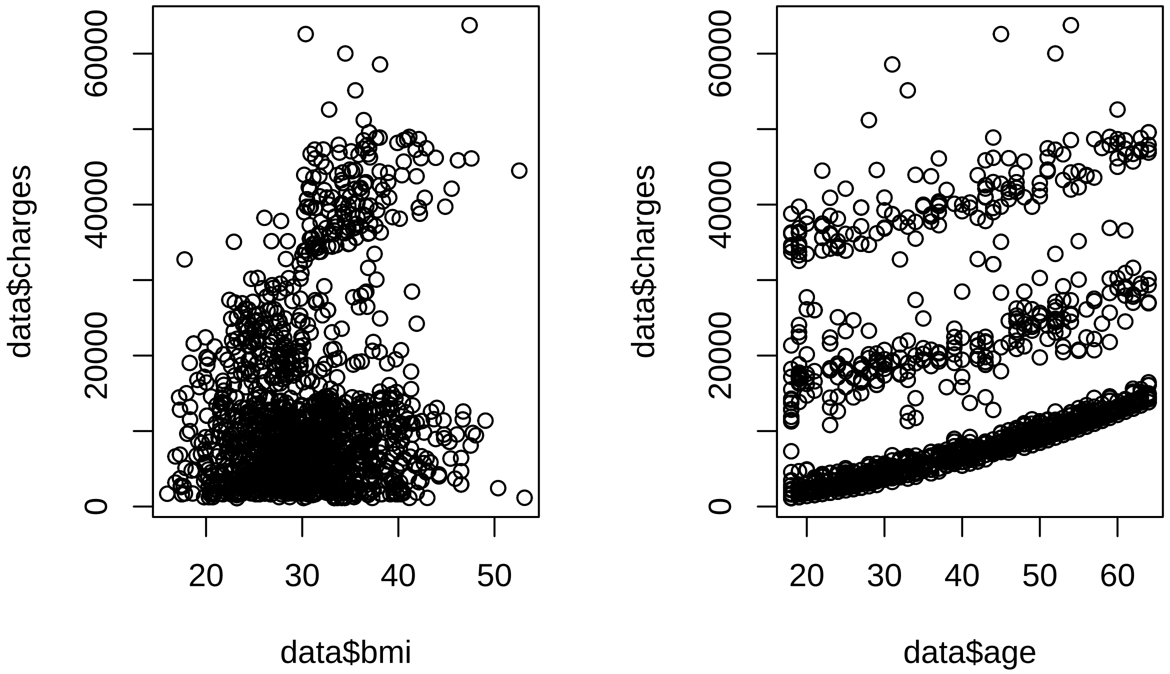
Here, I am graphing the pairs of different numerical variables and trying to see if there is any clear correlation between any two variables.

```
pairs(data[c("age","bmi","children","charges")])
```



The two noticeable pairs are bmi/charges and age/charges. As it seems like there are cohorts within those observations.

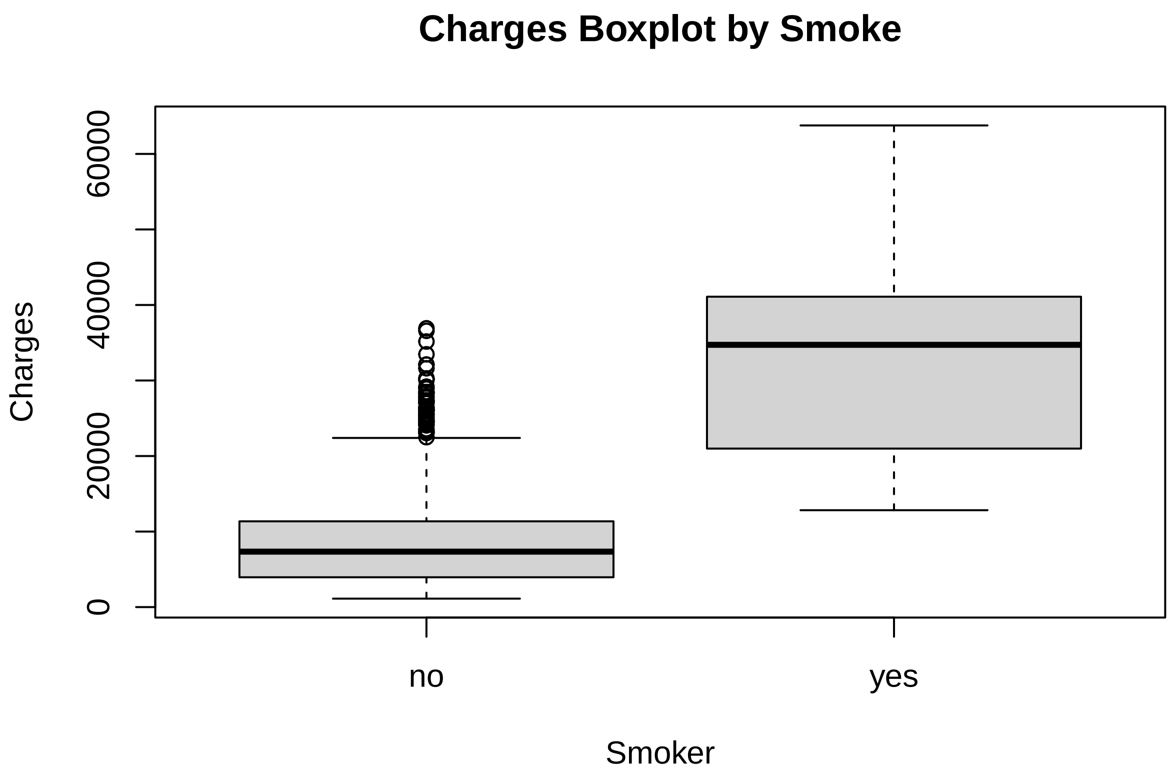
```
par(mfrow=c(1,2))
plot(data$bmi,data$charges)
plot(data$age,data$charges)
```



Here we can infer that age is not the major factor when determining medical charges. This is because people with the same age are broken down into three different charge levels. This means that there is some other major factor. Similarly, looking at the bmi/charges plot we can see two things.

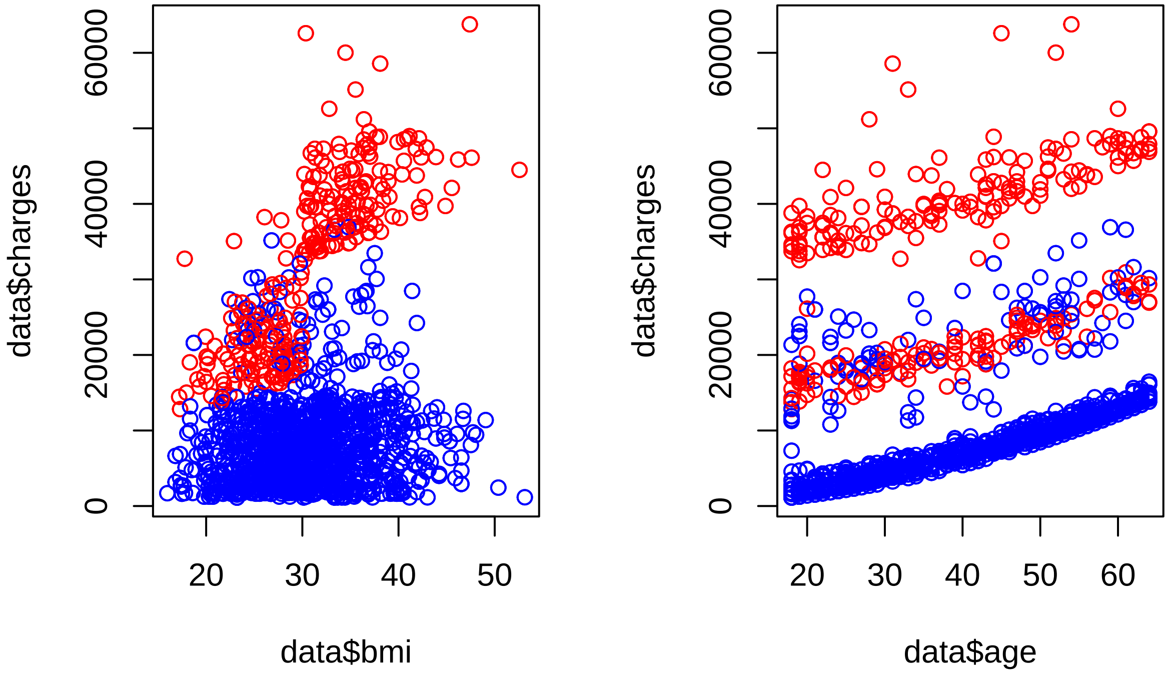
1. Charges seem to start increasing past a 30 bmi. This is also characterized as being obese
2. There is a cohort of individuals who are consistently being charged more then the others. Having some experience in the insurance space, my hunch is that this has something to do with smokers. Lets explore further and see whether being a smoker has a significant impact on medical charges

```
boxplot(charges ~ smoker, data = data,
        xlab = "Smoker", ylab = "Charges",
        main = "Charges Boxplot by Smoke")
```



Here we can see that in fact, smokers pay more in medical fees then non-smokers. Lets look at our previous two plots again but this time we'll colour the points blue if the individual is a non-smoker and red if they are a smoker

```
colours <- ifelse(data$smoker == "yes", "red", "blue")
par(mfrow=c(1,2))
plot(data$bmi,data$charges,col=colours)
plot(data$age,data$charges,col=colours)
```



From the data we can see that smoking has a significant impact on medical charges. Notice, that while smokers tend to have higher medical charges, smokers who are obese (bmi >= 30) have significantly higher medical charges. This is something we should keep in mind as we further analyze the data. Further, when comparing age to charges, it appears that the cohort made up entirely of smokers, by far has the highest medical expenses. This once again reassures me of the fact that smoking is a very important factor for determining medical charges.

Applying a model

First, lets look at a model with all predictors and see if the p-values can tell us anything about the significance of certain predictors

```
model1 <- lm(charges~.,data=data)
summary(model1)
```

```
##
## Call:
## lm(formula = charges ~ ., data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11489  -2789  -1016   1340  29867
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11635.451    686.885 -16.939  < 2e-16 ***
## age                255.577      8.268  30.913  < 2e-16 ***
## sexmale            -56.944    231.866  -0.246  0.80602
## bmi                330.015     19.869  16.609  < 2e-16 ***
## children           506.343     95.164   5.321 1.12e-07 ***
## smokeryes        23976.197    288.461  83.118  < 2e-16 ***
## regionnorthwest   -331.841    334.380  -0.992  0.32109
## regionsoutheast  -1078.362    334.418  -3.225  0.00128 **
## regionsouthwest  -1055.254    333.121  -3.168  0.00155 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6073 on 2763 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7502
## F-statistic:  1041 on 8 and 2763 DF,  p-value: < 2.2e-16
```

Here we see that

- Sex is not having a significant impact on the model
- The fact that someone lives in the northwest region does not have a significant impact. The other regions however do have some significance.

Another thing I'd like to note is that there seems to be a strong relationship between smokers and people with a bmi over 30. Namely, we see a large jump in medical charges for smokers who have a bmi over 30. After some research, it appears that a bmi over 30 is indicative of obesity. Thus, I will create a new categorical variable that characterizes someone as obese vs not obese. Further, from the chart above there is a case to be made about obesity and smoking being interaction terms. We will include this in our model.

```
data$obese <- ifelse(data$bmi >= 30, 1, 0)
```

From the p-values of the full model, we have determined that sex does not play a major role in determining medical charges. Thus, we will drop it from the model.

```
#New model with obesity as a categorical variable, sex dropped
model2<- lm(charges~age+obese+children+smoker,data=data)
summary(model2)
```

```
##
## Call:
## lm(formula = charges ~ age + obese + children + smoker, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13502.5  -3539.8     33.4   1312.1  27350.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4494.485    369.201 -12.174  < 2e-16 ***
## age           259.363      8.187  31.682  < 2e-16 ***
## obese        4104.172    230.725  17.788  < 2e-16 ***
## children      508.705     94.511   5.383 7.96e-08 ***
## smokeryes   23961.870    285.029  84.068  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6039 on 2767 degrees of freedom
## Multiple R-squared:  0.7534, Adjusted R-squared:  0.753
## F-statistic:  2113 on 4 and 2767 DF,  p-value: < 2.2e-16
```

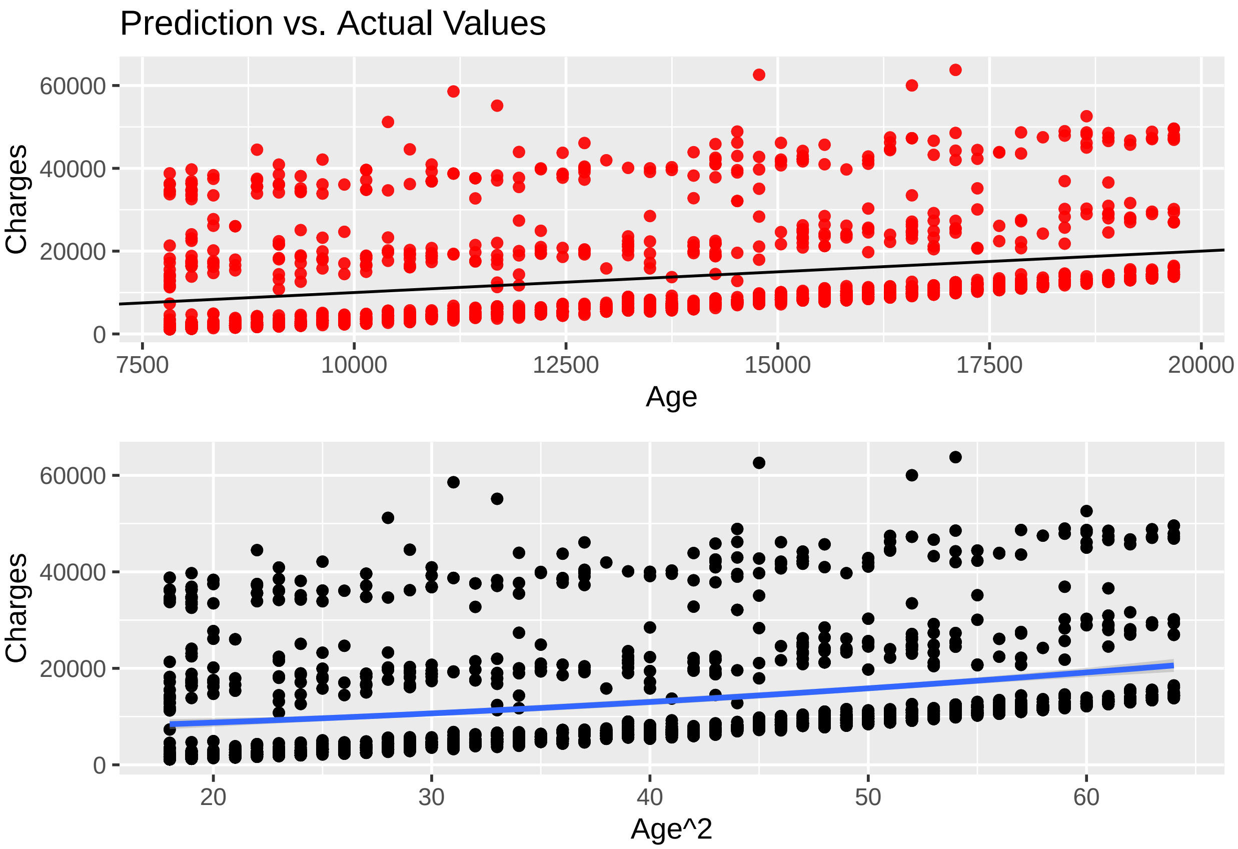
Here we can see that our model is a little bit better, but there are still other things I'd like to check. For example, whether or not age should be polynomial, specifically quadratic.This is because from personal experience, as age increases, the size of medical charges tends to go up as health problems become more serious.

```
library(gridExtra)

a<-ggplot(data, aes(x = predict(lm(charges~age),data=data), y = charges)) +
  geom_point(color = "red", alpha = 0.7) +
  geom_abline(color = "black") +
  ggtitle("Prediction vs. Actual Values") +
  xlab('Age') +
  ylab('Charges')

b<-ggplot(data, aes(x=age, y=charges)) +
          geom_point() +
          stat_smooth(method='lm', formula = y ~ poly(x,2), linewidth = 1) +
          xlab('Age^2') +
          ylab('Charges')

grid.arrange(a,b)
```



Looking at the plots, I believe it is fair to say that there is in fact a quadratic relationship but lets see if there is anything more we can say from an F-test. Referring to appendix figure 2 we can see that there isn't much significance in adding the polynomial term vs the linear term as we get a p-value > 0.05 (specifically, 0.07375). While the value is greater than 0.05, it is still fairly close. Given this fact and the fact that the graph does look slightly more quadratic than it does linear, mixed with my knowledge of real life, I will stick with having age as a quadratic predictor, I believe that for a larger sample my assertion would be true.

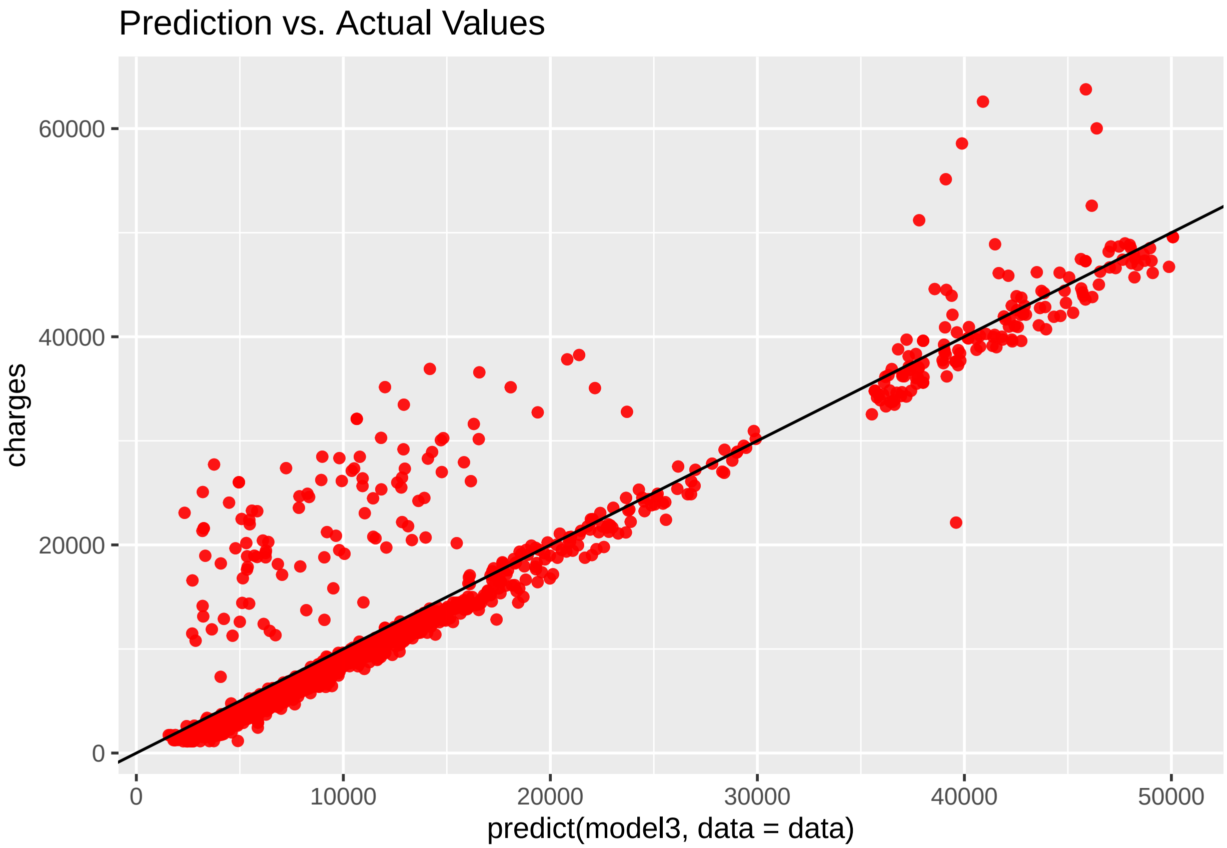
Putting it all together

```
model3<- lm(charges~poly(age,2) + bmi + children + smoker + smoker*obese + region, data=data)
summary(model3)
```

```
##
## Call:
## lm(formula = charges ~ poly(age, 2) + bmi + children + smoker +
##     smoker * obese + region, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17456.8  -1679.4  -1274.1   -662.1  23969.4
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5271.24     637.93   8.263  < 2e-16 ***
## poly(age, 2)1      193945.25    4512.80  42.977  < 2e-16 ***
## poly(age, 2)2       33237.87    4695.83   7.078 1.84e-12 ***
## bmi                   114.61      23.97   4.782 1.83e-06 ***
## children            674.16      73.39   9.186  < 2e-16 ***
## smokeryes         13340.48     310.81  42.922  < 2e-16 ***
## obese              -987.19     296.63  -3.328 0.000886 ***
## regionnorthwest    -224.15     246.33  -0.910 0.362924    
## regionsoutheast    -928.18     246.81  -3.761 0.000173 ***
## regionsouthwest   -1311.80     245.40  -5.345 9.75e-08 ***
## smokeryes:obese   19804.37     423.79  46.731  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4473 on 2761 degrees of freedom
## Multiple R-squared:  0.865,  Adjusted R-squared:  0.8645 
## F-statistic:  1769 on 10 and 2761 DF,  p-value: < 2.2e-16
```

```
library(ggplot2)

ggplot(data, aes(x = predict(model3,data=data), y = charges)) +
  geom_point(color = "red", alpha = 0.7) +
  geom_abline(color = "black") +
  ggtitle("Prediction vs. Actual Values")
```



# Outliers

Now, lets look at the outliers in our data

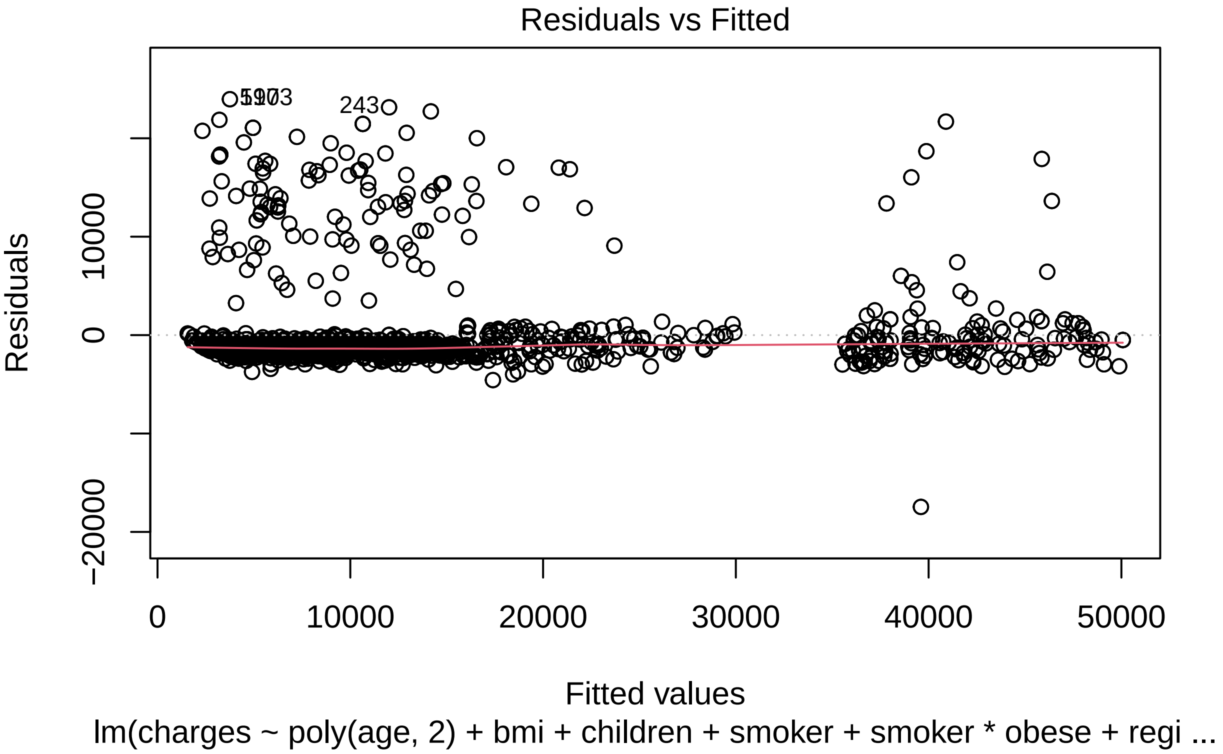
```
# Find the standardized residuals
stand_residuals <- rstandard(model3)

# Identify the outliers observations
```

```
outlier_observations<-stand_residuals[abs(stand_residuals) > 3]

# Outlier_observations (hidden)

#Plotting Residuals vs Fitted
plot(model3,1)
```



Looking at the residuals to predicted values, I can see that there are some outliers. While it may be tempting to drop those outliers it's not correct to do so as we don't know whether those are true outliers to the whole population. They may be just outliers to the specific sample and thus incorporating them into our model will give us a better predictive model for the whole population.

Note: Refer to appendix figure 1 for influential points, calculated using cook's distance and high leverage. There are no influential points in our data however there are high leverage points. Similar to the outliers, we will not drop them as we cannot say that they are completely out of the ordinary for a population, however it is important to note that they exist and have influence on our model.

#Final Model

$$y = age^2 + bmi + children + smoker + region + smoker * obese$$

Overall, I think my model is fairly effective at predicting medical charges. I believe there is enough justification for the predictors I have chose to include. I tried to keep the report fairly concise and not include every possible interaction term but from exploration I found that including smoker*obese yields almost the same model as including other interaction terms with smokers. Furthermore, with an adjusted R^2 of 0.8645 I believe that my model accounts for much of the variance in the observations.

Appendix

Figure 1

```
#Part c)
# Find cook's distance
cooks_dist <- cooks.distance(model3)

# Find the influential observations
influential_observations<-cooks_dist[cooks_dist > 1]

influential_observations
```

```
## named numeric(0)
```

```
# There is no influential observations in the dataset

# Find the leverage value of each observation
hat_values <- hatvalues(model3)

threshold_lev <- (2*length(coef(model3))/length(hat_values))
# The following code will identify the observations with high leverage and assign it to high_leverage v
high_leverage<-hat_values[hat_values> threshold_lev]

high_leverage
```

```
##          33          93          95         104         167         245
## 0.009010043 0.008189698 0.008706467 0.008704722 0.008294257 0.007939109
##         251         439         495         665         848         861
## 0.008546468 0.008733661 0.008818968 0.008855936 0.008081850 0.009397614
##         891         985        1048        1086        1205        1266
## 0.008441747 0.008913689 0.012221095 0.010181490 0.008312186 0.008690965
##        1302        1318        1419        1479        1481        1490
## 0.008765869 0.011916879 0.009010043 0.008189698 0.008706467 0.008704722
##        1553        1631        1637        1825        1881        2051
## 0.008294257 0.007939109 0.008546468 0.008733661 0.008818968 0.008855936
##        2234        2247        2277        2371        2434        2472
## 0.008081850 0.009397614 0.008441747 0.008913689 0.012221095 0.010181490
##        2591        2652        2688        2704
## 0.008312186 0.008690965 0.008765869 0.011916879
```

Figure 2

```
anova(lm(charges~age,data=data),lm(charges~age+poly(age,2),data=data))
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ age
## Model 2: charges ~ age + poly(age, 2)
##   Res.Df        RSS Df Sum of Sq      F  Pr(>F)
## 1   2770 3.7269e+11
## 2   2769 3.7226e+11  1 430188233 3.1999 0.07375 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```