SUMAN ANAND LOHIT

NET ID: SAL200005

BUAN/ OPRE 6359 : Advanced Statistics for Data Science

ANALYSIS OF MEDICAL COSTS

INTRODUCTION **1**

PROBLEM STATEMENT **1**

PRELIMINARY ANALYSIS **2**

SOLUTION **4**

CONCLUSIONS **7**

APPENDIX 1 : INITIAL ANALYSIS **8**

APPENDIX 2 : MULTIVARIATE REGRESSION **11**

APPENDIX 3 : ANALYSIS OF CHARGES **15**

BIBLIOGRAPHY **18**

INTRODUCTION

Insurance companies need to estimate the medical costs to fix the appropriate premiums to be collected from their clients. For that purpose, we analyze the factors driving medical costs. In this paper, I have used the Medical Costs Personal dataset[1] which is a simulated dataset containing medical expenses for patients in the United States by Brett Lantz. Data was created with demographic statistics from the U.S. Census Bureau[2].

Questions that I seek to answer in this paper are as follows:

1. What features/characteristics of the beneficiaries affect their medical costs?
2. How does the age of the beneficiary affect the changes in their medical costs?
3. How does smoking impact the medical costs of the beneficiaries?

I have used multivariate regression focusing on important variables, and on some specific interactions to understand their combined effect on costs. I have also explored the relationship of age and its combined effect with other explanatory variables on medical costs.

PROBLEM STATEMENT

The focus of this paper is to analyze and identify what variables provide a significant explanation for the changes in medical costs; how the age and the smoking habits of a beneficiary affect their medical costs.

Here is a brief description of the variables in the dataset :-

Age: Age of the primary beneficiary

Sex: Sex of the policy holder, male or female

BMI: Body mass index (BMI). Ideal BMI is in the range 18.5 to 24.9

[1] Miri Choi, "Medical Cost Personal Datasets," Kaggle, February 21, 2018, https://www.kaggle.com/mirichoi0218/insurance.
[2] Lantz, Brett. In *Machine Learning with R: Learn...Real-World Applications*, Chapter 6,p 173. Packt Publ., 2015.

Children: Number of children or dependents covered by the plan

Smoker : Whether the beneficiary is a regular smoker (yes/no)

Region : Beneficiary's place of residence - northeast, southeast, northwest or southwest

Charges: Total medical expenses charged to to the plan in a year (response variable)

PRELIMINARY ANALYSIS

The dataset has 1338 records, 7 variables, and no missing values. To ensure that the regression model assumptions are satisfied, I have computed summary statistics, correlation matrix, skewness, and kurtosis tests. (Refer to Appendix 1 for a detailed analysis)

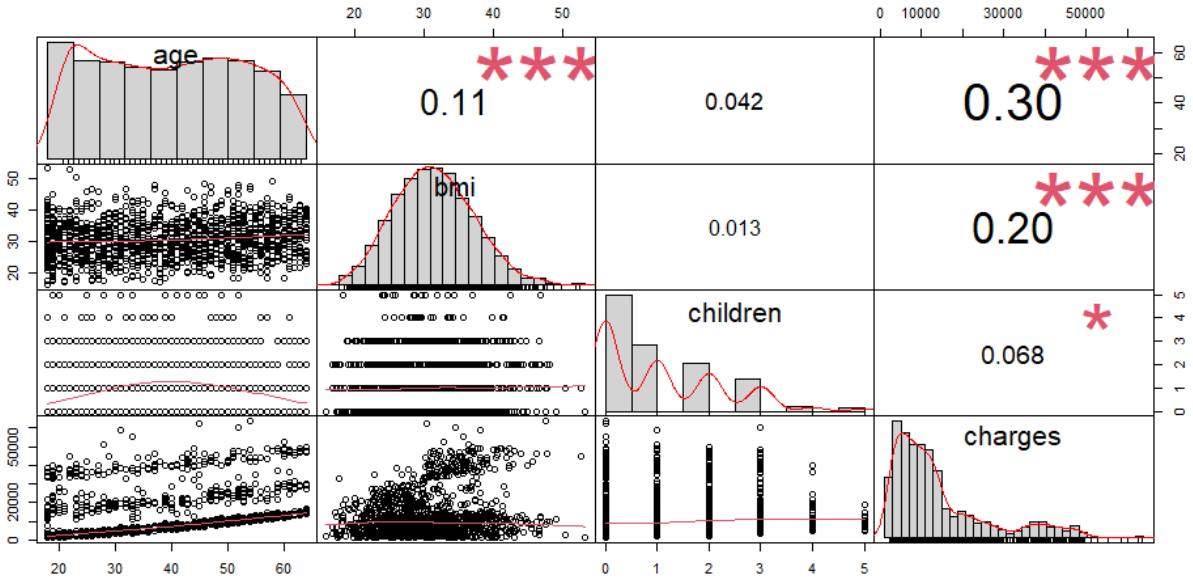
Summary statistics:

	Minimum	Median	Mean	Maximum	Std. Deviation
Age	18.00	39.00	39.21	64.00	14.05
BMI	15.96	30.40	30.66	53.13	6.10
Children	0.00	1.00	1.09	5.00	1.20
Charges	1,122	9,382	13,270	63,770	12110

Sex	Minimum	Median	Maximum	Mean	Std. Deviation	Count
female	1607.51	9412.96	63770.43	12569.58	11128.70	662
male	1121.87	9369.62	62592.87	13956.75	12971.03	676

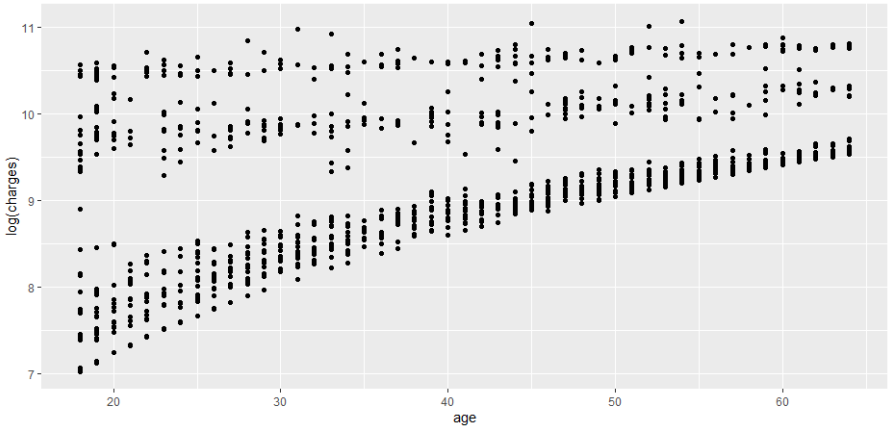
Region	Minimum	Median	Maximum	Mean	Std. Deviation	Count
northeast	1694.80	10057.65	58571.07	13406.38	11255.80	324
northwest	1621.34	8965.80	12417.58	11072.28	11072.28	325
southeast	1121.87	9294.13	64770.43	14735.41	13971.10	364
southwest	1241.57	8798.59	52590.83	12346.94	11557.18	325

Fig 1: Correlation chart



A matrix of scatterplots has been done to understand the relationship of non-categorical variables with *Charges* (Fig 1). There are no highly correlated variables. The highest correlation is between *Charges* and *Age* at 0.30. It can be seen that *Charges* is skewed to the right and thus, must be log-transformed to make it approximately normal. Another cause for concern is the scatterplot of *Age* vs. *Charges*, which shows three distinct lines of data points.

Fig 2: Scatterplot of "Age" vs. "Log Charges"



A squared term for *Age* has been adopted to address the curved portion in the scatterplot (Fig.2). A multivariate regression has been done to study the residuals for outliers. Four influential outliers have been removed. Two dummy variables have been added - one for BMI to check if there is a significant impact on *Charges* when the BMI was over 30; another for *Children* with five levels by combining all levels with over 3 children into one level to check if having over three children impacted the costs significantly. (Appendix 2)

SOLUTION

A multivariate regression model has been run with the following parameters :

The original variables - *age*, *smoker*, *region,* and *sex*; a squared term for age (*Age.2*), the dummy variables of *BMI* and *children*. Interaction terms were added to get the combined effect of *age* with *smoker*. Another set of interaction terms were added to get the combined effect of *smoker* with *BMI* (dummy), *children,* and *region*.

ANALYSIS OF CHARGES

Charges has been analyzed with respect to age for smokers and non-smokers separately at ages 20 and 60 years to see if the behavior changes with age. The variation in *Charges* has been analyzed for smokers and non-smokers, keeping all other factors constant, and iterated over both BMI levels, and number of children at a particular age.

Charges~ age

$LogCharges = ß_0 + ß_1*Age + ß_2*Age.2 + ß_3*BmiHigh + ß_4* Children + ß_5*Smoker + ß_6*Region + ß_7*BmiHigh*Smoker + ß_8*Smoker*Children + ß_9*Age*Smoker$

For smokers, % change in *Charges* with an increase of 1 year in *Age* : When age is 20 years, the percentage change in the median of *Charges* when age goes from 20 to 21 years is estimated to be 1.82% increase (95% Confidence Interval: 0.01% to 3.69%). When age is 60 years, for an

increase of 1 year in age from 60 to 61 there is an estimated decrease in the median of *Charges* by 0.05% (95% C.I : 2.79% less to 2.76% more), keeping other variables constant. The increasing trend seems to be reversed when age > = 59 years.(For a detailed analysis, refer Appendix 3)

For non-smokers, % change in *Charges* with an increase of 1 Year in age : When age is 20 years, the percentage change in the median of *Charges* when age moves from 20 to 21 years is 5.28% (95% C.I : 3.75% to 6.84%). When age is 60 years, an increase of 1 year in age from 60 to 61 leads to an increase in *Charges* by 3.33% (95% C.I : 0.85% to 5.88%), keeping other variables constant.

As observed here, this percentage change is different for smokers and non-smokers and the values depend on the age of the subject. For smokers, the increase in median charges is significant in the younger years but this change has decreased to a negative value in later years.

Charges ~ smoker

When age is 50 years, the ratio of median medical charges for smokers to that of non-smokers is tabulated below.

	Children				
BMI	0	1	2	3	>3
≤ 30	2.66	2.16	1.85	1.95	1.56
> 30	5.38	4.36	3.73	3.94	3.14

We can observe from above that, when BMI ≤ 30 and with no children, Median medical charges for smokers are 2.66 times that of non-smokers, while controlling for other variables. Similar observations can be made for all other combinations of BMI and children. Notice that the ratio doubles for BMI values above 30 when compared to those below 30, keeping *Children* constant.

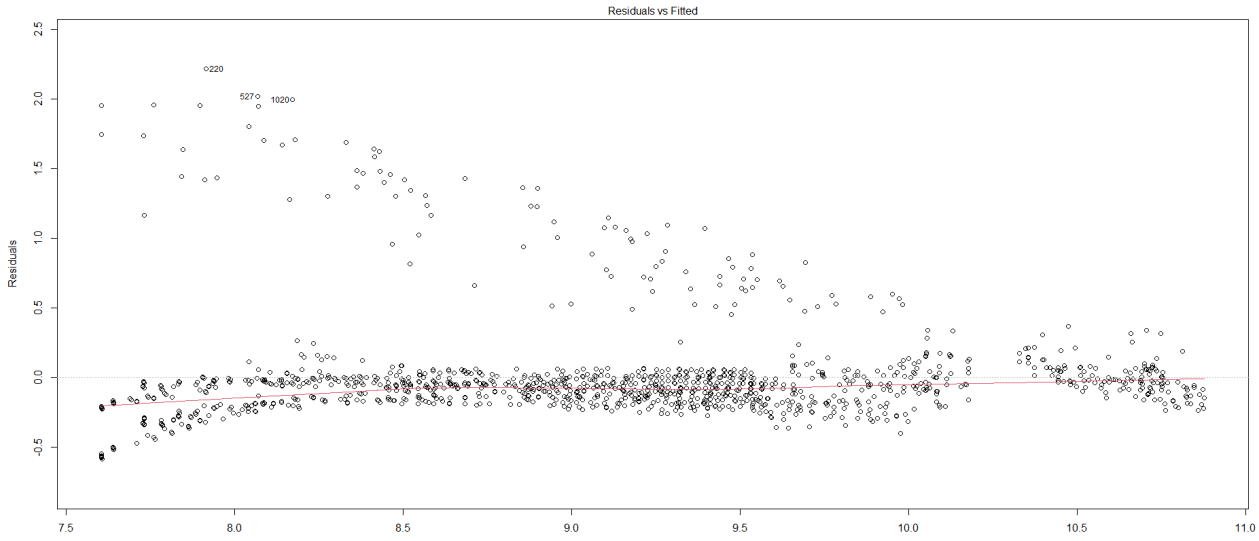
Summary of the regression model provides us the following:

Both *Age* and *Age.2* are significant with very small p-values. The combined effect of *Age* with *Smoker* is also significant in explaining the variation in *Charges*. This makes sense as the age of the beneficiary does have a bearing on their medical charges.

The new *BMI* (dummy) is not significant. However, when *BMI* is combined with *Smoker*, the effect is significant. In fact, both *Smoker* and all its interaction terms are significant, indicating that whether a beneficiary is a smoker or not plays an important role in estimating their medical charges.

The Residual Standard Error of the model is 0.35 on 1316 degrees of freedom. The Adjusted R-squared value is 0.85, i.e., 85% of the variation is explained by the model. The p-value of the F-statistic of the model is <0.0001. This indicates that the model is significant in explaining the changes in log-transformed *Charges*.

Fig 3: Residuals vs. Fitted values of the regression model



The residual plot (Fig 3) resulting from the regression model has a horizontal red line close to 0 on the y-axis which is an indication of a good model. However, we can also see a

pattern above the line. This could be because the relationship of *Age* with *Charges* was not captured completely in the model. While we could improve the adjusted R-square by considering an extensive model (Refer to Appendix 2), it would be a small change for multiple added interactions. I chose the simpler model taking into account the principle of parsimony.

CONCLUSIONS

It has been observed that the Medical charges have been significantly influenced by *Age*, *Children*, *Smoker,* and *Region*. Additionally, *BMI* (as a factor) also influences the charges in the form of an interaction term with *Smoker.* The linear model discussed in this paper accounts for a majority of variation. But from the Residuals vs Fitted values in *Fig 3*, we see a portion of residuals follow a linear pattern. This could be the result of a confounding variable, which is not a part of the dataset, and its interaction with *Age* and *Charges*.

Median Medical charges seem to be increasing with age for non-smokers over the years, while the rate of increase in charges has been decreasing with age. For smokers, It has been observed that the Median Medical charges have been increasing with age for younger people and start decreasing for people older than 59 years old. This could possibly be attributed to already high medical charges for smokers and to the fact that smoking drives a major chunk of their charges as compared to age.

Median Medical charges for smokers are significantly higher than that of non-smokers. It was about 85% to 438% higher than that of Median Medical Charges for non-smokers depending on the BMI and number of children for a 50-year-old beneficiary, keeping other variables constant.

APPENDIX 1 : INITIAL ANALYSIS

Variables *sex, smoker*, and *region* have been converted to factors.

A skewness test is done to check the normality of the Dependant variable.

Skewness and Kurtosis of Charges:

```
kurtosis(charges)
[1] 4.595821
skewness(charges)
[1] 1.51418
```

Charges ~ Sex

An F-test was done to compare the variances of the male and female groups. Since, F-test is not robust to non-normality of *Charges*[3], logged version is used.

H_0 : True ratio of variances of the groups female and male is equal to 1

$H_{1:}$: True ratio of variances is not equal to 1

```
var.test(log(charges) ~ sex, data= medCost, alternative = "two.sided")
F test to compare two variances

data:  log(charges) by sex
F = 0.73388, num df = 661, denom df = 675, p-value = 6.709e-05
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6305578 0.8542413
sample estimates:
ratio of variances
      0.7338819
```

From the above F test, we have evidence to say that the variances of charges in male and female groups are unequal.

[3] Carol A. Markowski and Edward P. Markowski, "Conditions for the Effectiveness of a Preliminary Test of Variance," *The American Statistician* 44, no. 4 (1990): p. 322, https://doi.org/10.2307/2684360.

Next, I have performed a Welch two sample t - test to compare the means of charges in female and male groups.

H_0 : Difference in the means of female and male groups is greater than or equal to 0

$H_{1:}$: Difference in the means of female and male groups is less than 0

```
t.test(log(charges) ~ sex, data= medCost)

        Welch Two Sample t-test

data:  log(charges) by sex
t = -0.20619, df = 1312.9, p-value = 0.8367
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.10886767  0.08815968
sample estimates:
mean in group female   mean in group male
        9.093428           9.103782
```

Since P- value is >0.05 , we fail to reject the null hypothesis and conclude that true difference in means is equal to zero. Hence the median medical charges for female are the same as median medical charges for male.

Anova test for Charges with respect to Region

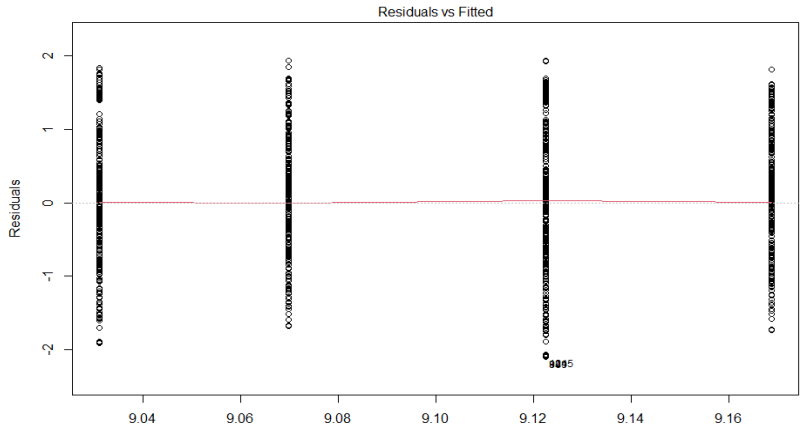
```
myAov<-aov(log(charges) ~ region)

               Df    Sum Sq   Mean Sq  F value  Pr(>F)
region          3     3.6      1.1844   1.402    0.241
Residuals    1334  1126.9      0.8448
```

The p-value is greater than 0.05(alpha). We fail to reject the null hypothesis. All the group means are equal.

Residual vs. Fitted values of the Anova model is plotted below:

plot(myAov, which=1)



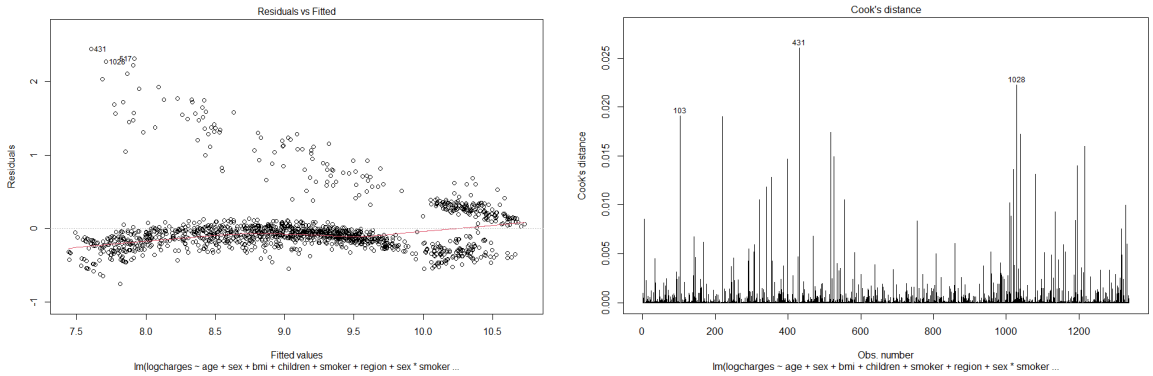
The red line is flat and exactly at 0 on the y-axis. There are no visible patterns in the plot which is good.

APPENDIX 2 : MULTIVARIATE REGRESSION

Squared term for age is taken and a regression model is run taking all interaction terms of age.

Diagnostic plots are run and outliers are found.

```
medCost$age.2 <- medCost$age^2
charges_lm_age2 <- lm(logcharges~age+sex+bmi+children+smoker+region+sex*smoker
+ age*smoker +age*children +age*sex + age*smoker*children,
                      data = medCost)
summary(charges_lm_age2)
plot(charges_lm_age2, which=1)
plot(charges_lm_age2, which=4)
```



There are four outliers and upon removing them, the residual plot's pattern has reduced and the adjusted R-squared value has improved from 0.82 to 0.84. BMI is modified into a dummy variable with 2 levels, less than or equal to 30 and greater than 30.

```
medCost2= medCost1
medCost2$bminew<- ifelse(medCost2$bmi > 30, 1, 0)

medCost2$factorChildren <- ifelse(children==0, "0",ifelse(children ==1, "1",ifelse(children
== 2,"2", ifelse(children==3, "3","Other"))))
charges_lm4 <- lm(logcharges ~ age + smoker + factorChildren + bminew +
                  region + sex + age.2 + age:smoker + age:factorChildren + smoker:bminew +
                  smoker:factorChildren + age:sex + age:region + smoker:region +
                  smoker:sex + age:smoker:factorChildren + age:smoker:sex,
```

data = medCost2)
summary(charges_lm4) #86.44

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.892e+00	1.048e-01	65.760	< 2e-16 ***
age	5.859e-02	4.978e-03	11.769	< 2e-16 ***
smokeryes	2.476e+00	1.245e-01	19.880	< 2e-16 ***
factorChildren1	4.136e-01	8.325e-02	4.968	7.66e-07 ***
factorChildren2	1.024e+00	9.934e-02	10.310	< 2e-16 ***
factorChildren3	7.765e-01	1.209e-01	6.421	1.89e-10 ***
factorChildrenOther	1.327e+00	1.945e-01	6.821	1.39e-11 ***
bminew	1.220e-02	2.150e-02	0.568	0.570319
regionnorthwest	-1.816e-01	8.153e-02	-2.227	0.026129 *
regionsoutheast	-4.226e-01	7.945e-02	-5.319	1.23e-07 ***
regionsouthwest	-4.680e-01	8.271e-02	-5.658	1.88e-08 ***
sexmale	-3.626e-01	6.232e-02	-5.819	7.47e-09 ***
age.2	-2.359e-04	5.918e-05	-3.987	7.07e-05 ***
age:smokeryes	-3.390e-02	2.826e-03	-11.994	< 2e-16 ***
age:factorChildren1	-5.917e-03	2.003e-03	-2.954	0.003197 **
age:factorChildren2	-1.731e-02	2.404e-03	-7.202	1.00e-12 ***
age:factorChildren3	-1.155e-02	2.769e-03	-4.171	3.24e-05 ***
age:factorChildrenOther	-2.011e-02	4.877e-03	-4.123	3.97e-05 ***
smokeryes:bminew	6.712e-01	4.765e-02	14.085	< 2e-16 ***
smokeryes:factorChildren1	-3.643e-01	1.915e-01	-1.902	0.057351 .
smokeryes:factorChildren2	-9.594e-01	2.030e-01	-4.727	2.52e-06 ***
smokeryes:factorChildren3	-8.519e-01	2.389e-01	-3.566	0.000375 ***
smokeryes:factorChildrenOther	-9.752e-01	8.344e-01	-1.169	0.242714
age:sexmale	6.104e-03	1.492e-03	4.092	4.54e-05 ***
age:regionnorthwest	2.498e-03	1.920e-03	1.301	0.193429
age:regionsoutheast	6.616e-03	1.865e-03	3.548	0.000402 ***
age:regionsouthwest	7.050e-03	1.930e-03	3.652	0.000271 ***
smokeryes:regionnorthwest	1.435e-01	6.895e-02	2.080	0.037679 *
smokeryes:regionsoutheast	2.057e-01	6.359e-02	3.235	0.001248 **
smokeryes:regionsouthwest	2.337e-01	7.042e-02	3.318	0.000932 ***
smokeryes:sexmale	4.012e-01	1.378e-01	2.911	0.003661 **
age:smokeryes:factorChildren1	4.281e-03	4.620e-03	0.927	0.354273
age:smokeryes:factorChildren2	1.462e-02	4.947e-03	2.956	0.003174 **

```
age:smokeryes:factorChildren3         1.394e-02  5.634e-03   2.475  0.013469 *
age:smokeryes:factorChildrenOther  8.198e-03  2.499e-02   0.328  0.742888
age:smokeryes:sexmale                -7.153e-03  3.343e-03  -2.140  0.032566 *

Residual standard error: 0.3386 on 1298 degrees of freedom
Multiple R-squared:  0.8679,  Adjusted R-squared:  0.8644
F-statistic: 243.7 on 35 and 1298 DF,  p-value: < 2.2e-16
```

The model with all interaction terms is giving a good Adjusted R-squared. However, adding too many terms is making the model hard to interpret. It is also against the principle of parsimony which states that the simpler the viable model, the better.

Final Model

The model can be made leaner by removing all the double interaction terms, *sex* and its interaction terms, and the interaction terms of *Region* with *Age* and *Smoker*. I have also removed the interaction term of *Children* with *Age* as it doesn't improve the Adjusted R-squared significantly.

```
lm6<- lm(logcharges~ age+age.2+bminew+factorChildren+smoker+region
+bminew*smoker + smoker*factorChildren +age*smoker, data = medCost2)
summary(lm6)

Coefficients:
                                Estimate      Std. Error    t value     Pr(>|t|)
(Intercept)                   6.709e+00    8.993e-02   74.603     < 2e-16 ***
age                             6.106e-02  4.950e-03   12.336      < 2e-16 ***
age.2                          -2.337e-04  6.163e-05   -3.792     0.000156 ***
bminew                        -2.722e-03  2.238e-02  -0.122      0.903239
factorChildren1            1.779e-01  2.873e-02     6.193      7.87e-10 ***
factorChildren2             3.402e-01  3.230e-02  10.534     < 2e-16 ***
factorChildren3             3.117e-01  3.763e-02   8.284     2.89e-16 ***
factorChildrenOther         5.443e-01  6.003e-02   9.067      < 2e-16 ***
```

smokeryes	2.646e+00	7.796e-02	33.936	< 2e-16 ***
regionnorthwest	-5.392e-02	2.811e-02	-1.918	0.055290 .
regionsoutheast	-1.246e-01	2.768e-02	-4.502	7.33e-06 ***
regionsouthwest	-1.438e-01	2.822e-02	-5.096	3.97e-07 ***
bminew:smokeryes	7.025e-01	4.868e-02	14.431	< 2e-16 ***
factorChildren1:smokeryes	-2.093e-01	6.311e-02	-3.317	0.000935 ***
factorChildren2:smokeryes	-3.663e-01	6.641e-02	-5.517	4.15e-08 ***
factorChildren3:smokeryes	-3.109e-01	7.585e-02	-4.099	4.40e-05 ***
factorChildrenOther:smokeryes	-5.370e-01	1.914e-01	-2.806	0.005084 **
age:smokeryes	-3.332e-02	1.749e-03	-19.050	< 2e-16 ***

Residual standard error: 0.3558 on 1316 degrees of freedom
Multiple R-squared: 0.8521, Adjusted R-squared: 0.8502
F-statistic: 446.1 on 17 and 1316 DF, p-value: < 2.2e-16

This model is a lot better than the previous one as it has reduced interaction terms and the adjusted R-squared has reduced only by about 1% which is a reasonable sacrifice made for a leaner model.

APPENDIX 3 : ANALYSIS OF CHARGES

A. Charges ~ Age

Change in the percentage of median charges at two different ages- 20 and 60 years for both smokers and non-smokers have been found as follows:

$Log(Charges|Age= A+1) - Log(Charges|Age= A) = ß_1(A+1-A)+ß_2*((A+1)^2-A^2)+ß_9*smoker*(A+1-A)$

$= ß_1+ß_2+ß_9*Smoker+2*A*ß_2$

$=> Charges|Age= A+1 = (Charges|Age= A)*exp(ß_1+ß_2+ß_9*smoker+2*ß_2*age)$

$=> \%\ Change\ in\ Charges\ per\ year\ change = 100*(exp(ß_1+ß_2+ß_9*smoker+2*ß_2*age)-1)$

```
ages<- c(20,60)
k1 <-lm6$coefficients["age"]
k2<-lm6$coefficients["age.2"]
k3<-(lm6$coefficients["age:smokeryes"])

100*(exp(k1+k2+k3+2*k2*ages)-1) #smoker % increase
 1.82,  -0.05%

100*(exp(k1+k2+2*k2*ages)-1 )      #non-smoker % increase
 5.28    3.33%
```

Confidence intervals for the percentage change in median charges:

```
100*(exp(kConf1+kConf2+kConf3+2*kConf2*20)-1) #smoker % increase
          2.5 %         97.5 %
         0.007          3.69
100*(exp(kConf1+kConf2+2*kConf2*20)-1)           #non-smoker % increase
          2.5 %         97.5 %
          3.75           6.84
100*(exp(kConf1+kConf2+kConf3+2*kConf2*60)-1) #smoker % increase
          2.5 %         97.5 %
         -2.79           2.76
```

```
100*(exp(kConf1+kConf2+2*kConf2*60)-1 )#non-smoker % increase
          2.5 %        97.5 %
          0.85          5.88
```

At what point does the percentage change in median *Charges* due to a unit change in *Age* turn negative?

For smokers = (k1+k2+k3)/(-2*k2) ≈ 59 years

For non-smokers= (k1+k2)/(-2*k2) ≈ 130 years

i.e. for smokers, percentage change in median charges starts decreasing with age around 59 years old. But for non smokers, that doesn't happen practically anytime. (The data set has a maximum age of 65 year old)

B. Charges ~ Smoker

Ratio of Median Medical Charges for smokers to that of non-smokers at an age of 50 Years and various combinations of Children/ dependants and BMI levels is found using the following R code :

```
k4 <- lm6$coefficients["smokeryes"]
k5 <- lm6$coefficients["bminew:smokeryes"]
k6 <- lm6$coefficients["age:smokeryes"]
k7 <- lm6$coefficients["factorChildren1:smokeryes"]
k8 <- lm6$coefficients["factorChildren2:smokeryes"]
k9 <- lm6$coefficients["factorChildren3:smokeryes"]
k10 <- lm6$coefficients["factorChildrenOther:smokeryes"]

bmis <-c(0,1)  # 2 BMI Levels
childrenV <- c(0,k7,k8,k9,k10) # Coeff. of interaction terms of children with smoker
```

```
for (bmiValue in bmis)
{
    for(childValue in childrenV)
    {
        cat("\n \n bmi Value :", bmiValue)
        cat(" ; children :", match(childValue,childrenV) -1)
       cat(" ; Ratio:",(round(exp(k4+k5*bmiValue+k6*50+childValue),2)))
    }
}
```

Output:

bmi Value : 0 ; children : 0 ; Ratio: 2.66

bmi Value : 0 ; children : 1 ; Ratio: 2.16

bmi Value : 0 ; children : 2 ; Ratio: 1.85

bmi Value : 0 ; children : 3 ; Ratio: 1.95

bmi Value : 0 ; children : 4 ; Ratio: 1.56

bmi Value : 1 ; children : 0 ; Ratio: 5.38

bmi Value : 1 ; children : 1 ; Ratio: 4.36

bmi Value : 1 ; children : 2 ; Ratio: 3.73

bmi Value : 1 ; children : 3 ; Ratio: 3.94

bmi Value : 1 ; children : 4 ; Ratio: 3.14

BIBLIOGRAPHY

1. Choi, Miri. “Medical Cost Personal Datasets.” Kaggle, February 21, 2018. https://www.kaggle.com/mirichoi0218/insurance

2. Lantz, Brett. “Forecasting Numeric Data - Regression Methods.”Chapter. In *Machine Learning with R: Learn How to Use R to Apply Powerful Machine Learning Methods and Gain an Insight into Real-World Applications*, 173. Packt Publ., 2015.

3. Markowski, Carol A., and Edward P. Markowski. “Conditions for the Effectiveness of a Preliminary Test of Variance.” *The American Statistician* 44, no. 4 (1990): 322. https://doi.org/10.2307/2684360.

4. Ramsey, Fred L., and Daniel W. Schafer. *The Statistical Sleuth: A Course in Methods of Data Analysis*. Australia: Brooks/Cole, Cengage Learning, 2013.