Medical Insurance Analyses Report

Ryota Nishida

Executive Summary

This report summarizes a graphical and statistical analysis of the effect of various factors such as age,sex, and children on a dependent variable called charges.The analysis is done using data statistics taken from the U.S. Census Bureau. The analysis indicates that factors such as smoking,age,bmi,and children have a statistically significant effect on medical insurance charges.

Data Description

The data set comes from a file called "insurance.csv" which has been retrieved from the data website Kaggle and has been entered into R using the read.csv function. The data was originally derived from a book by Brett Lantz called Machine Learning with R. The data was collected using demographic statistics from the U.S. Census Bureau. It includes data from 1338 individuals who are enrolled in an insurance plan. It features different characteristics of these individuals. It contains 7 categories made up of the following:

- **age**: The age of the individual (excludes those above 64 years as their insurance is often covered by the government)
- **Sex**: Gender of individual, consisting of male or female
- **bmi**: The body mass index of the individual to see whether or not they are at a healthy weight for their size.
- **children**: This shows how many of the individuals dependents are covered by the insurance plan
- **smoker**: Indicates whether or not the individual smoke, the choices are yes or no
- **region**: This is the area in which the individual is located, it includes the regions: northeast, southeast, southwest, or northwest.
- **charges**: This is the Individual medical costs billed by health insurance.

Table 1: Insurance Data

age	sex	bmi	children	smoker	region	charges
19	female	27.900	0	yes	southwest	16884.924
18	male	33.770	1	no	southeast	1725.552
28	male	33.000	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.471
32	male	28.880	0	no	northwest	3866.855
31	female	25.740	0	no	southeast	3756.622
46	female	33.440	1	no	southeast	8240.590
37	female	27.740	3	no	northwest	7281.506
37	male	29.830	2	no	northeast	6406.411
60	female	25.840	0	no	northwest	28923.137
25	male	26.220	0	no	northeast	2721.321
62	female	26.290	0	yes	southeast	27808.725
23	male	34.400	0	no	southwest	1826.843
56	female	39.820	0	no	southeast	11090.718
27	male	42.130	0	yes	southeast	39611.758
19	male	24.600	1	no	southwest	1837.237
52	female	30.780	1	no	northeast	10797.336
23	male	23.845	0	no	northeast	2395.172
56	male	40.300	0	no	southwest	10602.385
30	male	35.300	0	yes	southwest	36837.467
60	female	36.005	0	no	northeast	13228.847
30	female	32.400	1	no	southwest	4149.736
18	male	34.100	0	no	southeast	1137.011
34	female	31.920	1	yes	northeast	37701.877
37	male	28.025	2	no	northwest	6203.902
59	female	27.720	3	no	southeast	14001.134
63	female	23.085	0	no	northeast	14451.835
55	female	32.775	2	no	northwest	12268.632
23	male	17.385	1	no	northwest	2775.192
31	male	36.300	2	yes	southwest	38711.000

The table 1 showcases the first 30 rows from the insurance data frame. The data frame has not been modified and is identical to the original. No modifications were made due to the data already being tidy as it already features columns with distinct attributes and unique variable names. Furthermore the rows showcase unique observations with random ordering and the table already has a unique identifier.

Table 2: Insurance Data Statistics

	attr.no	n	Mean	sd	min	max	range	se
age	1	1338	39.207025	14.049960	18.000	64.00	46.00	0.3841024
sex	2	1338	NaN	NA	Inf	-Inf	-Inf	NA
bmi	3	1338	30.663397	6.098187	15.960	53.13	37.17	0.1667142
children	4	1338	1.094918	1.205493	0.000	5.00	5.00	0.0329562
smoker	5	1338	NaN	NA	Inf	-Inf	-Inf	NA
region	6	1338	NaN	NA	Inf	-Inf	-Inf	NA
charges	7	1338	13270.422265	12110.011237	1121.874	63770.43	62648.55	331.0674543

The table 2 showcases some summary statistics on the data frame which includes a measure of central tendency such as the mean and a measure of statistical dispersion such as the standard deviation.(Note: sex,smoker, and region feature non-numeric values because they are categorical variables)

- **Attr.NO**: shows the number of the attribute
- **n**: shows the number of valid cases
- **mean**: shows the central or typical value within the data set($\overline{x}$)
- **sd**: This is the standard deviation which measures the dispersion of the data set relative to its mean ($\hat{\sigma}$)
- **max**: Shows the maximum value of the dataset
- **min**: Shows the minimum value of the dataset
- **range**: Shows the difference between the largest and smallest values
- **se**: This is the standard error which shows the standard deviation of its sampling distribution

Objectives

The insurance data will be used to see how different factors affect the amount of medical expenses that an individual posseses. It will also be used to see what factors impact medical expenses the most and least. In order to do this the relationship between the independent variables and dependent variable(charges) will be explored.A model will be built that can determine possible factors that influence the amount of medical expenses an individual procures. The variables that will be observed for their impact on the amount of charges are age,sex,bmi,children,smoker, amd region. Then some conclusions about the possible factors that influence the amount of medical expenses can be drawn from this study and applied to the larger population from which the sample was taken, giving us an idea about how different socio-economic factors impact medical expenses and thus medical charges billed by health insurance.

Background

In order for the report to be understood, some background theory and mathematical equations will be required. The necessary information will be listed below:

1. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i + \epsilon$

The above formula is known as an multiple linear regression equation. y is the response variable and showcases the effect of the model. β_i shows the estimated effect for each of the regressors. The x symbol is to denote the regressor. The i symbol simply denotes the regressor in question. This formula will be used to see the relationship between y and the x_i's.

2. Hypotheses Testing

This is a method of statistical inference used to decide whether the data at hand sufficiently supports a particular hypothesis. This will be used to evaluate the relationships between charges and the predictor variables. It will help show which regressors have a statistically significant effect on charges.

3. Null Hypotheses vs. Alternative Hypotheses

The null hypotheses is a statement about the population parameter and the alternative hypotheses is a contradicting hypotheses. The null hypotheses used will be that all coefficients in the model are equal to zero and the alternative hypotheses will be that not every coefficient is simultaneously equal to zero.

4.T-statistic

$$T = \overline{x} - \mu_0/(\hat{\sigma}/\sqrt{n})$$

The above formula is known as the t-statistic. This formula takes the difference between the sample mean $\overline{x}$ and the population mean μ_0 and divides it by the standard error which is the product of the standard deviation $\hat{\sigma}$ divided by the square root of the sample size $\sqrt{n}$. This formula will be used to test the linearity of the relationship between the response variable(charges) and the different predictor variables.

5. p-value

Once the t statistic is calculated, the corresponding p-value can be found using a t-distribution table. A p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct.

6.Significance level(α)

The significance level is the probability of rejecting the null hypothesis when it is true. The significance level is compared to the p-value. When the p-value is smaller than the significance level then the null hypotheses can be rejected.

7. F-statistic

The F-statistic is used to see whether or not a model has coefficients that all equal to zero. The formula for the F-statistic is

$$F = \frac{SSR_{restricted} - SSR_{unrestricted}/q}{SSR_{unrestricted}/(n - k_{unrestricted-1})}$$

$SSR_{restricted}$ is the restricted residual sum of squares and $SSR_{unrestricted}$ is the unrestricted sum of squares.The symbol n is the number of observations and k is the number of regressors. q is the number of restrictions. Once the F-statistic is calculated, it's p-value can then be found using a F-table. Just as in the case for the t-statistic, if the p-value is less than the significance level, then we can say that one or more of the coefficients have a mean significantly different from zero.

8. R-Squared

The R-Squared value is a statistical measure that shows the percentage of the variation in the dependent variable that is explained by the regressors of a model. The R-Squared value will help give an idea about the accuracy of the multi-linear regression model. The formula for R-Squared is

$$R^2 = 1 - \frac{SSR}{TSS}$$

SSR is the sum of squared residuals and TSS is the total sum of squares.

9. Correlation coefficient

Correlation is a statistical measure that describes the size and direction of a relationship between two or more variables. The formula for correlation is

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}$$

x_i is the values of the x-variables within the model,$\overline{x}$ is the sample mean,y_i is the values of the response variable. $\overline{y}$ is the mean of the response variable. The correlation coefficient will help give an understanding of the strength and direction of relationship between the regressors and the dependent variable. It will also be able to give the relationship between to predictors as well.

Methods

A few graphical techniques will be employed; these will include box-plots,compound residual plots,scatter plots,and histograms.Box plots are used for displaying basic distributional features of uni-variate data, but they can be extended to multivariate settings by being placed side-by-side. We use side-by-side box-plots to compare means and outliers in relation to the dependent variable Charges and the regressors of age,sex,bmi,children,smoker, and region.

Scatter plots help to showcase the effects of the regression model in a visual manner. The use of a curve of fit helps to also explain the data within a scatter plot by helping the viewer to observe the relationship between the dependent variable and the regressors and showcasing things like whether the model is linear or non linear in nature.Scatter plots are useful for displaying bivariate data, but they can also be extended in

a limited way to reveal multivariate patterns through the use of pairwise scatter plots, where each pair of variables in the data set is used to produce a scatter plot.

The compound residual plot will showcase: a residual vs fitted,scale location,constant leverage, and QQ plot. These plots will give insights about the distribution and linearity of the insurance data set. Along with this, an understanding of the spread and scedasticity of the data set can also be gained in order to test out different assumptions. The histogram will provide a visual interpretation of the numerical data by showing the number of data points that fall within a specified range of values (called "bins"). The histogram will be used to obtain a first overview of the dependent variable.

A multiple linear regression model will also be implemented in order to examine the relationship between the predictors and the response variables. These will help showcase which factors have the highest impact on charges and which ones have no impact at all.

Results and Discussion

We had previously calculated the mean of the response variable charges within table 2, we would now like to calculate the variance of charges as well in order to get an idea of the dispersion within the model. The mean of charges was 13270.42 and its variance is 146652372.The variance is much greater than the mean, which suggests that there will be over-dispersion within the model.

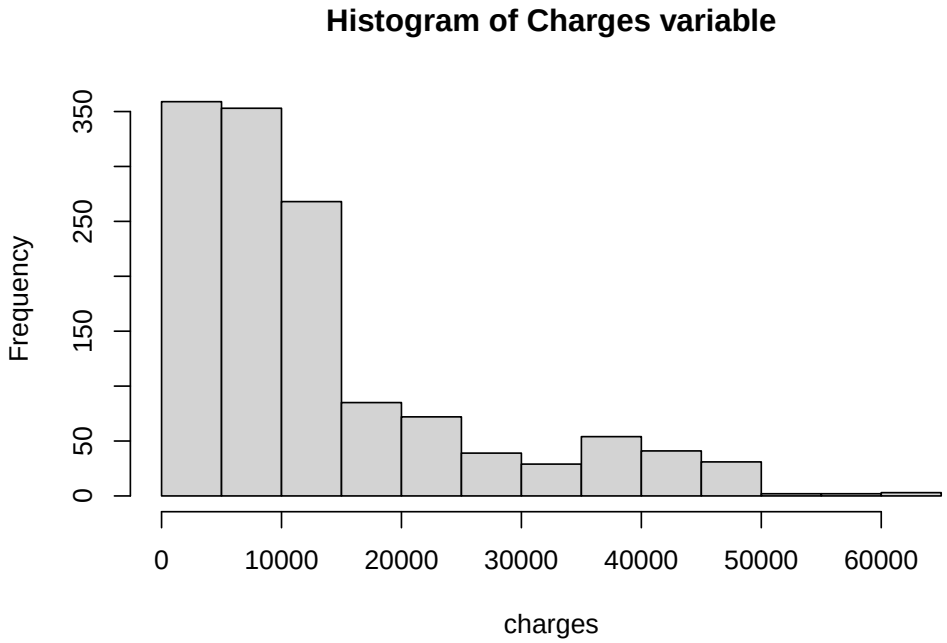

Figure 1: Histogram of the dependent variable charges's data continuity

In figure 1, the histogram showcases the frequency of charges for the range of medical expenses billed by insurance. By looking at the histogram it clearly looks like the data does not follow a normal distribution. The data is skewed to the right.

The compound residual plot in figure 2 suggests that the variance may not be constant and that the data is heteroskedastic in nature. This is evident from the dispersion of the residuals viewed in both the the Scale location plot and the Residual vs Fitted plot. The QQ plot suggests that the points diverge away from the dashed line and therefore the data most likely does not follow a normal distribution.On the right hand side of the graph, the points lie above the line which suggests a 'fat tail' on the right hand side of the distribution. The scale location plot also shows that the average magnitude of the standardized residuals increases with the predicted values as can be seen through the fact that the red line has a positive slant. The Residuals vs

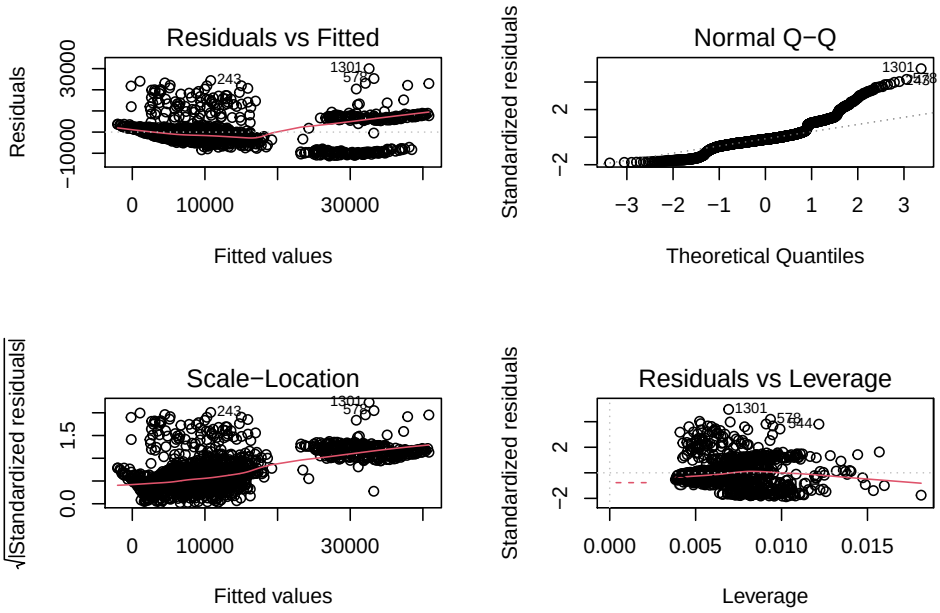

Figure 2: Compound residual plot

Leverage plot shows no points lie outside the cooks distance, which shows that the data may not have high influence points.

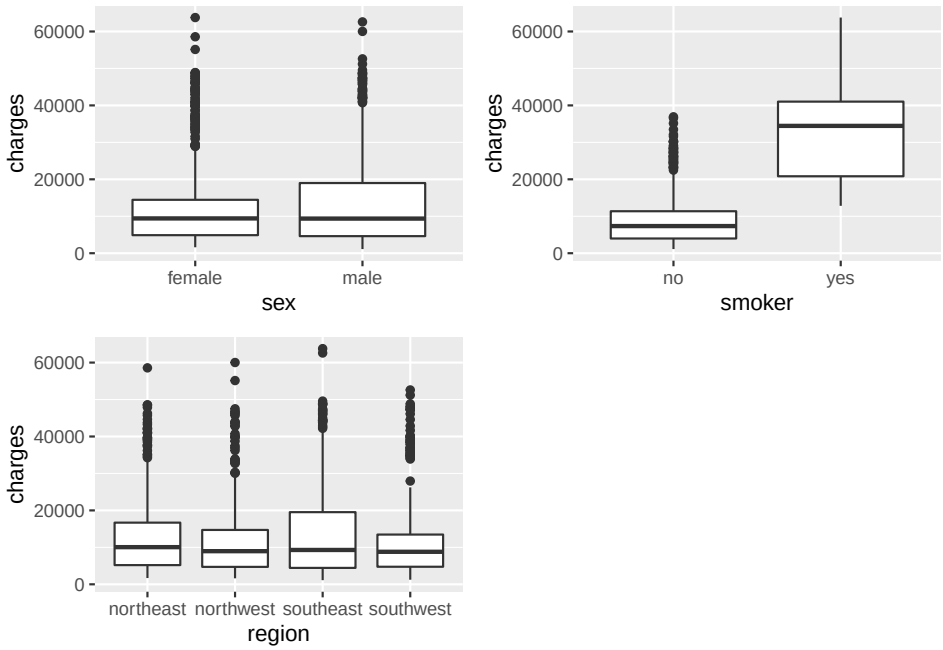

Figure 3: Sideby-side box plots for the dependent variable(Charges) against the categorical regressors

The boxplots in figure 3 show that the average amount of charges is higher for smokers than non-smokers. It also shows that smokers have a higher variation in the amount of charges that are billed by insurance. Average charges seem to be even when it comes to the sex and region boxplots. Males seem to have more variation in their charges than females and the southeast region has the the most variation out of the other regions. There seems to be a fair amount of outliers in the data aswell.

```
##
## Call:
## lm(formula = charges ~ ., data = insurances)
##
## 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 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## 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
```

Since charges is a numeric variable and we want to see possible factors that influence the amount of charges billed by medical insurance, the multi-linear regression model will be used as it allows to see the effects of multiple variables on a response variable. A summary of the multi-linear regression featuring charges as the dependent variable shows that sex and regionnorthwest are statistically insignificant regressors at the 0.05 significance level, meaning that there is not enough evidence to say that the two regressors have a meaningful effect on charges.

Smoking seems to have the highest positive impact on charges with a coefficient value of 23848.5. This means that if a person is a smoker, their charges billed by insurance increase by 23848.5 dollars. The number of children who are covered by the insurance plan is the factor with the second highest positive impact on charges. A 1 unit increase in the number of dependents causes a 475.5 dollar increase in charges. Bmi is third with an effect of 339.2.

The factor that causes the largest decrease in charges seems to be region. Specifically, being from the south east region of the US. Those hailing from that region experience a decrease in insurance by 1035 dollars. The southwest is not too far behind with an negative effect of 960.0.

The model features an F-statistic less than 2.2e-16. This is an statistically significant F-statistic as it is lower than the significance level of 0.05. This means that it can be said with 95 percent confidence that one or more of the regressors have a mean that is significantly different from zero.It also shows that the model is better than one with no regressors.

It can also be seen that the model showcases an adjusted R-squared value of 0.7494. This means that 74.94 percent of the variation in charges can be explained by the model.

The fact that smoking was the biggest contributor to medical charges billed by insurance is no surprise as even the American Journal of Preventive Medicine states that,"By 2010, 8.7% of annual healthcare spending in the U.S. could be attributed to cigarette smoking, amounting to as much as $170 billion per year"[1]. The fact that the southeast of the US is associated with the largest decrease in health insurance charges makes sense as well. This is because according to the CDC's national health statistics report, "The percentage of adults aged 18–64 who were uninsured was significantly higher than the national average (13.9%) in Florida (19.5%), Georgia (25.4%), North Carolina (20.3%), and Texas (28.1%)"[2]. Most of theses states happen to be in the southeast. Many uninsured individuals living in the southeast may be the reason why the southeast region has the largest decrease in charges. The model also showcased bmi as an statistically significant coefficient on charges. The CDC states,"BMI appears to be as strongly correlated with various metabolic and disease outcome as are these more direct measures of body fatness"[3]. Analysis from the CDC has also shown a correlation between bmi and disease which further explains the positive relationship between bmi and medical insurance charges.

Figure 4, shows multiple scatter plots along with the multi-linear regression line. They showcase the individual relationships between charges and the the various coefficients. As can be seen, there are a few plots that have observations that seem to move in clusters in an vertical direction. This makes sense because they are the obeservations for the set of variables that are categorical in nature, the children variable also showcases this pattern because it is of the integer type. Also, scatterplots are best suited for numeric data, such as that of age and bmi. The scatter plot for bmi seems to show a positive correlation between bmi and charges, however this is not clear and further inspection may be necessary. The scatter plot for age shows a more apparent relationship between age and charges. It can be seen that there is a positive relationship between age and charges.

Table 3: Insurance Data Statistics

	Age	bmi	charges
age	1.0000000	0.1092719	0.2990082
bmi	0.1092719	1.0000000	0.1983410
charges	0.2990082	0.1983410	1.0000000

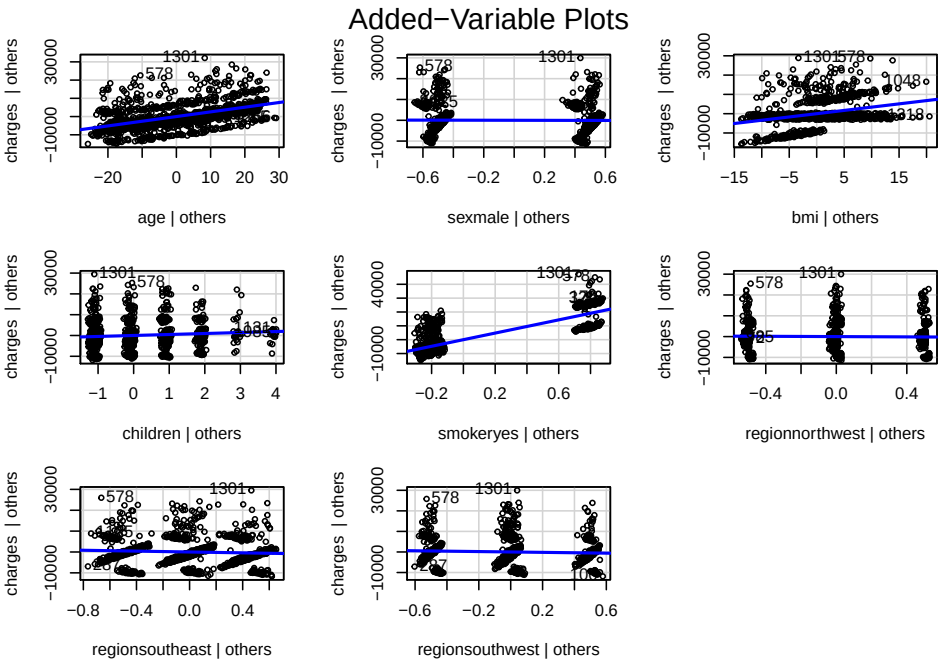

Figure 4: Scatter plots

The correlation of age and bmi against charges can be further examined to get a better understanding of the pattern seen within the scatter plots. The above table in table 3, shows a matrix of the correlations between age,bmi, and charges. Age seems to have a positive correlation with charges with a value of 0.2990082, this means that although it has a positive relationship, it is a weak one in nature.Bmi has a weaker positive relationship with charges with a value of 0.1983410. Age and bmi also seem to be positively correlated with each other with a value of 0.1092719.

Conclusion

In general, whether someone smoked or not seemed to be the biggest factor that impacted health insurance charges. The factor that had the largest negative impact was someone being from the southeast region. The model showed the presence of heteroskedasticity and a distribution that was not normal. However, according to the central theorem a sufficiently large enough sample size will make the sampling distribution of the mean for a variable approximate a normal distribution. Since the data features more than 35 observations, we can assume the data takes a normal distribution. The statistical significance within the regressors showed that only sex and the northwest region were the only statistically insignificant factors. The fact that the results of the model did not contradict other studies on health care shows that the model is accurate. The R-Squared value of the model also showed a high level of correlation.

Recommendations

The insurance data was analyzed using a multiple linear regression model. However, there are other models that could be used that may possibly provide different results and insights. Some of these models include non-linear regression models such as a quadratic model. Examining the effects of the of the interactions between the regressors could also give further insights into the data.

Appendices

```
knitr::opts_chunk$set(echo = TRUE)
library(gridExtra)
library(kableExtra)
library(psych)
insurances <- read.csv("insurance1.csv")
knitr::opts_chunk$set(echo = TRUE)

 knitr::kable(insurances[1:30,1:7], "latex",caption = 'Insurance Data')  %>%
  kable_styling(latex_options = "HOLD_position")

 df<-describe(insurances,fast = TRUE)

df1<-kable(x = df,
      col.names = c("attr.no", "n", "Mean","sd","min","max","range","se"),
      caption = "Insurance Data Statistics") %>%
  kable_styling(latex_options = "HOLD_position")
df1

insurances$sex<- as.factor(insurances$sex)
insurances$smoker<- as.factor(insurances$smoker)
insurances$region<- as.factor(insurances$region)
med.insurance <- lm(charges ~., data = insurances)

var((insurances$charges))

hist(insurances$charges,main ="Histogram of Charges variable",xlab="charges")
```

```
 par(mfrow=c(2,2))
plot(med.insurance)
```

```
par(mfrow=c(1,1))

library(ggplot2)
  q<-ggplot(insurances, aes(x=sex, y=charges)) +
    geom_boxplot()
w<- ggplot(insurances, aes(x=smoker, y=charges)) +
    geom_boxplot()
e<-ggplot(insurances, aes(x=region, y=charges)) +
    geom_boxplot()
grid.arrange(q, w,e, ncol=2)
```

```
smi<-summary(med.insurance)
smi

library(car)
avPlots(med.insurance)
```

```
df2<-cor(insurances[c("age", "bmi", "charges")])
dfc<-kable(x = df2,
      col.names = c("Age", "bmi", "charges"),
      caption = "Insurance Data Statistics") %>%
  kable_styling(latex_options = "HOLD_position")
dfc
```

References

[1] Annual Healthcare Spending Attributable to Cigarette Smoking https://www.ajpmonline.org/article/S0749-3797(14)00616-3/fulltext,December 09, 2014.

[2] Geographic Variation in Health Insurance Coverage: United States, 2020 https://www.cdc.gov/nchs/data/nhsr/nhsr168.pdf,February 11, 2022.

[3] About Adult BMI https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html,June 03, 2022.