# Linear regression model for predicting medical expenses based on insurance data

Akhil Alfons Kodiyan (19210912), Kirthy Francis (19210588)
Engineering and Computing, Dublin City University, Dublin, Ireland

***Abstract*—Medical expenses is one of the major recurring expenses in a human life. Its a common knowledge that one life style and various physical parameters dictates diseases or ailments one can have and these ailments dictates medical expanses. According various studies, major factors that contribute to higher expenses in personal medical care include smoking, aging, BMI. In this study, we aims to find a correlation between personal medical expenses and different factors, and compare them. Then we use the prominent attributes as predictors to predict medical expenses by creating linear regression models and comparing them using ANOVA. In research, we found that smoking, age and higher BMI have a high correlation with higher medical expenses indicating they are major factors in contributing to the charges and the regression can predict with more than 75% accuracy the charges.**

***Index Terms*—medical expenses, smoking, BMI, linear regression, multiple regression**

## I. Introduction

According to WHO, personal expenditure on medical and healthcare has been increasing faster than the overall economy globally[1]. This increase in expenditure has been attributed to many causes, major of which include smoking, ageing and increased BMI. In this study, we aim to find a correlation between medical expenses and different factors using insurance data of different people with attributes such as smoking, age, number of children, region and BMI.

Here, we first find the correlation of medical charges with each of the attributes and use these attributes to predict charges. The method used is regression analysis, which tries to try to fit a predictive model to the data and then use that model to predict an outcome variable from one or more independent predictor variables. We used multiple regression to create different models and then used ANOVA to compare the different models and find the best-fit model.

## II. Related Work

There have been various studies in this field, as the increase in healthcare expenditure has been dominant over the years. Obesity has similar health conditions as that which come with twenty years ageing as Sturm discusses in this paper [2]. Here he compares the effects of obesity, overweight, smoking, and problem drinking based on national survey data in the USA. He concludes obesity is more dependent on healthcare expenditure when compared to smoking or problem drinking. A very similar paper uses multivariate analysis on online Health Risk Assessment (HRA) conducted on a sample of South Africans [3]. While An used retrospective data analysis to estimate personal health care expenses associated with smoking and obesity among US adults by gender, ethnicity and age group [4]. Regression analysis is a popular method used for estimating medical charges as demonstrated by David et al. in his paper which estimates the medical costs with respect to obesity in the United States [5]. They used linear regression and logistic regression among any other methods to estimate the medical costs. Another similar paper used multiple regression and hierarchical multiple regression to find the determinants and associated factors that contribute to healthcare expenditures in Korea [6]. The study categorises determinants into negotiable and non-negotiable factors and finds that the proportion of elderly in the population is the major contributor to medical costs.

## III. Dataset & Exploratory analysis

### A. Dataset

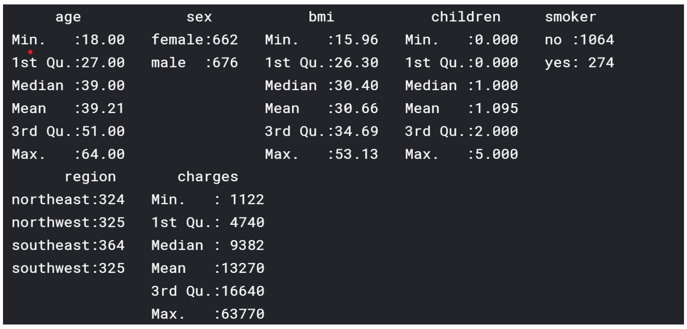
The data set we use here is the medical cost personal dataset from Kaggle which consists of peoples anonymous information and annual insurance premiums given to them. The attributes in this dataset is as follows.

| Column | Description |
|---|---|
| Age | Age of primary beneficiary |
| Sex | Insurance contractor gender. female/ male |
| BMI | Body mass index, provides an understanding of body. |
| Children | Number of children covered by insurance |
| Smoker | If Insurance primary smokes |
| Region | The beneficiary's residential area in the US. |
| Charges | Individual medical costs billed by health insurance. |

The variable age and BMI are continuous variables, the variables sex, smoker and region are categorical variables. In the summary statistics below, we can see that there are no missing values in the dataset.

### B. Data Exploration

In this section, we shall explore the data by correlations and scatter plots to find out any visible relation between various attributes, so as to use such findings in the later phase for predicting medical charges.



```
      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
```

Fig. 1. Summary of Data

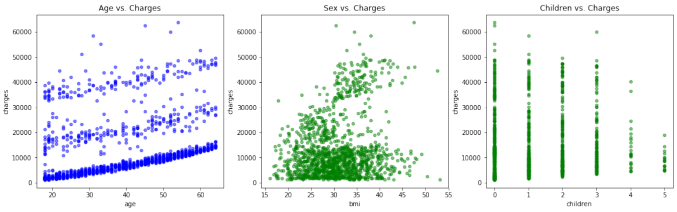
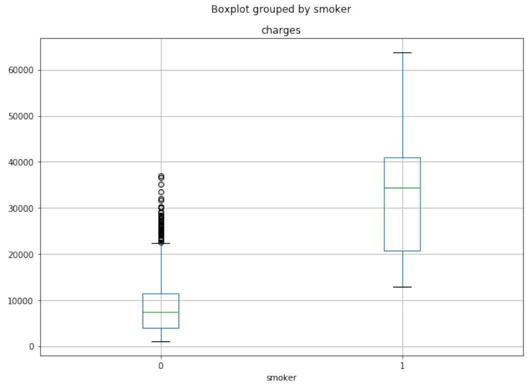


Fig. 2. Charge vs Age, BMI, Children

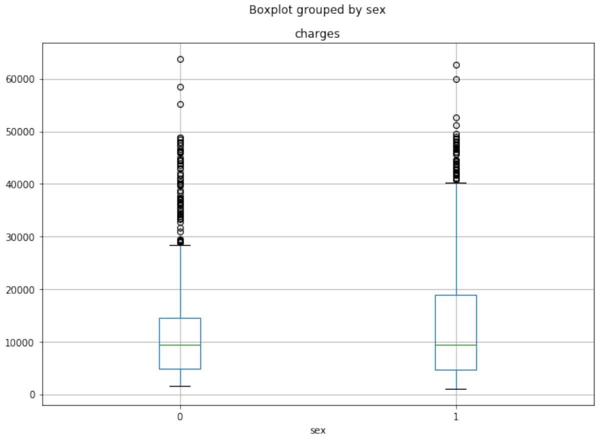
1) *Visualising data spread with respect to charge:* Here we will explore visually various attributes and how they affect charges.

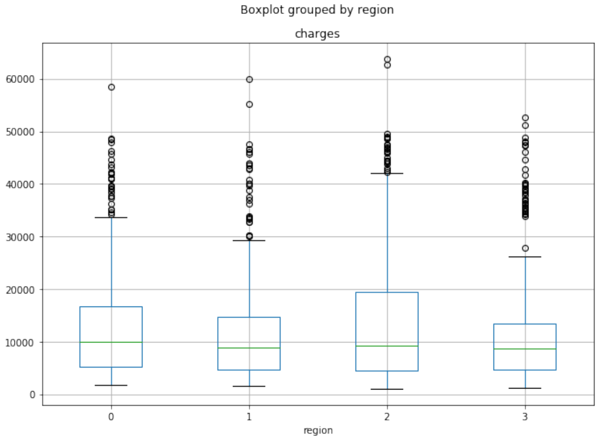
Clearly, a trend of increasing charges can be observed from the age vs charge plot. In BMI vs charge, there seems to be no discernible pattern, but an increase in BMI seems to hold some correlation to charges. Here a correlation between the number of children and charge is not clear from the scatter plot.



There seems to be a strong correlation between smoking and medical charges, from the box diagram one can clearly make out the fact that a smoker nearly pays 4 times the medical expenses when compared to a non-smoker.

Gender seems to have next to no influence on the charge as the mean medical expenses of both sexes seems nearly equal. Similarly, the region also seems to have no effect on the medical charges in any significant way. The influences each attribute hold on each other would be more clear on evaluating correlation.





2) *Correlation of attributes:* Here we will find the correlation matrix between the columns and then visualize the same using a Heatmap. Hence, we will be able to figure out the relationship between attributes more clearly and visualize them.

A strong correlation can be observed between smoker attribute and charge from the heat map. Which is followed by age and BMI, other factors seem to have less correlation charge. This is a useful fact in getting started with the regression model as we can start our model using these attributes.

## IV. Hypotheses and or Research Questions

In this section, we will try to come up with a regression model, to predict with a reasonable amount of certainty the personal medical charges. In previous section, we discovered that there is a strong correlation between smoking, age, BMI to charges, so in the initial model we create a multiple regression model with these attributes. i.e the predictor formula would be smoking+age+bmi.

In the summary of this model, Multiple R-squared is 0.7475. This value represents the square R between age and charges and indicates that 74.75% of the variation in the outcome variable charges can be explained by the predictors. The adjusted R-square is almost equal to the multiple R-square,

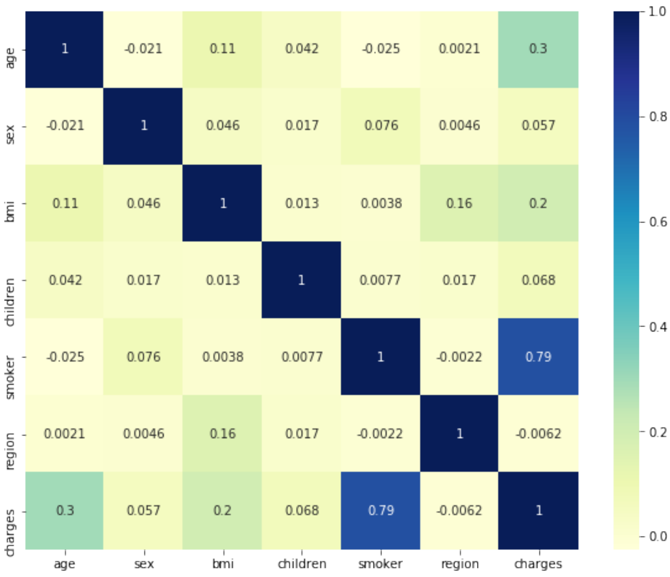
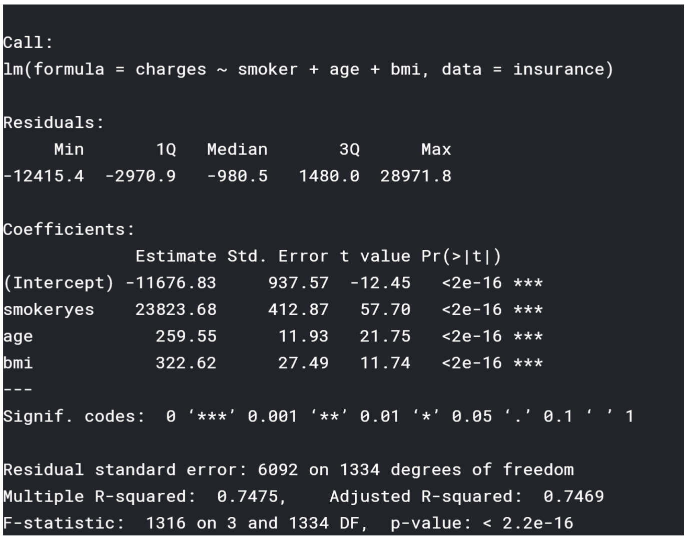


Fig. 3. Correlation heatmap of attributes



```
Call:
lm(formula = charges ~ smoker + age + bmi, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-12415.4  -2970.9   -980.5   1480.0  28971.8 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11676.83     937.57  -12.45   <2e-16 ***
smokeryes    23823.68     412.87   57.70   <2e-16 ***
age            259.55      11.93   21.75   <2e-16 ***
bmi            322.62      27.49   11.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6092 on 1334 degrees of freedom
Multiple R-squared:  0.7475,	Adjusted R-squared:  0.7469 
F-statistic:  1316 on 3 and 1334 DF,  p-value: < 2.2e-16
```

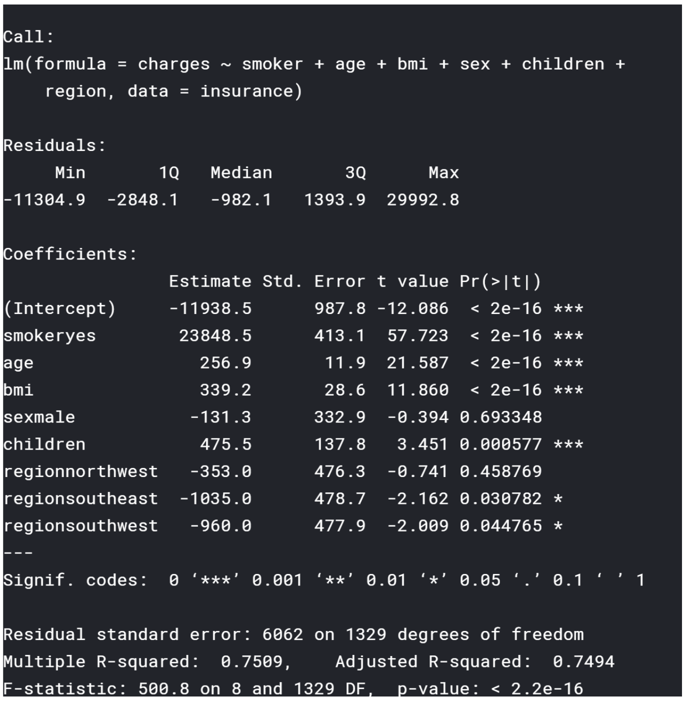
Fig. 4. Summary of Model (smoking+age+bmi)

this tells us that the model has good cross-validity. From the F-statistic (P¡0.001) we can conclude that the model is good and that the predictor variables are all significant. The coefficient for age is 259.55, BMI is 322.62 and smoker is 238223.68. The intercept is at -11676.83. Thus the model for predicting charges using smoking,age,BMI becomes,

charges = -11676.83 + (238223.68 * smoker) + (259.55 * age) + (322.62 * BMI)+ error.

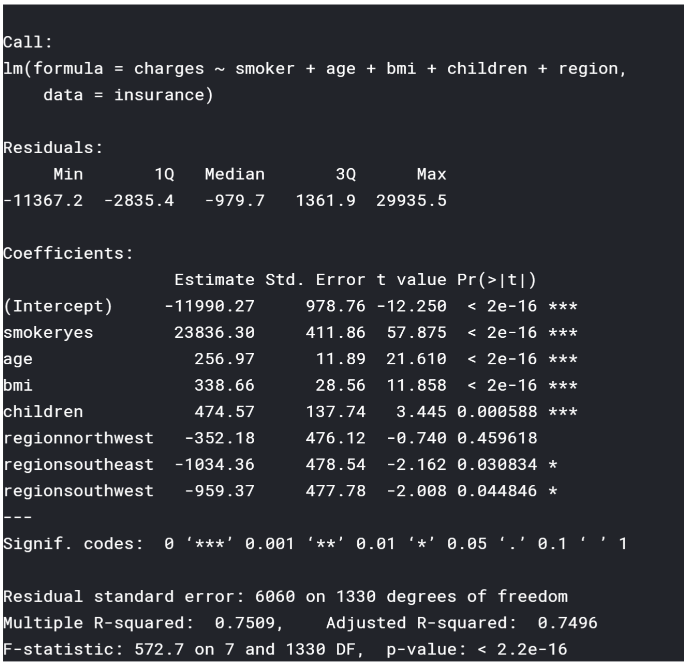
Next let’s try to improve this model by including the predictor variables sex, children and region.

In this updated model the multiple R square is 0.7509, meaning that the new model explains 75.09 of the variation in charges. While looking at the coefficients & significant codes we can see that sex is not significant. Now we remove the sex attribute from the model and reevaluate model.



```
Call:
lm(formula = charges ~ smoker + age + bmi + sex + children + 
    region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-11304.9  -2848.1   -982.1   1393.9  29992.8 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
bmi                339.2       28.6  11.860  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348    
children           475.5      137.8   3.451 0.000577 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769    
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *  
regionsouthwest   -960.0      477.9  -2.009 0.044765 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,	Adjusted R-squared:  0.7494 
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

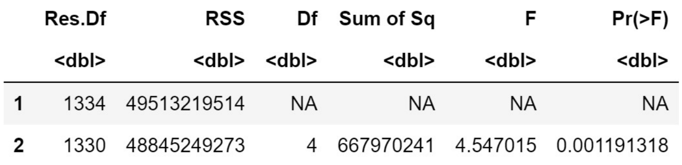
Fig. 5. Summary of Model (smoker + age + bmi + sex + children + region)



```
Call:
lm(formula = charges ~ smoker + age + bmi + children + region, 
    data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-11367.2  -2835.4   -979.7   1361.9  29935.5 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11990.27     978.76 -12.250  < 2e-16 ***
smokeryes        23836.30     411.86  57.875  < 2e-16 ***
age                256.97      11.89  21.610  < 2e-16 ***
bmi                338.66      28.56  11.858  < 2e-16 ***
children           474.57     137.74   3.445 0.000588 ***
regionnorthwest   -352.18     476.12  -0.740 0.459618    
regionsoutheast  -1034.36     478.54  -2.162 0.030834 *  
regionsouthwest   -959.37     477.78  -2.008 0.044846 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6060 on 1330 degrees of freedom
Multiple R-squared:  0.7509,	Adjusted R-squared:  0.7496 
F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

Fig. 6. Summery of Model (smoker + age + bmi + children + region)

After removing sex from the model, we can observe that the model still explains 75.09 per cent of the variation in charges. Let compare the initial model and the current one using ANOVA.

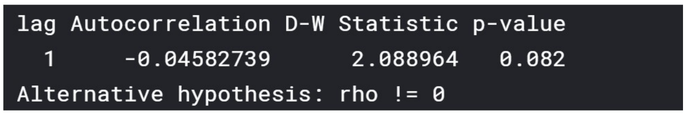


| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|---|
| | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | 1334 | 49513219514 | NA | NA | NA | NA |
| 2 | 1330 | 48845249273 | 4 | 667970241 | 4.547015 | 0.001191318 |

Fig. 7. Summary of ANOVA (smoker + age + bmi) vs (smoker + age + bmi + children + region)

Let's examine the case of outliers. One way of detecting outliers is to look at the differences between the observed and predicted values (aka residuals). An outlier would show a large difference between the predicted and observed value. And on calculating The percentage of cases with a standardized residual greater than 2 is 5.01 per cent and greater then 2.5 is 3.29 per cent. So it is expected that 95% of the cases are within the boundaries of -2 and 2, which is not too much.

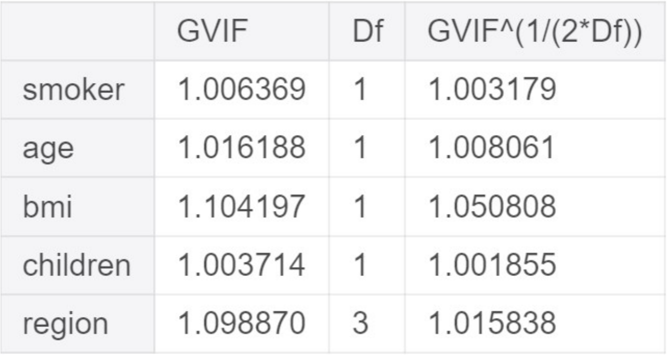
Checking the assumption of independence using the Durbin-Watsons test. The D-W statistic is very close to 2 and we see a p-value larger than 0.5. This means that the assumption of independence is met.



```
lag Autocorrelation D-W Statistic p-value
  1      -0.04582739      2.088964   0.082
Alternative hypothesis: rho != 0
```

Fig. 8. Summery of DWT Test

Checking the assumption of multicollinearity, One of the easier methods to verify this assumption is to consider the variance inflation factor (VIF). The VIF should not be higher than 10, which is the case here.

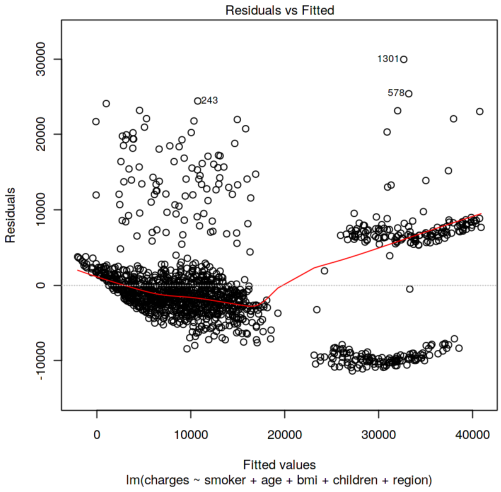


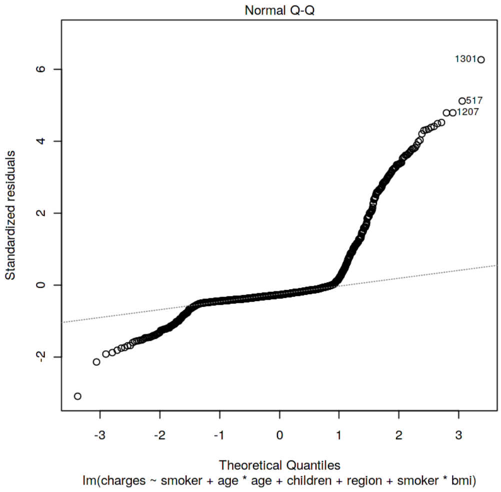
| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| smoker | 1.006369 | 1 | 1.003179 |
| age | 1.016188 | 1 | 1.008061 |
| bmi | 1.104197 | 1 | 1.050808 |
| children | 1.003714 | 1 | 1.001855 |
| region | 1.098870 | 3 | 1.015838 |

Fig. 9. Summery of VIF Test

Finally, to check, assumptions about residuals to plot the standardized residuals. This plot should look like random dots evenly distributed amongst the zero lines.

In this chart, the dots should be randomly placed around the horizontal zero lines, which is no the case here. It seems that there are identifiable 3 groups. Thus as per each group, there are differences invariance across the residuals.





Also there is an increasing variance across the residuals and there might indicate a non-linear relationship between the charge and the predictor. This non-linear relationship is also reflected by the second plot (Q-Q plot) since the dots deviate from the dotted line.

## V. Methods Used and Why

Multiple linear regression is used when there is more than one independent variable used for the prediction of a response variable. For our model, we used smoking, age, BMI, number of children and region as predictors. Since there is more than one explanatory variable we used multiple linear regression.

We used the lm() function or the linear model function to create simple regression models. The lm() function accepts parameters and returns a linear model. We saved each model to a variable so that they can be used in following calculations and comparisons. One of the two parameters used is a formula, which describes the model - usually in the form of YVAR $\sim$

XVAR, where YVAR is the dependent, or predicted variable and XVAR is the independent, or predictor variable and the other is a variable that contains the dataset.

We created more than one model to compare and identify the best-fit model using Analysis of Variance (or deviance) to find the best fit model. Here we created the first model containing the predictor's smoker, age and BMI and then a second model adding sex, region and children to the first model. From the coefficients in the summary of the second model, we find that sex is not a significant variable. Hence, we created a third model removing sex predictor from the second model. Then we compare the first and third model using ANOVA to find the model which gives a better prediction.

Once we found the model which gives the best prediction, we proceeded to check the assumptions of linear regression on our model.

Linear regression is sensitive to outlier effects, therefore it is important to check for outliers in the model. Outliers can be detected using residuals i.e. the differences between the observed and predicted values. We used standardized residuals to detect the outliers. Standard residual is essentially the ratio of the residual and the predicted value of a data point.

The Durbin-Watson test was used to measure the autocorrelation in the residuals. Auto-correlation occurs when there is no independence between the residuals. The Durbin-Watson test requires the D-W Statistic to be very close to 2 and the p-value to be larger than 0.5.

To check if there is any correlation between the predictors in the regression analysis, we calculated the Variance Inflation Factor(IVF). If the IVF is greater than 10, it implies the correlation between the predictors.

In Q-Q plot, which is a graphical method for examining two probability distributions by plotting their quantiles against each other, was used to check if all the variables are multivariate normal. When the data is distributed in a linear transformation, roughly a straight line is formed.

## VI. Results and Findings

In our research, we find that region and gender do not bring significant difference on charges among its peers. Whereas Age, BMI, number of children and smoking are the ones that drive the charges. Among this Smoking seems to have the most influence on the medical charges. Though the final model explains 75% of the data, there seems to be an indication of a non-linear relationship as seen from the Q-Q diagram.

## VII. Conclusion

Through building linear regression models and comparing the fitted models using ANOVA, we were able to predict with relatively high degree of accuracy medical expenses, healthcare costs also have non linear dependency significantly across Age, BMI. While Other factors such as region and gender have least relevance in the medical expenses based on insurance data.In real insurance much more variables are considered like, various disease conditions, prior medical history etc. Due to lack of data we were unable to research on such data in this paper. Also we have observed some indications of non-linear relationships between charges and BMI and age which we were unable to explore in this work.

## References

[1] World Health Organization. *Public spending on health: a closer look at global trends.* No. WHO/HIS/HGF/HFWorkingPaper/18.3. World Health Organization, 2018.

[2] Sturm, Roland. *The effects of obesity, smoking, and drinking on medical problems and costs.* Health affairs 21, no. 2 (2002): 245-253.

[3] Sturm, Roland, Ruopeng An, Josiase Maroba, and Deepak Patel. *The effects of obesity, smoking, and excessive alcohol intake on healthcare expenditure in a comprehensive medical scheme.* South African Medical Journal 103, no. 11 (2013): 840-844.

[4] An, Ruopeng. *Health care expenses in relation to obesity and smoking among US adults by gender, race/ethnicity, and age group: 19982011.* Public Health 129, no. 1 (2015): 29-36.

[5] Kim, David D., and Anirban Basu. *Estimating the medical care costs of obesity in the United States: systematic review, meta-analysis, and empirical analysis.* Value in Health 19, no. 5 (2016): 602-613.

[6] Han, Kimyoung, Minho Cho, and Kihong Chun. *Determinants of health care expenditures and the contribution of associated factors: 16 cities and provinces in Korea, 2003-2010* Journal of Preventive Medicine and Public Health 46, no. 6 (2013): 300.

[7] Finkelstein, Eric A., Ian C. Fiebelkorn, and Guijing Wang. *National Medical Spending Attributable To Overweight And Obesity: How Much, And Who's Paying? Further evidence that overweight and obesity are contributing to the nation's health care bill at a growing rate.* Health affairs 22, no. Suppl1 (2003): W3-219.

[8] Wolf, Anne M., and Graham A. Colditz. *Current estimates of the economic cost of obesity in the United States.* Obesity research 6, no. 2 (1998): 97-106.

[9] Van Baal, Pieter HM, Johan J. Polder, G. Ardine de Wit, Rudolf T. Hoogenveen, Talitha L. Feenstra, Hendriek C. Boshuizen, Peter M. Engelfriet, and Werner BF Brouwer. *Lifetime medical costs of obesity: prevention no cure for increasing health expenditure.* PLoS medicine 5, no. 2 (2008): e29.

[10] Haynes, George, and Tim Dunnagan. *Comparing changes in health risk factors and medical costs over time.* American Journal of Health Promotion 17, no. 2 (2002): 112-121.

[11] Warner, Kenneth E., Thomas A. Hodgson, and Caitlin E. Carroll. *Medical costs of smoking in the United States: estimates, their validity, and their implications.* Tobacco control 8, no. 3 (1999): 290-300.

[12] Flegal, Katherine M., Margaret D. Carroll, Cynthia L. Ogden, and Clifford L. Johnson. *Prevalence and trends in obesity among US adults, 1999-2000.* Jama 288, no. 14 (2002): 1723-1727.

[13] Wang, Houli, Tengda Xu, and Jin Xu. *Factors contributing to high costs and inequality in China's health care system.* JAMA 298, no. 16 (2007): 1928-1930.