AGENO SCHOOL OF BUSINESS

Golden Gate University

MSBA 320- Advanced Statistical Analysis with R & Python

Dr. Siamak Zadeh

Summer 2020

Statistical Analysis on Factors Affecting Medical Expenses

Fig:1

Reena Sehitya
SID: 0597218

Abstract

The aim of this final project is to explore factors that affect long-lasting medical costs, the costs that constitute a recurring portion of expenses of any individual's life. I chose this topic for my final project as I was interested in understanding what factors might lead to an increase in the insurance or medical costs for any individual. Since, every individual has to incur medical costs at some point of their life, the analysis of the factors that led to an increase in medical expenses can highlights what those factors are and individual's can be aware of them so that he/she can avoid paying high medical expenses. The factors I will analyze in this project to see if they are significantly related to medical expenses or not are age, BMI, children, gender, smoker/non-smoker, and region. In order to look for the answer to the above questions, I will work on a dataset "insurance.csv" downloaded from "Kaggle.com". In the first part, I will run descriptive analysis on each factor such as age, BMI, children, gender, smoker/non-smoker, region and medical expenses as well as how they have related each other. In the second part, I will run regression analysis and ANOVA tests to analyze if any of these factors significantly impact medical expenses for an indivisual.

TABLE OF CONTENT

Introduction

Health insurance premiums have been increasing significantly in the United States over the last sixty years. This rising medical expenses had led to many people in the U.S. leading a life without medical coverage and as a result end up paying higher medical bills when they fall sick. The United States as a nation spends a major portion of its budget on healthcare as compared to other nations and this continues to be growing significantly. According to the Peter G. Peterson foundation report, U.S has spent around $3.6 trillion on healthcare in the year 2018, which is around $ 11,000 per person. The healthcare costs in the United States were 5% of gross domestic product (GDP) in the year 1960 and that costs had risen to 18% of gross domestic product (GDP) in the year 2018. (Peter G. Peterson foundation, 2020). As per the Centers for Medicare and Medicaid Services (CMS), medical costs are expected to grow at an average annual rate of 5.4 percent for 2019-28 and will reach $6.2 trillion by 2028. These rising medical expenses will negatively be impacting the overall debt of this country. This can also lead to public health crisis in the US in the near future.

According to the Peter G. Peterson foundation report, aging population of the United States is one of the prime reasons for the rising medical costs. As people grow older, their medical expenses usually tend to increase. U.S. Census Bureau predicted that population of people in the age group of 65 and above, is expected to grow from 16% to 20% of the total population by the year 2030. As per the Sterling Price (2020), medical expenses vary depending upon which state or county a person lives in the U.S. According to the study conducted by ValuePenguin, monthly medical costs are highest in the states like Alaska, Wyoming, New Jersey, and Florida. The state that has the lowest monthly medical expenses is Utah (the Sterling Price, 2020). People who smoke, can pay up to 50% higher medical insurance costs than a non-

smoker (the Sterling Price, 2020). Number of people covered by any insurance plan also affects the monthly insurance costs. For example, a married couple with kid will pay higher insurance premium than an individual (the Sterling Price, 2020). University of Illinois kinesiology and community health professor Ruopeng An, reports that the healthcare costs for people who smoke and have obesity is about $1,360 and $1,046 per person per year in the year 2015 that was way above average annual health care expenditures of non-obese and nonsmoking Americans.

According to the University of Illinois at Urbana-Champaign (2015), health care costs are higher for smokers and people with higher BMI and this additional cost are higher for women. According to the Centers for Medicare & Medicaid Services, women in the U.S. women has spent $1.4 trillion on personal healthcare in the year 2014. Overall, women have been spending more than men in every age groups with exception from the age 0-18. As per RegisteredNursing.org, medical expenditure of women is 84% higher than men for the age range of 18 to 44. For the age range of 44 to 64, medical expenses of women are 24% higher than that for men. Health care expenses also vary by race in the United States. For instance, medical expenses for white people is higher than Native American followed by Asian/Pacific Islander, Black, Hispanic for the age group of 18 to 44 (R. Writers, 2020).

In this final paper, I will use my dataset that has all these variables such as age, gender, smoking, children, region, and medical expenses. I will analyze each of these variables to see how these variables are distributed and related to each other. Through hypothesis testing, I will try to find if each of these factors are contributing towards increasing or decreasing medical expenses. The goal of this project is statistically analyzing how these factors such as age, gender, smoking, children, region and BMI are significantly related to medical expenses.

Definition of Insurance Premium Prediction Dataset

The data set that I will be using for the term project is “Insurance Premium Prediction”. It is a publicly available dataset on Kaggle.com and originally obtained from GitHub. The data set is very clean, and I did not need to do any modification on the data. The data set is in CSV format with 7 columns and 1338 unique records. Columns or attributes are sub-divided into continuous variables and categorical variables. Among 7 attributes, “Medical Expenses” is my dependent/target variable and rest of the features/attributes will be considered as independent factors.

Continuous variables

Continuous variables include age, BMI, children and expenses. The definition of all four continuous variables are given below:

- Age: This attribute refers to the age of each insured person considered in this dataset
- BMI: BMI stands for body mass index. This attribute gives idea about weights of each insured person considered in this dataset, relative to their height
- Children: This attribute refers to the number of children of each insured or the number of dependents on the person insured considered in this dataset
- Expenses: I renamed this attribute as Medical Expenses to give the context that what kind of data that attribute holds. This attribute refers to the medical expenses borne by each person considered in the dataset. This is the expenses that is billed by Insurance Companies for that insured.

Categorical variables

Categorical variables include sex, smoker and region. The definitions of all categorical variables are given below:

- Sex: I renamed this attribute as gender. This attribute is sub-categorized into two level i.e. insured male and female.
- Smoker: This attribute is also sub-categorized into two level to indicate one who smokes and one who doesn't smoke.
- Region: This attribute includes four different regions of the United States such as southwest, southeast, northeast, and northwest where the insured live.

These are the attributes I will explore to answer how the explanatory variables such as gender, age, BMI, region, smoker, and children are impacting our response variable "medical expenses" using python jupyter notebook editor. As of now, I don't see that I need any secondary datasets to supplement the data set that I am going to use for my final project.

Exploratory data analysis of the Insurance Premium Prediction

In this chapter, I am going to run descriptive analysis first on each attribute including both predicator variables as well as response variable to get to know their central tendency, variability and distribution. My primary focus here is exploring the data and identifying trends as well as relationships of various attributes with medical expenses. I will also explore how each of these attributes are related to each other. The entirety of this analysis run be done in python jupyter notebook.

To begin, I have uploaded the Insurance.csv dataset in the jupyter notebook kernel. After uploading the dataset, I run the code to check if there's any missing or NAN values in the data-frame. Fortunately, there is no missing values in my dataset. I also ran the code to know total number of columns and rows and if all attributes are identified as a correct data type. The goal here is to confirm if my data is clean and in good shape or not.

Next, I have examined the central tendency and variability of each continuous variable.

Attribute	count	mean	std	min	25%	50%	75%	max
age	1338.00	39.21	14.05	18.00	27.00	39.00	51.00	64.00
bmi	1338.00	30.66	6.10	15.96	26.30	30.40	34.69	53.13
children	1338.00	1.09	1.21	0.00	0.00	1.00	2.00	5.00
MedicalExpenses	1338.00	13270.42	12110.01	1121.87	4740.29	9382.03	16639.91	63770.43

Table 1: Description of continuous variables

From the above table, it is clear that maximum age of insured in my dataset is 64 and median and mean age is more or less same at 39. If we look at BMI, there is a huge difference between minimum and maximum BMI. BMI at 50 and above considered to be extremely obesity condition as per Marina del Rey Hospital. This shows that variability in BMI factor is very high. In case of children, mean and standard deviation is not giving much information about the behavior of this factor, since the maximum number of children can be 5 and the least is 0. Medical expenses as our response variable is showing much difference between minimum and maximum medical charges charged by Insurance companies. Through visualization it will be clearer about the outliers, if present in medical expenses.

The description of all categorical variables is shown below:

Attribute	gender	smoker	region
count	1338	1338	1338
unique	2	2	4
top	male	no	southeast
freq	676	1064	364

Table 2: Description of categorical variables

The above table gives us a brief idea of top frequency of each categorical attribute. Number of males is higher than number of females i.e. male- 676 and female- 662. But the difference between the number of male and female is not much. Number of non-smoker (1064) is much higher than number of smoker (274). There are 364 people from Southeast region which is the highest among the four different regions of the U.S.

In order to check symmetry and density in the tail regions of all Numerical variables of my data-frame, I run the kurtosis and skewness functions on them.

As per *NIST/SEMATECH e-Handbook of Statistical Methods*-

"Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point"

"Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers".

The kurtosis of all Numeric columns:	**Value**
Excess kurtosis of "Medical Expenses" (should be 0):	1.59582
Excess kurtosis of "bmi" (should be 0):	-0.055
Excess kurtosis of "children" (should be 0):	0.19722
Excess kurtosis of "age" (should be 0):	-1.2449
Skewness of Numeric columns:	
skewness of "Medical Expenses" (should be 0):	1.51418
skewness of "bmi" (should be 0):	0.28373
skewness of "children" (should be 0):	0.93733
skewness of "age" (should be 0):	0.05561

Table 3: Kurtosis and skewness of Numerical variables

The above table shows that excess kurtosis for "medical expenses" (response variable) – it is greater than 0 that means more outliers are present in the tail region of the distribution. Statistical term for this type of distribution is Leptokurtic. Whereas, excess kurtosis of "bmi" and "children" is close to zero that gives us a sense that overall distribution of these two variables are

similar to normal distribution. On the other hand, excess kurtosis of age is lower than zero that shows lack of outliers in the tail region of the distribution. From skewness point of view, it is pretty clear that “medical expenses” is heavily right skewed. Other variables such as bmi and age is symmetrical i.e. close to zero. I am skeptical about the skewness of “children”.

Visualization of numerical variables such as medical expenses, BMI, children and age gives us better understanding of how each of these variables are distributed.

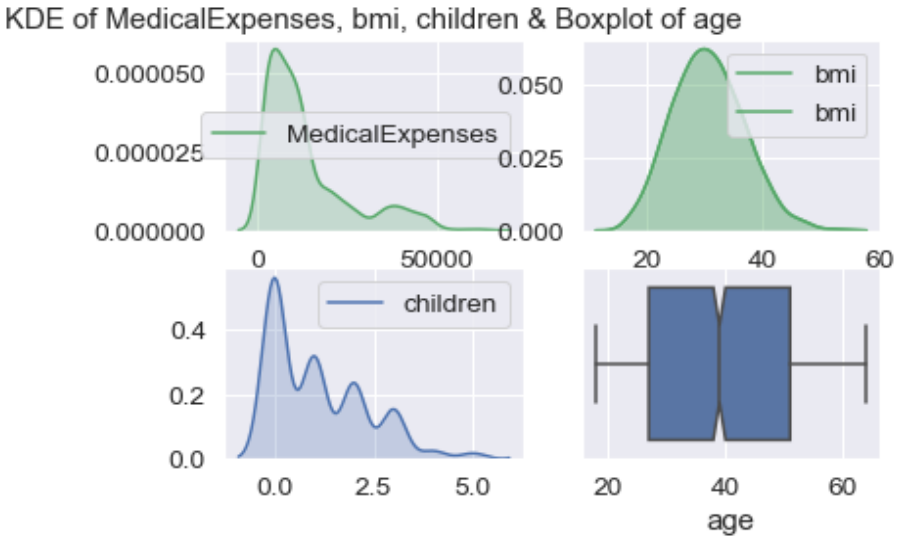


Fig:2

The Kernel density estimation plot of "MedicalExpenses" and "children" show that these variables are right skewed whereas KDE of "bmi" clearly shows that it is normally distributed. The boxplot of "age" shows that it doesn't have skewness, with no outliers and fairly distributed. The Kernel density estimation in the short form KDE, is the non-parametric way of smoothening the curve to get to know the shape of the data. It is used in place of discrete histogram (Matthew Conlen, n.d.).

Since, the distribution of my response variable “MedicalExpenses” is right skewed and leptokurtic. I applied log to this variable to see if there are any changes in its distribution. After

transforming the variable through applying log, I found KDE of "MedicalExpenses" has transformed into more or less symmetrical.

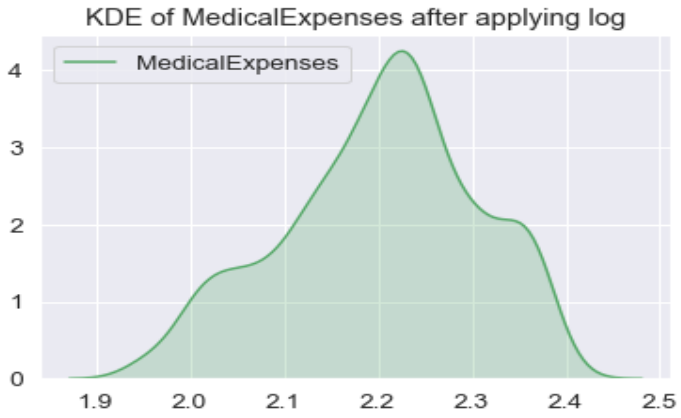


Fig:3

The above Kernel density estimation plot of "MedicalExpenses" has transformed into as close to as symmetrical in shape after applying logarithm to this entire variable. Now there are lesser number of outliers are in the tail regions. Transforming the response variable might also affect outcome of regression analysis that we will see later in our analysis.

Categorical variable analysis

Analysis of categorical columns 'gender', 'smoker', 'region' with respect to Medical Expenses (response variable). The aim of this analysis is to find out how categorical variables are related to medical expenses.

1) Medical expenses vs gender.

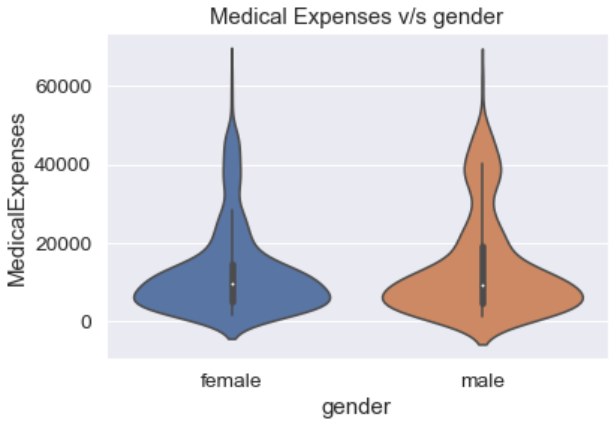


Fig: 4

Takeaway: The above violin plots are giving us the distribution of medical expenses for both male and female. By looking at the graph, it is clear that overall distribution of medical expenses for both male and female are more or less the same with majority of people having medical bills around 10k. Towards the tail, it seems that proportion of male are higher than female that means more number of male than female have medical expenses in the range from 35k to 50k. This finding contradicts with the literature review where it is stated that from age group 18-44, medical expenses for female exceeds that of the male. I will run one-way ANOVA analysis to see if there are no mean differences in the medical expenses between male and female.

2) Medical Expenses vs smoker

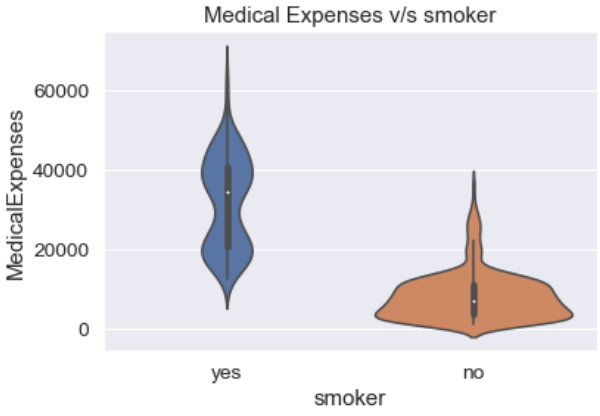


Fig: 5

Takeaway: The above violin plots are showing us the distribution of medical expenses for smokers and non-smokers. By looking at the graph, we can see that the distribution of medical expenses for smokers is bimodal that means higher proportion of smokers have medical expenses around 20k and 40k. It is pretty apparent by seeing the distributions that the medical expenses for non-smokers are way lower than that of smokers. If we look at the tail regions for both smokers and non-smokers, we can easily see that the medical expenses of non-smokers are around 40k whereas the medical expenses for smokers are lying above 70k. Boxplot inside the graph clearly shows the huge difference between the mean/median medical expenses between smokers and non-smokers.

3) Medical Expenses vs region

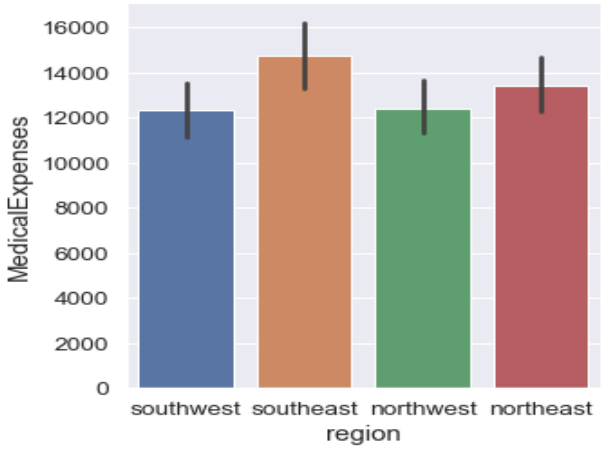


Fig: 6

Takeaway: The above violin plots are the distribution of medical expenses for different regions in the U.S.- southwest, southeast, northwest, northeast. By looking at the graph, we can see that the distribution of medical expenses for southeast region is the highest followed by northwest, northeast and southwest. I need to dig further deeper to find out the reason behind higher medical expenses in the southeast region.

4) Medical Expenses vs gender with smoker attribute

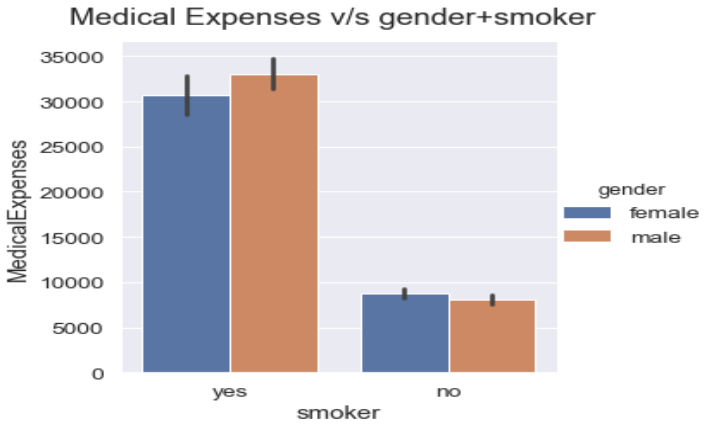


Fig:7

Takeaway: The above bar plots are clearly showing that number of male smokers are more as compared to number of female smokers. As a result, total medical expenses are also higher for male than female. In order to verify this, I will run ANOVA test to see if the mean medical expenses are different between male smokers and female smokers or not. This explains the reason the reason for medical expense for males being higher than females in the "Medical expenses vs gender" analysis.

5) Medical Expenses vs four different regions of the United States and smoker attribute

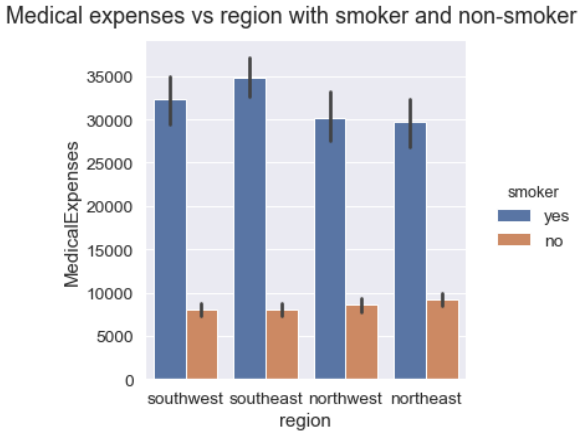


Fig: 8

Takeaway: Earlier violin plots give us a picture that the distribution of medical expenses for southeast region is the highest followed by northwest, northeast and southwest. The above bar plots are showing that number of smokers are also the highest in southeast region that kind of gives us a hint that smoking might be a reason for the highest medical expenses in southeast region. Southwest is the region where there are second highest number of smokers as well as medical expenses associated with that region. However, as per violin plot, southwest was the last region in terms of medical expenses. Rest other orders with northwest and northeast are matching in both box plots and violin plots. To get to the conclusion if smoking lead to increase in the medical expenses in any of the four regions, I will run ANOVA to see if the mean medical expenses are different among the four regions are not.

Continuous variable analysis

Now comparing the numerical_columns such as age, bmi with our response variable i.e. "Medical Expenses". Since, in this case both explainatory and response variables are continuous, a scatter plots usually give better understanding of the relationships.

1) Medical expenses vs BMI

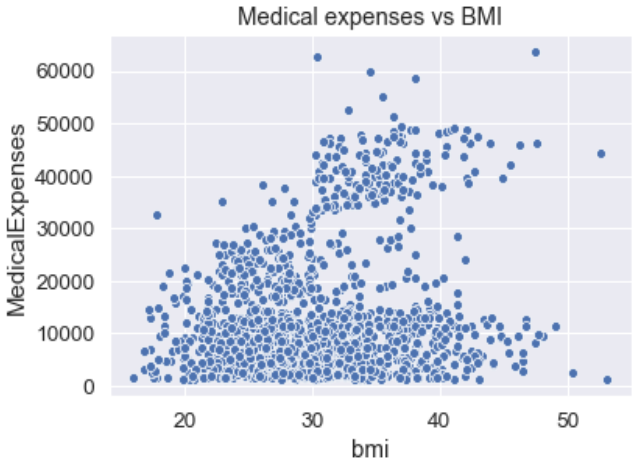


Fig: 8

Takeaway: The above scatterplot gives us another big picture associated with higher medical expenses i.e. people with higher BMI typically also tend to pay more medical bills or expenses. In other words, BMI and medical expenses are positively correlated. To confirm if this relationship holds true, we add a regression line, or the line that best fits the data.

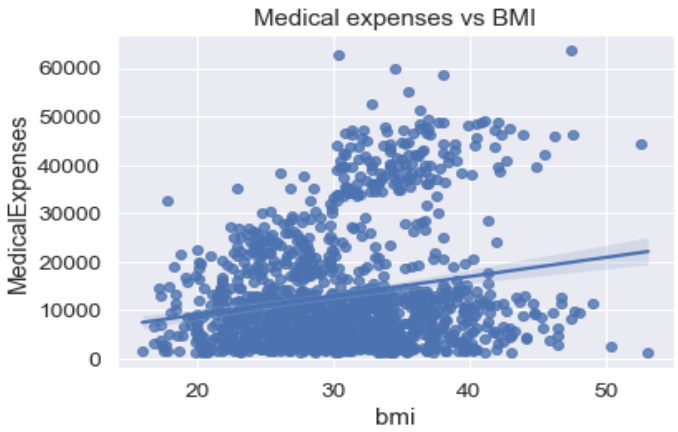


Fig: 9

Takeaway: The above scatterplot with regression line suggests that medical expenses increases with higher BMI as people with higher BMI are more prone to chronic disease. Let's, dig further to check how smoker and non-smoker along with BMI are related to medical expenses.

2) Medical expenses vs BMI and non-smoker

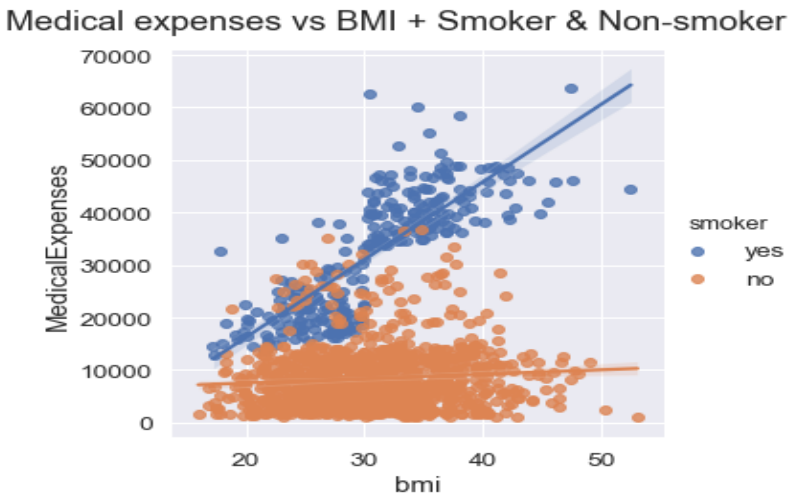


Fig: 10

Takeaway: The slope of regression line of smoker is much steeper than the slope of the regression line of non-smoker with a given BMI. This clearly suggests that at a given level of BMI, smokers end up paying more medical expenses that non-smokers.

3) Medical expenses vs age

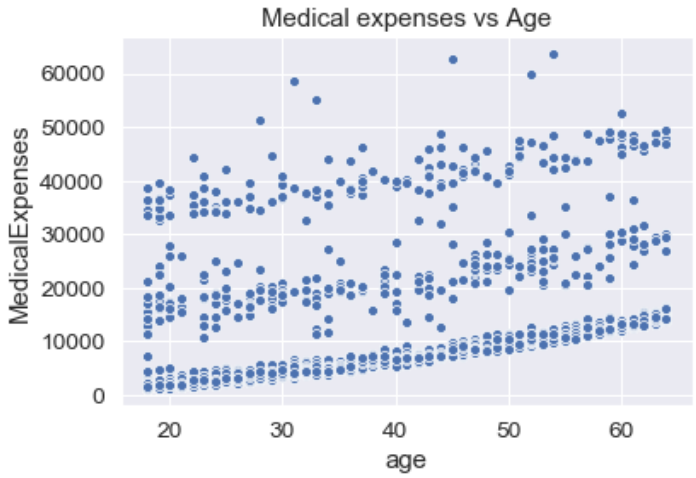


Fig: 11

Takeaway: The above scatter plot suggests that medical expenses increases with the people become older. In other word, medical expenses are positively correlated with age factor. To confirm if this relationship is really true, we add regression line that passes through the data.

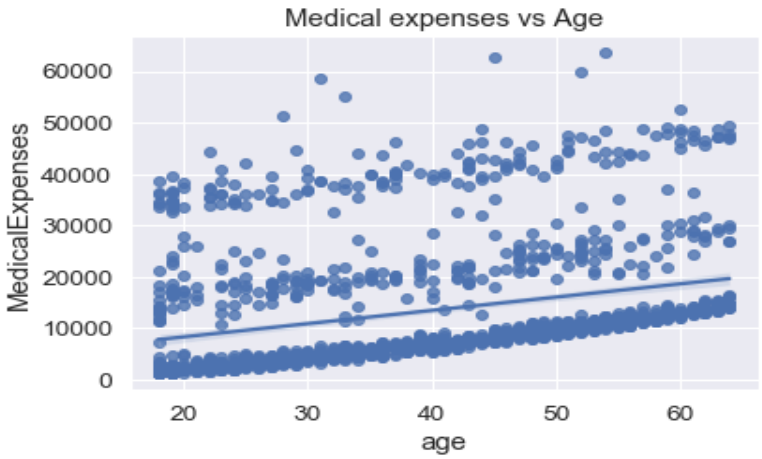


Fig: 12

Takeaway: The above scatterplot along the regression line passing through the data suggest that medical expenses are positively related to age. On an average, as people grow older, medical bills also goes up. In the plot, we can see three layers of scatterplots that is indicating might be there are three factors forming these three classifiers. If I had a variable with three classifiers, I would have got better picture as what factors are involved in forming three layers of medical expenses.

4) Medical expenses vs Age & Smoker:

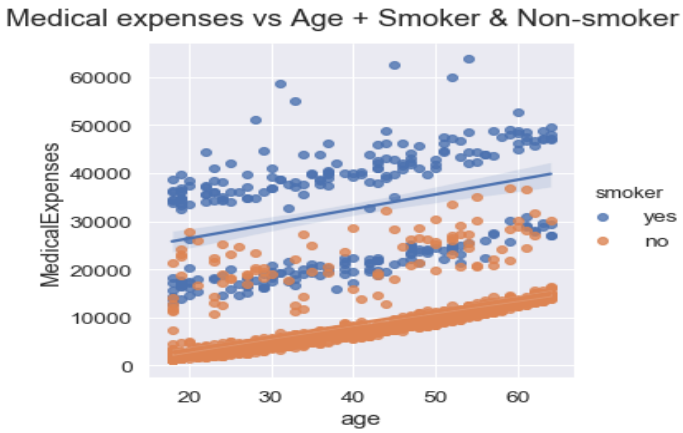


Fig: 13

Takeaway: The above scatterplot with regression lines are indicating that smokers pay more medical bills as they grow older as compare to non-smokers. Let's see if there's any differences in medical expenses between male and female with respect to age. But three layers in the figure above is due to some factors that is missing in my dataset.

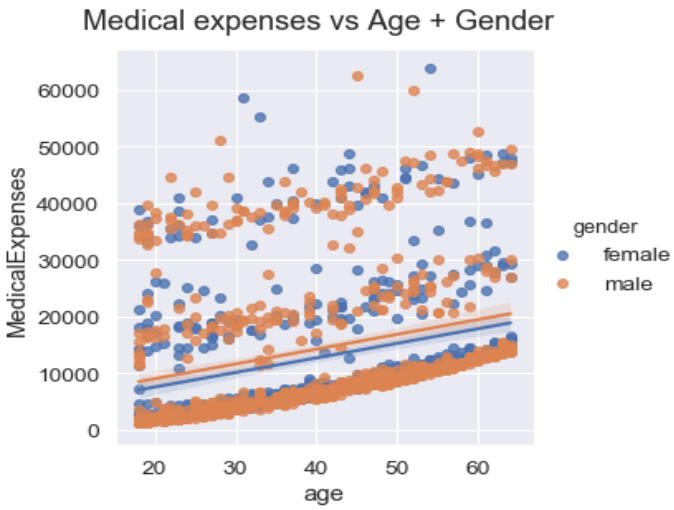


Fig: 14

Takeaway: The above scatterplot with regression lines doesn't show any difference in medical expenses between male and female with respect to their age.

5) Correlation among the continuous variables

	age	gender	bmi	children	smoker	region	Medical Expenses
age	1.0000	0.0209	0.1093	0.0425	0.0250	0.0063	**0.2990**
gender	0.0209	1.0000	-0.0464	-0.0172	0.0762	0.0128	-0.0573
bmi	0.1093	-0.0464	1.0000	0.0128	-0.0038	**-0.2628**	**0.1983**
children	0.0425	-0.0172	0.0128	1.0000	-0.0077	0.0012	0.0680
smoker	0.0250	0.0762	-0.0038	-0.0077	1.0000	0.0389	**-0.7873**
region	0.0063	0.0128	**-0.2628**	0.0012	0.0389	1.0000	-0.0395
Medical Expenses	**0.2990**	-0.0573	**0.1983**	0.0680	**-0.7873**	-0.0395	1.0000

Table: 4 Correlation Matrices of all the attributes

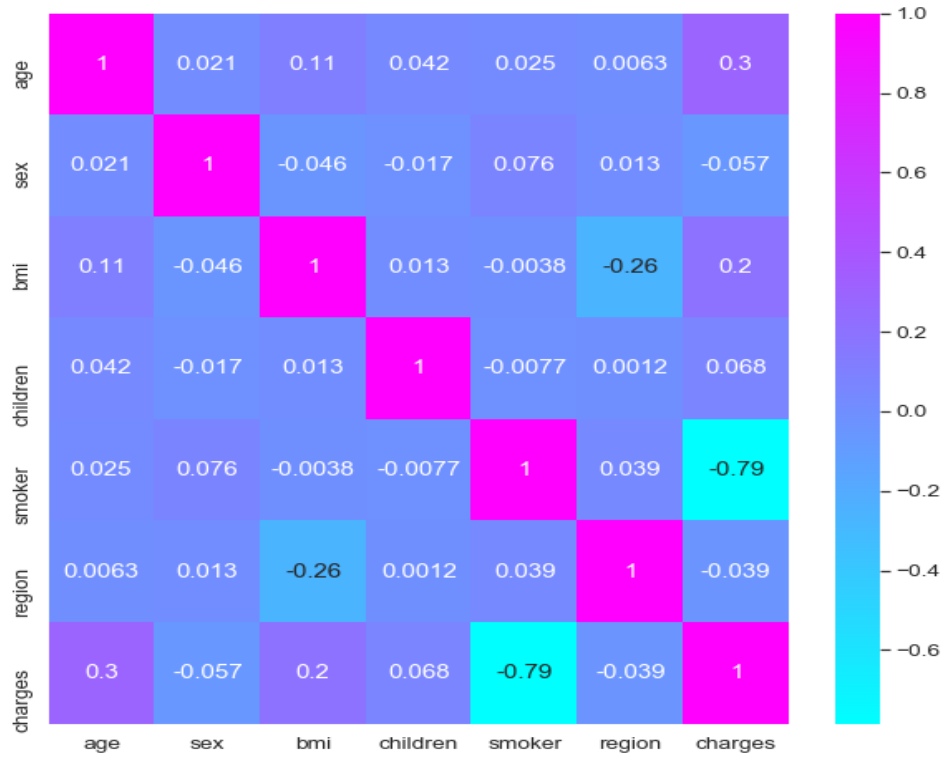


Fig: 15

Takeaway: Correlation matrix and correlation heatmap with correlation coefficients on it shows how all variables are related to each other as well as to the response variable i.e. Medical Expenses. Both matrix and heatmap suggest that age (0.3) and BMI (0.2) are positively correlated to Medical Expenses. To check how significant these factors are in terms of predicting response variable, I will run **regression analysis and ANOVA test** in the next section. Contradicting feature is smoker attribute is negatively correlated to charges i.e. "Medical Expenses". Although, through our descriptive analysis, we got the understanding that smokers have to pay higher medical expenses than non-smokers. Another striking feature is region is negatively correlated with BMI. To verify this, I will run **ANALYSIS OF COVARIANCE** to see if there is a mean difference in medical expenses for different bmi at four different regions. For rest other variables, I don't see any strong sign of collinearity.

Analysis and interpretation of the results

In this chapter, I will run one-way ANOVA analysis in the first phase to see if there are mean differences in the medical expenses between male and female, then between smoker and non-smoker, and then among four regions of the United States. Then I will run two-way ANOVA analysis tests to see if there are mean differences in medical expenses in two factors such as smoker with gender, smoker with region, and region with bmi. Then I will run the regression models one with log transform medical expenses with all other predictors and one without transformation with respect to all the predictors. I will see if all the predictors are significantly related to medical expenses (response variable) or not. I will also run ANOVA between two regression models to see which model does the best job in predicting variability of our response variable i.e. medical expenses.

ANOVA analysis

Model 1: One-way ANOVA test between Medical expenses vs smoker

The very **first model** is based on one-way ANOVA to see if there is any mean difference in medical expenses between smoker and non-smoker in the United States. So, my null and alternative hypotheses are:

H0: Smokers and Non-smokers have equal mean medical expenses

H1: Smokers and Non-smokers have not equal mean medical expenses

Table: 5 *One-way ANOVA between Medical expenses vs smoker*

	sum_sq	df	F	PR(>F)
smoker	5.7252	1.0	**917.7264**	**0.0**
Residual	8.3346	1336.0	NaN	NaN

Takeaway: The above table shows that the p-value of the F-statistic is equal to 0 that is **p < 0.05**. Thus, we reject null hypotheses and conclude that the mean medical expenses are not equal for smokers and non-smokers. In other words, the amount of medical costs charge by the Insurance companies significantly depend on whether insured is smoker or not.

Model 2: One-way ANOVA between Medical expenses vs region

My **second model** is also based on one-way ANOVA to see if there is any mean difference in medical expenses among four regions in the United States. So, null and alternative hypotheses are:

H0: Regions have equal mean medical expenses

H1: Regions have not equal mean medical expenses

Table: 6 *One-way ANOVA between Medical expenses vs region*

	sum_sq	df	F	PR(>F)
region	0.0440	3.0	**1.3975**	**0.2419**
Residual	14.0157	1334.0	NaN	NaN

Takeaway: The above table shows that the p-value of the F-statistic is greater than 0 (PR(>F) = 0.2419) i.e. **p > 0.05**. Thus, I failed to reject null hypotheses and conclude that I have insufficient evidence to accept that there are no equal mean medical expenses among four regions of the United States. This is surprisingly contradicting what we found in literature reviews that says that medical charges do vary across different states and counties in the United States such as Alaska is a state that charges the highest medical costs and Utah charges the lowest medical costs.

Model 3: One-way ANOVA between Medical expenses vs gender

My **third model** is again based on one-way ANOVA tests. This time I want to see if there is any mean difference in medical expenses between male and female in the United States. So, my null and alternative hypotheses are:

H0: Gender (male/female) have equal mean medical expenses

H1: Gender (male/female) do not have equal mean medical expenses

Table: 7 *One-way ANOVA between Medical expenses vs gender*

	sum_sq	df	F	PR(>F)
gender	0.000089	1.0	**0.008471**	**0.926679**
Residual	14.059679	1336.0	NaN	NaN

Takeaway: The above table shows that the p-value of the F-statistic is greater than 0 (PR(>F) = 0.926679) i.e. **p > 0.05**. Thus, I failed to reject null hypotheses and conclude that I have insufficient evidence to accept mean medical expenses are not equal between male and female in the United States. This is again contradicting what we found in literature reviews that says that medical charges do vary across between gender in the United States. Women of the age group 18-44, generally do pay higher medical costs than men.

Model 4: Two-way ANOVA - MedicalExpenses ~ C(smoker) + C(region)+ C(region):C(smoker)

My **fourth model** is based on two-way ANOVA tests. In this model, I am considering the **interaction** between the two factors as well. This time I want to see if there is any mean difference in medical expenses between smoker and non-smoker based in four regions in the United States. So, my null and alternative hypotheses are:

H0: Smokers and non-smokers have equal mean medical expense

Regions have equal mean medical expense

Regions and smokers have no interactions.

H1: Smokers and non-smokers have no equal mean medical expense

Regions have no equal mean medical expense

Regions and smokers have interactions

Table: 8 *Two-way ANOVA using Statsmodels- MedicalExpenses ~ C(smoker) + C(region)+ C(region):C(smoker)*

	sum_sq	df	F	PR(>F)
C(smoker)	5.7206	1.0	**923.5863**	**0.0000**
C(region)	0.0395	3.0	**2.1257**	**0.0952**
C(region):C(smoker)	0.0571	3.0	**3.0744**	**0.0268**
Residual	8.2379	1330.0	NaN	NaN

Takeaway: The above table shows that the p-value of the F-statistic of smoker is equal to 0 that is **p < 0.05**, for regions is **0.0952,** and interaction term C(region):C(smoker) is **0.0268**. The above result shows that smoker is significantly related to medical expenses. Region, if run along with factor "smoker", is close to significance level (less than -$alpha$ = 0.1). The PR(>F) of 0.0268 i.e. **p < 0.05,** shows that the interaction between these two factors is significant. In other word, medical expenses do vary based on smoker belonging to a particular region of the United States. Probably, additional variable listing states would have given clearer results than region.

Model 5: Two-way ANOVA using Statsmodels- MedicalExpenses ~ C(smoker) + C(gender) + C(smoker):C(gender)

My **fifth model** is also based on two-way ANOVA tests. In this model, I am considering the **interaction** between the two factors such as smoker and gender. This time I want to see if there is any mean difference in medical expenses between smoker and non-smoker based on gender. So, my null and alternative hypotheses are:

H0: Smokers and non-smokers have equal mean medical expense

Gender have equal mean medical expense

Regions and gender have no interactions.

H1: Smokers and non-smokers have no equal mean medical expense

Gender have no equal mean medical expense

Regions and gender have interactions

Table: 9 *Two-way ANOVA using Statsmodels- MedicalExpenses ~ C(smoker) + C(gender)+ C(smoker):C(gender)*

	sum_sq	df	F	PR(>F)
C(smoker)	5.7621	1.0	929.7577	0.0000
C(gender)	0.0370	1.0	5.9664	0.0147
C(smoker):C(gender)	0.0303	1.0	4.8837	0.0273
Residual	8.2673	1334.0	NaN	NaN

Takeaway: The above table shows that the p-value of the F-statistic of smoker and gender is close to 0 that is **p < 0.05**. The interaction term C(smoker):C(gender) is **0.0273 i.e. it is < 0.05**. The above result shows that smoker, gender and interaction between the two are significantly

related to medical expenses. Here, we reject null hypotheses. In the other words, medical expenses do differ depending on smoker is male or female.

Model 6: One-way ANOVA - MedicalExpenses ~ C(region) + C(bmi)

My **sixth model** is based on one-way ANCOVA tests. I will run one-way ANCOVA tests as one of the explanatory variables i.e. bmi is continuous and another explanatory variable i.e. region is categorical variable. That meet the condition of ANCOVA model (M.J. Crawley, 2008). So, my null and alternative hypotheses are:

H0: probability that there's no effect or relationship between two factors region and bmi

H1: probability that there's an effect or relationship between two factors region and bmi

Table: 10 *Two-way ANCOVA - MedicalExpenses ~ C(region) + C(bmi)*

	Source	SS	DF	F	p-unc	n2
0	region	0.070589	3	**2.277823**	**0.077911**	0.005011
1	bmi	0.246014	1	**23.815799**	**0.000001**	0.017465
2	Residual	13.769705	1333	NaN	NaN	NaN

Takeaway: The above table shows that the uncorrelated p-value of region is close to **0.05** and that of bmi is 0. F-statistics of region is low that means region is insignificant to the amount of medical costs charges by insurance companies. Whereas the F-statistics of bmi shows that mean medical expenses significantly related to bmi as a factor. In other word, insurance companies do charges higher medical costs to person with higher bmi and vice versa.

Regression analysis

In the second phase of this chapter, I will run multiple regression analysis to see how all the explanatory variables including categorical and continuous related to the response variable i.e. medical expenses. In this phase, I will run two model, one with log transformation of response variable and one without log transformation. Then I will compare the two model to see which one works the best. Before running multiple regression, I run the one-hot encoding to change all the categorical variable into dummy variables.

Model 1: Log (Medical Expenses) vs Age + BMI+ children + region+ smoker+ gender

Table 11: Log (Medical Expenses) vs Age + BMI+ children + region+ smoker+ gender

OLS Regression Results

Dep. Variable:	MedicalExpenses	R-squared:	**0.759**
Model:	OLS	Adj. R-squared:	**0.757**
Method:	Least Squares	F-statistic:	521.9
Date:	Wed, 12 Aug 2020	Prob (F-statistic):	0.00
Time:	13:00:35	Log-Likelihood:	2099.9
No. Observations:	1338	AIC:	-4182.
Df Residuals:	1329	BIC:	-4135.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.9068	0.004	252.656	0.000	0.900	0.914
age	0.0040	9.92e-05	40.434	0.000	0.004	0.004
I(bmi)	0.0013	0.000	5.642	0.000	0.001	0.002
I(children)	0.0124	0.001	10.817	0.000	0.010	0.015
I(gender_female)	0.4582	0.002	204.128	0.000	0.454	0.463
I(gender_male)	0.4486	0.002	195.703	0.000	0.444	0.453
I(smoker_no)	0.3701	0.002	157.223	0.000	0.365	0.375
I(smoker_yes)	0.5367	0.003	205.339	0.000	0.532	0.542
I(region_northeast)	0.2370	0.003	94.652	0.000	0.232	0.242
I(region_northwest)	0.2295	0.003	91.374	0.000	0.225	0.234
I(region_southeast)	0.2183	0.003	79.171	0.000	0.213	0.224
I(region_southwest)	0.2219	0.003	85.616	0.000	0.217	0.227

Omnibus:	402.465	Durbin-Watson:	2.034
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1360.999
Skew:	1.463	Prob(JB):	2.90e-296
Kurtosis:	6.981	Cond. No.	9.07e+17

Takeaway: The result of the above multiple regression model shows that r-square and adjusted r-square of this model is 0.759 and 0.757 respectively. That means 75.9% and 75.7 % as per r-square and adjusted r-square respectively, of the variability of our response variable is explained by the variabilities of all the independent factors considered in this model. If we look at the p-value (**p < 0.05)** of all the explanatory variables, we will find that all theses variables are significantly related to the medical expenses (response variable). Durbin-Watson statistics measures the autocorrelation of residuals of regression analysis. In this regression model, Durbin-Watson statistics is very close to 2 then i.e. 2.034, that means there is no presence of autocorrelation in the residuals.

Model 2: Medical Expenses vs Age + BMI+ children + region + smoker + gender

Table 12: Medical Expenses vs Age + BMI+ children + region+ smoker+ gender

OLS Regression Results

Dep. Variable:	MedicalExpenses	R-squared:	**0.751**
Model:	OLS	Adj. R-squared:	**0.749**
Method:	Least Squares	F-statistic:	500.8
Date:	Wed, 12 Aug 2020	Prob (F-statistic):	0.00
Time:	13:00:50	Log-Likelihood:	-13548.
No. Observations:	1338	AIC:	2.711e+04
Df Residuals:	1329	BIC:	2.716e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-296.4168	430.507	-0.689	0.491	-1140.964	548.130
age	256.8564	11.899	21.587	0.000	233.514	280.199
I(bmi)	339.1935	28.599	11.860	0.000	283.088	395.298
I(children)	475.5005	137.804	3.451	0.001	205.163	745.838
I(gender_female)	-82.5512	269.226	-0.307	**0.759**	-610.706	445.604
I(gender_male)	-213.8656	274.976	-0.778	**0.437**	-753.299	325.568
I(smoker_no)	-1.207e+04	282.338	-42.759	0.000	-1.26e+04	-1.15e+04
I(smoker_yes)	1.178e+04	313.530	37.560	0.000	1.12e+04	1.24e+04
I(region_northeast)	512.9050	300.348	1.708	0.088	-76.303	1102.113
I(region_northwest)	159.9411	301.334	0.531	**0.596**	-431.201	751.083
I(region_southeast)	-522.1170	330.759	-1.579	0.115	-1170.983	126.749
I(region_southwest)	-447.1459	310.933	-1.438	0.151	-1057.119	162.827

Omnibus:	300.366	Durbin-Watson:	2.088
Prob(Omnibus):	0.000	Jarque-Bera (JB):	718.887
Skew:	1.211	Prob(JB):	7.86e-157
Kurtosis:	5.651	Cond. No.	9.07e+17

Takeaway: The result of above multiple regression model- model 8, shows that r-square and adjusted r-square of this model is 0.751 and 0.749 respectively. That means 75.1% and 74.9 % as per r-square and adjusted r-square, of the variability of our response variable is explained by the variabilities of all the independent factors considered in this model. This model is underperforming as compared to model 1 in terms of explaining the variabilities of the dependent variable. If we look at the p-value ($p < 0.05$) of all the explanatory variables, we will find that all these variables except female, male, and northwest region, are significantly related to the medical expenses (response variable). In this regression model Durbin-Watson statistics is also close to 2 then i.e. 2.088 but not as close to as model 1.

Model 3: ANOVA comparison between regression models 1 and 2

When I compare both the models - model 1 and model 2 through ANOVA analysis test, I got the following result:

Table: 13 ANOVA: model 1 vs model 2

df_resid	ssr	df_diff	ss_diff	F	Pr(>F)	
0	1329	3.39E+00	0	NaN	NaN	NaN
1	1329	4.88E+10	0	-4.88E+10	inf	NaN

Takeaway: The above result is not easily interpretable. But by looking at Pr(>F) and F-statistics of **model 8**, it looks like that **model 1** is the best between the two.

Regression line for model 1

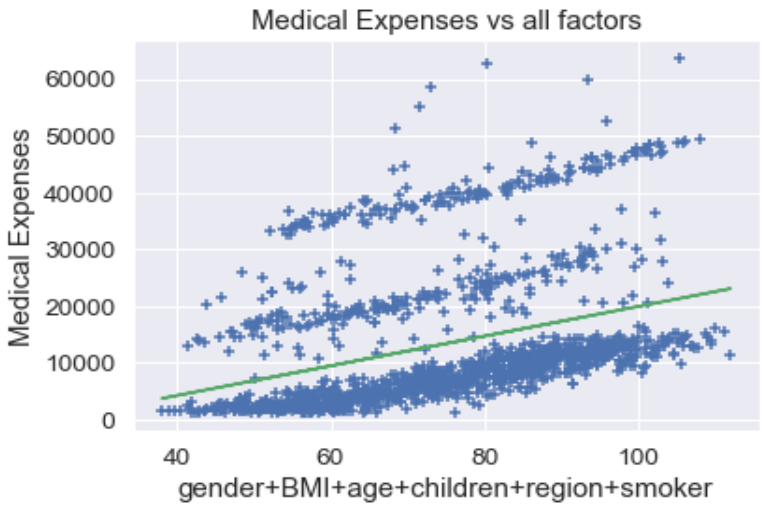


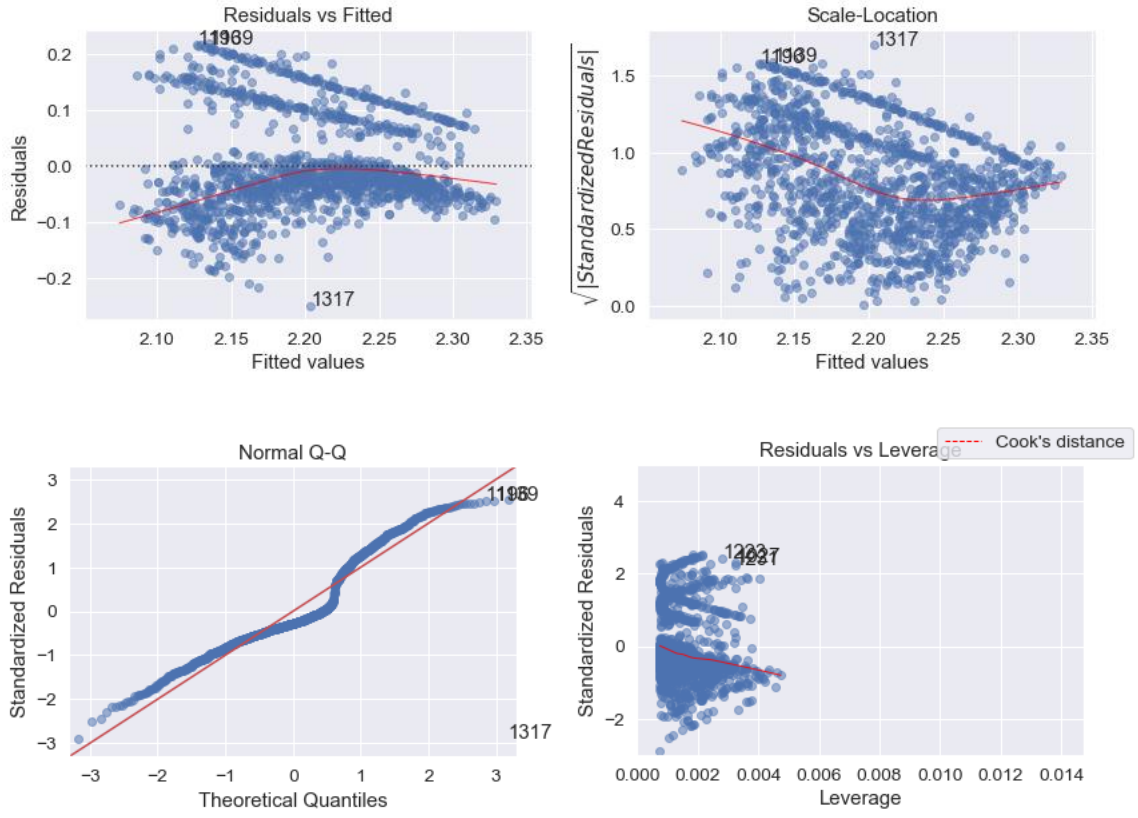
Fig: 16

Takeaway: The above scatter plot shows the regression line passing through the actual values of dependent variable. The regression line is showing that the predicted values of medical expenses are indeed based on the values of all the independent factors considered in the model.

Note: Predicted values of medical expenses are in log form. In order to get the values in integer values, we will have to apply antilog to those predicted values.

Diagnostic plot for model 1

Diagnostic plots of model 1 shown below gives us better understanding whether residuals are meeting the assumption of normal distribution.



The first two graphs i.e. residuals vs fitted value and square root of standardized residuals vs fitted values do not show any evident patterns in the residuals. The Normal Q-Q plot is giving us a picture that residuals is more or less following a straight red line. From residual vs leverage, I hardly can see any outlier values present that can affect the regression line.

Conclusion

The aim of my final project is to run the statistical analysis on each factor given in my dataset with respect to medical expenses (response variable) in python jupyter notebook environment. The key findings through statistical analysis can help both insured and policy maker in better planning for medical plan by avoiding factors that can lead to increase in insurance costs. Certain factors such as age, gender, regions can't be avoided or controlled. But the factors such as higher BMI, smoking, number of children can be looked upon to avoid increase in the medical expenses.

Summarization of the entire analysis through ANOVA tests and multiple regression models based on the dataset "Insurance Premium Prediction", into following key points:

- The amount of medical costs charge by the Insurance companies significantly depend on whether insured is smoker or not.
- Medical expenses do not vary based on four regions of the United States.
- In general, medical expenses do not vary based on whether a person is male or female
- On an average, medical expenses do vary based on smoker belong to which region of the United States
- On an average, medical expenses do differ depending on smoker is male or female
- Insurance companies in general, do charge higher medical costs to person with higher BMI and vice versa.
- In the regression model where response variable is transformed by applying log i.e. model 1, all the variables such as age, children, bmi, gender, region and smoker are significantly related to the medical expenses (response variable).

Further Improvements to Data Analysis

The model clearly shows that the factors considered in the multiple regression model: **model 1** such as gender, region, bmi, children, age, and smoker are significantly related to medical expenses. Hence, it is important to focus on these factors while evaluating medical expenses for any insured person. However, further improvements that I could have done in my analysis that could have improved the results are:

- If I had additional factor such as states, I could have got the results stating that medical expenses do vary among different states of the United States. In other words, some states do charges higher medical costs than the other states.
- Instead of applying log, I could have transformed the response variable through some other mathematical operations, for instance applying square root on the dependent factor to see if there are any improvements in the model
- Further, if had factors had a variable with three classifiers, I would have got better picture as what factors are involved in forming three layers of medical expenses with respect to age factor.

References

Conlen, B. (n.d.). Kernel Density Estimation. Retrieved August 11, 2020, from https://mathisonian.github.io/kde/

Crawley, MicR (Lenguaje de programación)hael J. (2008). *Statistics: An introduction using R*. Chichester: John Wiley and Sons.

Do you have a BMI 50? Weight Loss Information. (n.d.). Retrieved August 11, 2020, from https://www.marinahospital.com/weight-loss/bmi/50

Goldy-Brown, S. (2019, May 14). Average Healthcare Costs in 2019 - We Break Down The Stats. Retrieved August 11, 2020, from https://www.studentdebtrelief.us/news/average-healthcare-cost/

NHE Fact Sheet. (n.d.). Retrieved August 10, 2020, from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet

Price, S. (2020, June 12). Average Cost of Health Insurance (2020). Retrieved August 10, 2020, from https://www.valuepenguin.com/average-cost-of-health-insurance

The Cost of Health Insurance by State. (n.d.). Retrieved August 10, 2020, from https://www.healthmarkets.com/content/cost-of-health-insurance-by-state

Smokers, the obese, have markedly higher health-care costs than peers. (2015, January 06). Retrieved August 10, 2020, from https://www.sciencedaily.com/releases/2015/01/150106112542.htm

Why Are Americans Paying More for Healthcare? (n.d.). Retrieved August 10, 2020, from https://www.pgpf.org/blog/2020/04/why-are-americans-paying-more-for-healthcare

Writers, R. (2020, July 04). Healthcare Costs & Spend: Rising by Age, Gender, and Race. Retrieved August 11, 2020, from https://www.registerednursing.org/healthcare-costs-by-age/

(n.d.). Retrieved August 11, 2020, from https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

Fig. 1.10 Super-Smart Ways to Save on Medical Expenses.[Online Image]. (n.d.) Retrieved from https://toptenzilla.com/10-super-smart-ways-to-save-on-medical-expenses/