HW 12

Priya Shaji

11/9/2019

Using R, build a multiple regression model for data that interests you. Include in this model at least one quadratic term, one dichotomous term, and one dichotomous vs. quantitative interaction term. Interpret all coefficients. Conduct residual analysis. Was the linear model appropriate? Why or why not?

Dataset: Insurance dataset from kaggle('https://www.kaggle.com/mirichoi0218/insurance)

Description: This dataset looks at medical insurance costs charges for various people based on several factors like number of children, region of residency, age etc.

Step 1) Load the dataset

Step 2) Display first few rows of insurance dataset

```
insurance <- read.csv("https://raw.githubusercontent.com/PriyaShaji/Data605/master/week%2012/insurance.csv")
```

```
head(insurance)
```

	age	sex	bmi	children	smoker	region	charges
	<int>	<fctr>	<dbl>	<int>	<fctr>	<fctr>	<dbl>
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

6 rows

Step 3) Summarize the dataset and display the dimensions

```
summary(insurance)
```

```
##       age            sex           bmi           children       smoker
##  Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064
##  1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274
##  Median :39.00                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
##        region       charges
##  northeast:324   Min.   : 1122
##  northwest:325   1st Qu.: 4740
##  southeast:364   Median : 9382
##  southwest:325   Mean   :13270
##                  3rd Qu.:16640
##                  Max.   :63770
```

```
dim(insurance)
```

```
## [1] 1338    7
```

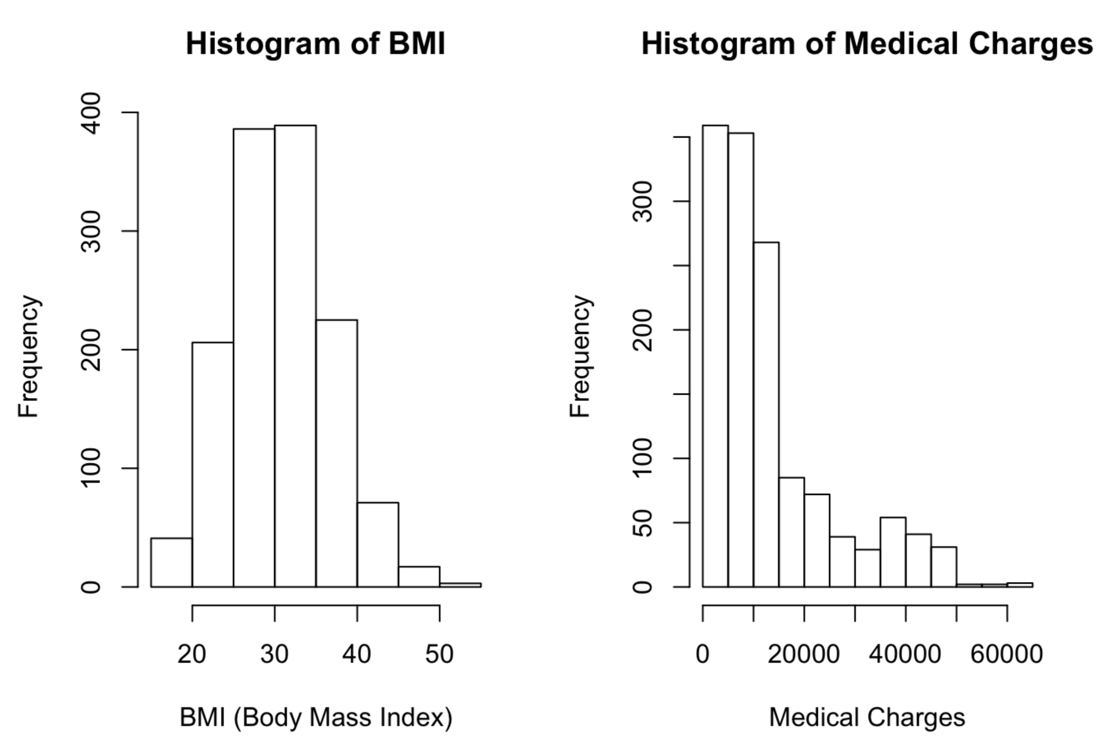
```
str(insurance)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

From the above steps we see that the dataset is tidy and clean.

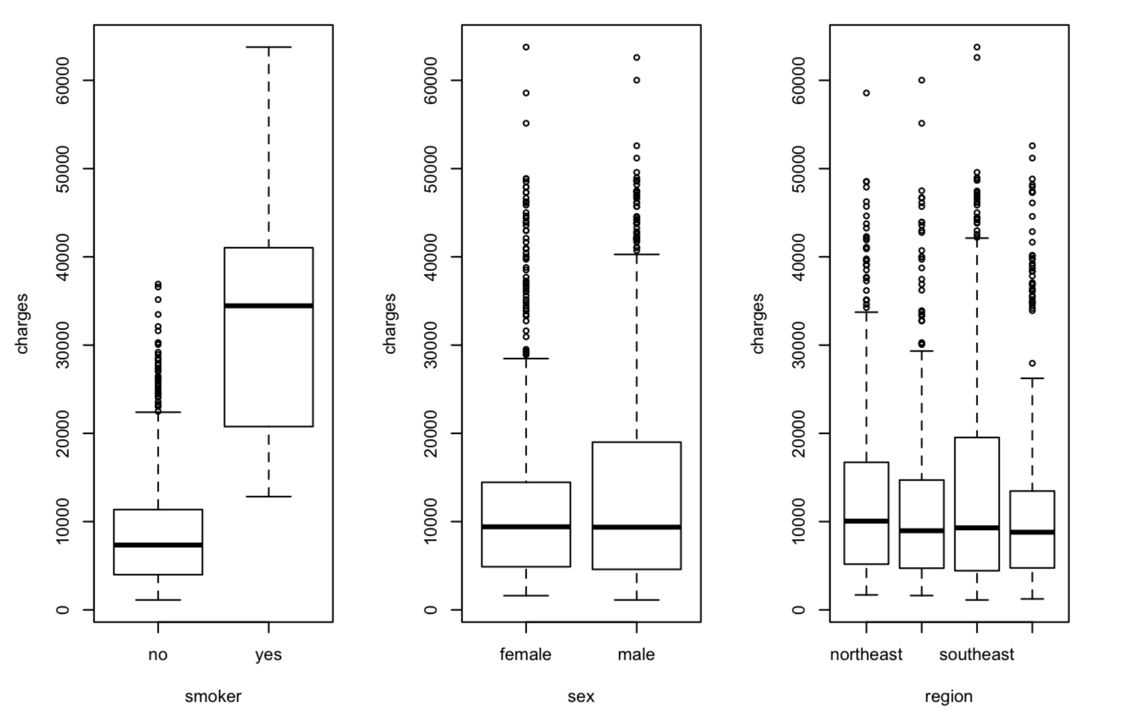
Step 4) Now, let's analyze it using graphs

```
par(mfrow=c(1,2))
hist(insurance$bmi, xlab = "BMI (Body Mass Index)",
     main = "Histogram of BMI")
hist(insurance$charges, xlab = "Medical Charges",
     main = "Histogram of Medical Charges")
```



Histogram of medical charges looks more right-skewed.

```
par(mfrow=c(1,3))
with(insurance, plot(charges ~ smoker + sex + region))
```



I have plotted above graphs using smoker, gender and region.

We see that BMI is nearly normally distributed, medical charges is right-skewed and there are many outliers for high medical charges against both genders and various regions.

We also see that the median is about the same for all regions, and genders.

Note that for smokers, medical charges are much higher than normal ones which we should expect.

Now, let's fit a multiple regression model, let have the explanatory variables as

sex (categorical)

bmi (numerical, continous)

age (numerical, discrete)

smoker (categorical)

charges (numerical, continous)

Step 5) Let's make a multiple regression model of the following equation:

$$charges = \beta0 + \beta1 * Sex + \beta2 * bmi + \beta3 * age + \beta4 * smoker + \beta5(bmi * sex)$$

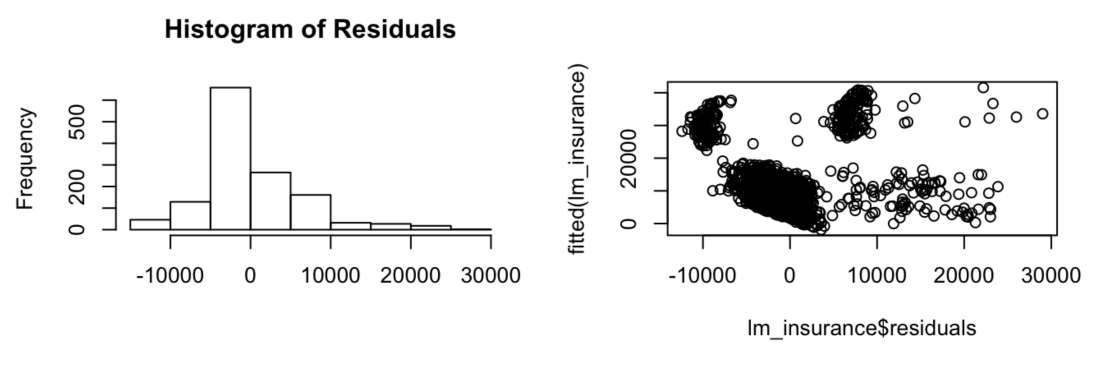
```
lm_insurance <- lm(charges ~ sex + bmi + age + smoker + bmi*sex, data = insurance)
summary(lm_insurance)
```

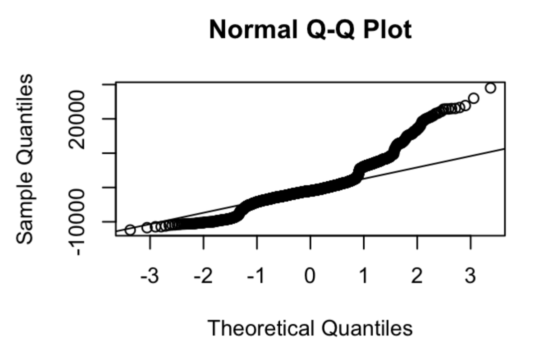
```
##
## Call:
## lm(formula = charges ~ sex + bmi + age + smoker + bmi * sex,
##     data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12396.7  -2983.0   -985.4   1478.3  29015.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11717.369   1282.175  -9.139  < 2e-16 ***
## sexmale         54.096   1712.963   0.032    0.975
## bmi            325.779     39.341   8.281 2.94e-16 ***
## age            259.469     11.947  21.718  < 2e-16 ***
## smokeryes    23836.067    414.958  57.442  < 2e-16 ***
## sexmale:bmi     -5.326     54.846  -0.097    0.923
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6097 on 1332 degrees of freedom
## Multiple R-squared:  0.7475, Adjusted R-squared:  0.7466
## F-statistic: 788.6 on 5 and 1332 DF,  p-value: < 2.2e-16
```

Conduct residual analysis

Step 6) Plot Histogram of Residuals

```
par(mfrow=c(2,2))
hist(lm_insurance$residuals, main = "Histogram of Residuals", xlab= "")
plot(lm_insurance$residuals, fitted(lm_insurance))
qqnorm(lm_insurance$residuals)
qqline(lm_insurance$residuals)
```





We see that the residuals histogram is somewhat normal but the residuals vs fitted values doesn't show constant variance which is not good for a multiple regression model.

The equation of this multiple regression model is as follows:

$$charges = -11717.37 + 54.1 * sex + 325.78 * bmi + 259.47 * age + 23836.07 * smoker - 5.33(bmi * sex)$$

Note the variables

sex = 1 for male and 0 for female

smoker = 1 for male and 0 for female

Interpret all coefficients

What does this tell us? Let's look at the details of the summary in more detail

Coefficients:

Intercept: This tells us that leaving all other terms constant, on average the estimated medical charge is about $-11717.36 which logically won't make sense and is good there are other terms in the model.

Sex: If a person is male and leaving all other terms constant, he can expect to pay about $54.1 in medical costs.

BMI: Leaving all other terms constant, a person can be expected to pay about $325.78 in medical charges per BMI value.

Age: Leaving all other terms constant, a person can be expected to pay about $259.47 in medical expenses multiplied by their age (A 31 year old will pay about $8043.57)

Smoker: A person who smokes and leaving all variables constant can expect to pay $23836.07

Sex*BMI: A male can expect to pay holding all other variables constant can expect to pay $-5.326 which doesn't make sense logically.

P-values of coefficients:

The p-values of the intercept, bmi, age and male smokers are very low and we can reject the null hypothesis (H_0 = 0) and favor the alternative (H_A != 0) that is the true coefficients is not 0

For Males and Male*bmi, we fail to reject the null hypothesis and thus these coefficients are very close to 0 and can be excluded in our model.

Residual Standard Error: The residual standard error of 6097 is the standard deviation and is a bit far from the good fit of points.

R-squared/Adjusted R^2: values of 0.7475 and 0.7466 respectively, this means that about 75% of the data fall into the regression line.

F-statistic: value of 788.6 with a small p-value < 2.2e-16 means that the features selected are better than the intercept-only model which as described before makes sense as a intercept only model gives a negative medical cost which doesn't apply or make sense.

Conclusions

The above model does not have much efficiency, as there are coefficients that can be removed or probably added for better accuracy and properly modeling and predicting medical costs. The residual standard error as well as the Q-Q plots show that the model is not a good fit for the data. One good thing I can say about the model is that the BMI and Age coefficients make sense as the more your BMI is and older, you are more likley to have more health problems and have more medical costs to pay.

Future work can be done to add more coefficients, transforms and possibly use non-linear regression to better predict medical costs.