Multivariate Linear Regression Project for the Insurance Dataset

Math 6010
Janaan Lake
u0987016

Introduction

The goal of this project is to create a linear model for the data in a medical insurance dataset[1] analyze how well the model fits the data, and draw conclusions about what affects insurance costs the most based on the results. This dataset includes data on insurance charges, age, BMI, sex, region, number of children and smoking status for over 1,300 insurance beneficiaries. Several linear models were created for this dataset, and they are explained in this paper.

Visualization

The first step to creating a linear model involves graphing the dataset to learn more about its characteristics and to get a basic overview of the relationships among the variables.
Violin plots were created to explore the relationship between the insurance charges and categorical/discrete variables, such as sex, region, smoking status and number of children. See Figures 1-4.

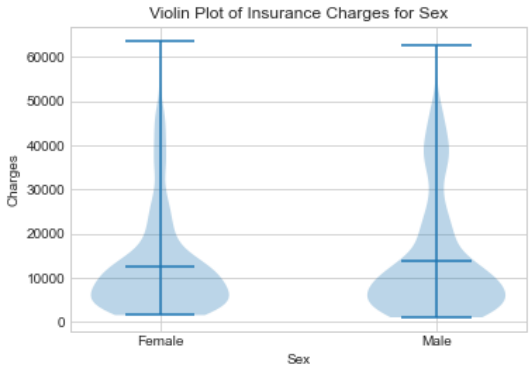

Figure 1- Insurance Costs by Sex

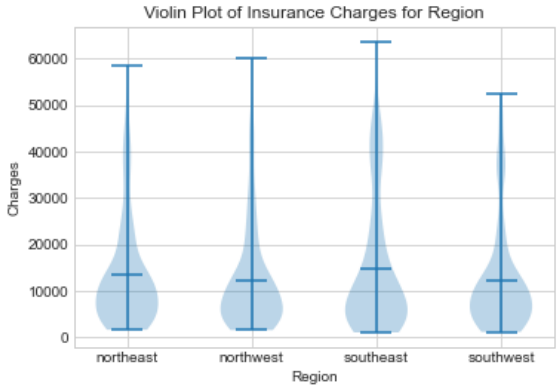

Figure 2 – Insurance Costs by Region

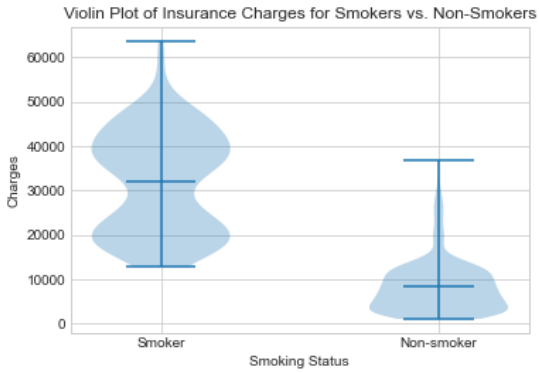

Figure 3 – Insurance Costs by Smoking Status

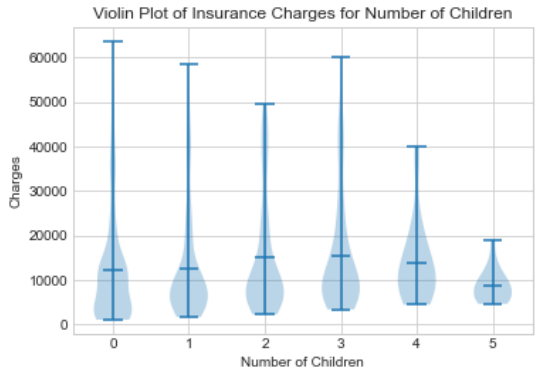

Figure 4 – Insurance Costs by Number of Children

[1] https://www.kaggle.com/datasets/mirichoi0218/insurance

Looking at the figures, the difference among categories is most pronounced for the smoking status. The mean insurance costs between these two categories differs significantly. The data also appears to have a bimodal distribution when the smoking status is positive. The rest of the categories have slight differences in their means and distributions, with men having a small increase in charges from women, the western regions having slightly lower insurance costs, and surprisingly the highest number of children resulting in the lowest insurance costs when compared to individuals with fewer children.

Next, the data of more continuous regressors, such as age and BMI were plotted using scatter plots. As can be seen in Figures 5, there is a correlation between age and insurance costs. However, age isn't necessarily the only factor as can be seen from the three distinct linear patterns that the data appears to follow. The BMI appears to have less correlation to the insurance costs as the scatter plot appears "messier".

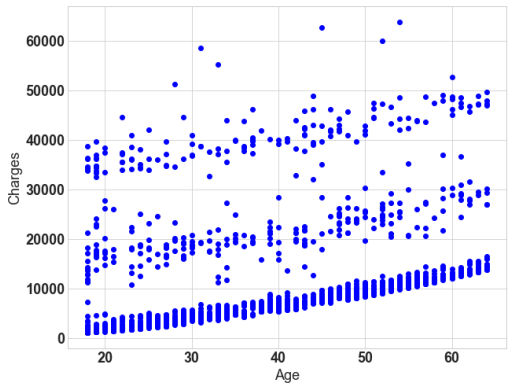

Figure 5 – Insurance Costs by Age

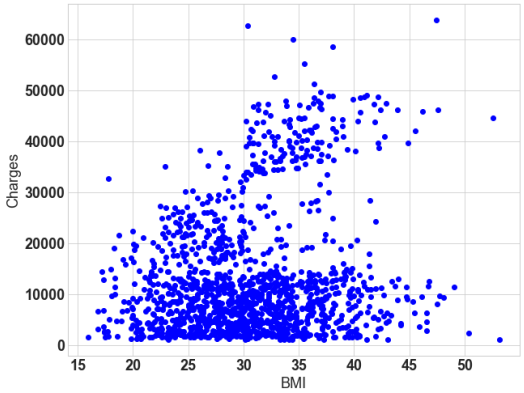

Figure 6 – Insurance Costs by BMI

To explore what other regressors may be affecting the relationship between costs and age, the age was broken out by category and then plotted. See Figures 7-9.

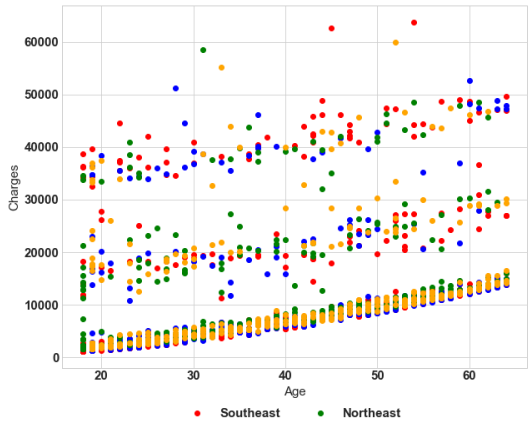

Figure 7 – Insurance Costs by Age and Regions

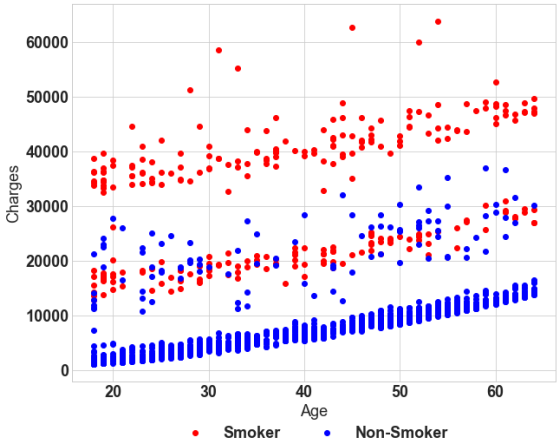

Figure 8 – Insurance Costs by Age and Smoking Status

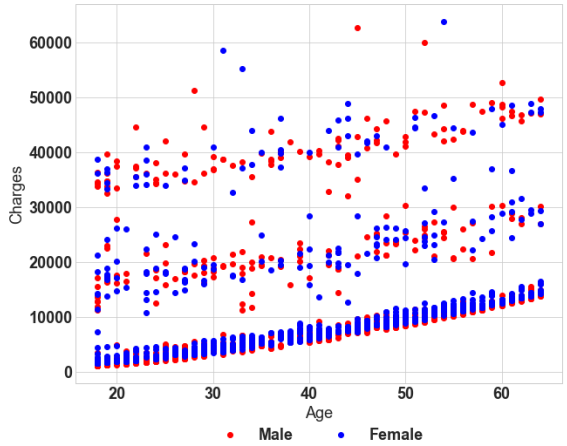

Figure 9 – Insurance Costs by Age and Sex

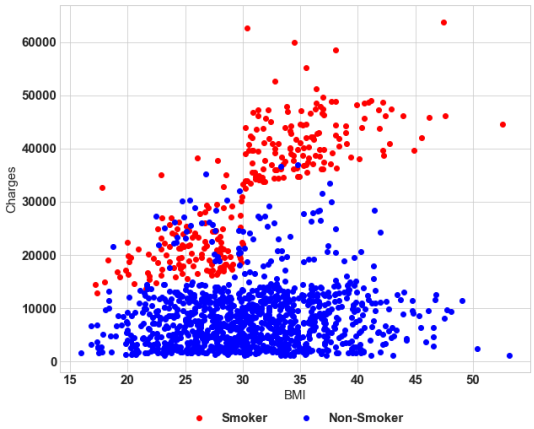

Figure 10 – Insurance Costs by BMI and Smoking Status

The region and sex didn't appear to have any noticeable linear relationship to the age and insurance costs. However, the smoking status obviously did. The top and bottom linear segments were obviously distinguished by smoking status, with the middle line segment being composed of both categories. This suggests that the individual's with the highest insurance costs and hence most likely the unhealthiest are those that are older and smokers. The individuals with the lowest insurance costs are those that are younger and non-smokers. However, that isn't the complete picture as there are a segment of people in the middle of the cost spectrum where smoking status isn't necessarily a determining factor. This can be seen by observing the middle linear segment in Figure 8, which is composed of both smokers and non-smokers. This implies that are other factors that can affect the cost for these individuals as well.

Lastly, the relationship between BMI and smoking status was examined since the smoking status appeared to have a strong influence on the insurance costs when combined with age. Figure 10 shows these results. Again, smoking status appears to segment the data and create a more defined linear relationship with the BMI data and insurance costs than the BMI data alone. This suggests that when smoking status is combined with other regressors, the linear relationships become more well-defined and is a strong predictor for insurance costs.

These data visualizations created a starting point for analyzing the data and exploring the linear relationships between the various regressors and the insurance costs. The next step involved creating a model by estimating the parameters of a linear model and then analyzing the residuals of the models.

Parameter Estimation

The continuous response variable (insurance charges) was modeled as a linear function of n numeric variables. The regressor variables, or x_i, consist of continuous, discrete and categorical variables. These variables are listed along with their type and range in Table 1. Note that the categorical variables are represented by $k - 1$ variables, where k is the number of categories. For example, the region category has 4 categories (Northeast, Northwest, Southeast, Southwest)

and is represented by three regressors as shown in Table 1. The last category is represented by a 0 value in the other region variables.

Regressor Name	Variable	Type	Range
Age	x_1	Discrete	18 - 64
Sex	x_2	Categorical	1 = male, 0 = female
BMI	x_3	Continuous	15.9 – 53.1
Children	x_4	Discrete	0-5
Smoker	x_5	Categorical	1 = smoker, 0 = nonsmoker
Northeast	x_6	Categorical	1 = northeast region, 0*
Northwest	x_7	Categorical	1 = northwest region, 0*
Southeast	x_8	Categorical	1 = southeast region, 0*
Southwest	None	Categorical	*Northeast=0, *Northwest=0, *Southeast=0

Table 1: List and types of Regressors

A linear model of the form

$$y_i = \beta_0 + \beta_1 x_1 +, \dots, +\beta_n x_n + \epsilon_i$$

was used. The matrix form is

$$Y = X\beta + \epsilon$$

where

$$Y = (y_1, y_2, \dots, y_n)^\top, \epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^\top$$

$$X = \begin{pmatrix} x_1{}^\top \\ x_2{}^\top \\ \vdots \\ x_n{}^\top \end{pmatrix}$$

β is the parameter that is estimated using the following closed-form solution:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

$\hat{\beta}$ is an unbiased estimator for β. In this dataset, X is a 1,338 x 9 matrix, with 1,338 samples and 9 regressors variables (including the intercept). The $X^\top X$ matrix is non-singular and so $\hat{\beta}$ can be calculated by the given formula, and the estimates of β are shown in Table 2.
The variance of $\hat{\beta}$ is $\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ where S_n^2 is an unbiased estimator for σ^2. We can calculate S_n^2as follows:

$$S_n^2 = \frac{1}{n-d} \left\| \boldsymbol{Y} - \boldsymbol{X}\hat{\beta} \right\|^2$$

The standard error of $\widehat{\beta_1}$ is therefore defined as σ multiplied by the square root of the *i*th diagonal entry of $(X^\top X)^{-1}$. The standard error gives an idea of the accuracy of $\widehat{\beta_i}$ as an estimator of β_i. The standard errors of each $\widehat{\beta_i}$ are shown in Table 2.

Regressor Name	Beta	Beta Value	Standard Error
Intercept	$\widehat{\beta_0}$	-12,899	1,021
Age	$\widehat{\beta_1}$	257	12
Sex	$\widehat{\beta_2}$	-131	333
BMI	$\widehat{\beta_3}$	339	29
Number of children	$\widehat{\beta_4}$	476	138
Smoking status	$\widehat{\beta_5}$	23,849	413
Northeast	$\widehat{\beta_6}$	960	478
Northwest	$\widehat{\beta_7}$	607	477
Southeast	$\widehat{\beta_8}$	-75	471

Table 2: Estimated beta values and standard errors for regressors

Residuals

When using the linear model described above, several assumptions are made. Some of these assumptions are related to the error terms, ϵ_i and include the normality, independence and constant variance of the error terms. Since we cannot directly observe the error terms, the residuals ε_i become a way of mimicking the error terms and observing if these conditions are met. The residuals $\hat{\varepsilon}$ are calculated as $Y - X\hat{\beta}$ and are plotted in the histogram below:

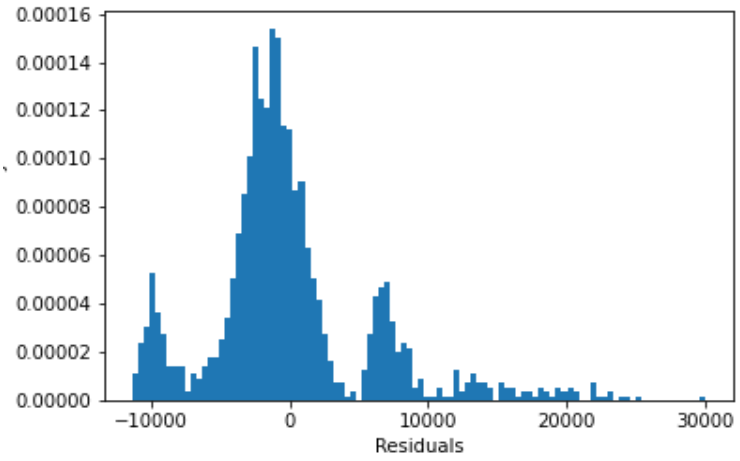

Figure 11: Distribution of residuals

As can be seen, they are not normally distributed so we cannot assume the error terms are normally distributed as well. Next, we plot the residuals against the fitted values $\hat{Y}$. This plot is shown in Figure 12.

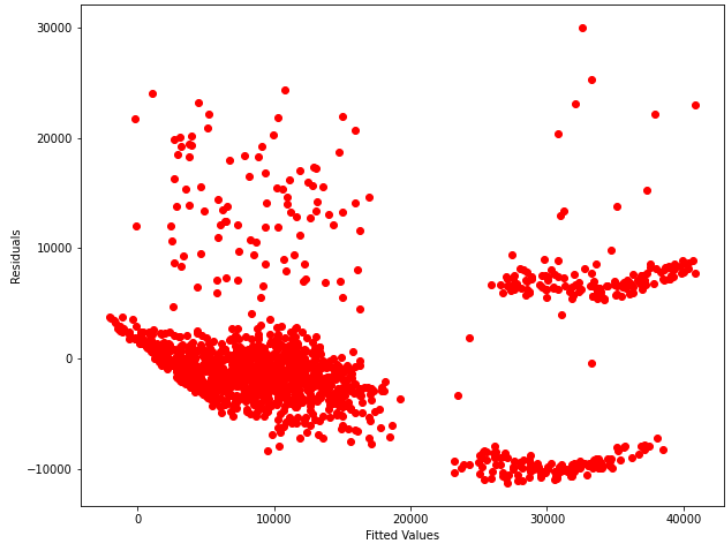

Figure 12: Residuals plotted against Fitted Values

In a correct model, the residuals should be somewhat equally centered around a 0-mean line and display random fluctuations. When that doesn't occur, either the model is wrong, one or many of the assumptions are not met. In this case, the residuals are do not appear to be random, and they are not evenly centered around a 0 mean. Therefore, we cannot assume the model is correct and that the model assumptions are met.

We further examine the residuals by plotting them against the age and BMI variables. These plots are shown in Figures 13 and 14. When we examine these plots, we can see that the behavior of the residuals when plotted against the BMI variable displays some behavior that does not appear to be random and should be examined further.

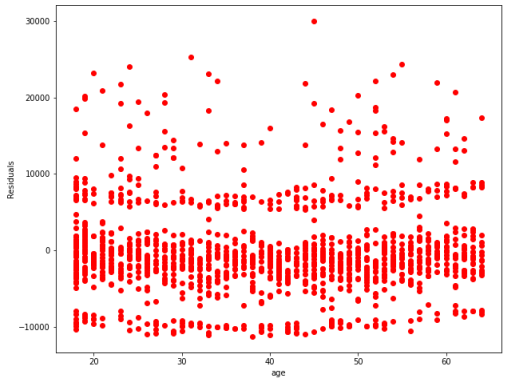

Figure 13: Residuals plotted against age

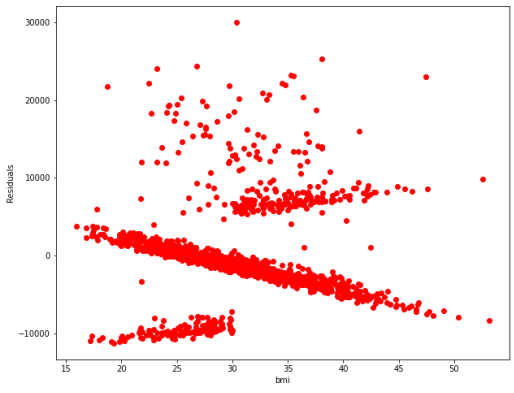

Figure 14: Residuals plotted against BMI

Further Analysis

Because the residuals displayed behavior that wasn't expected given the assumptions, the data was further analyzed to determine if the reason for the residual behavior could be explained and corrected. The smoking status appeared to have a significant effect on the data as can be seen from Figures 3, 8, and 10. Therefore, the dataset was split into two groups based on the smoking status, and a separate model was created for each dataset.

Nonsmoker Status

First, we examined the dataset where the smoking status was nonsmoker. This reduced the dataset to 1,064 individuals. The same parameter estimates as was done for the full dataset were performed, and the residuals were examined. A histogram plot of the residuals is shown in Figure 15. Next, the residuals were plotted against the fitted values, age, and BMI in Figures 16 – 18.

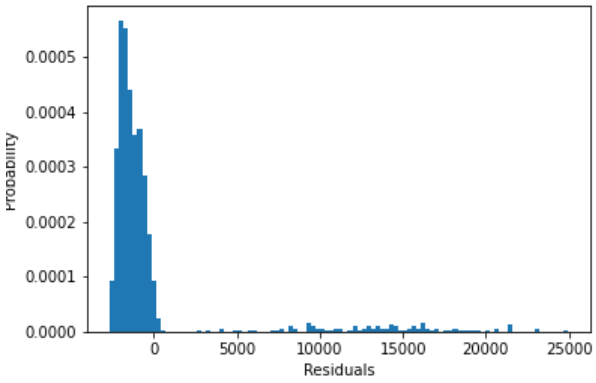

Figure 15: Distribution of Residuals for Nonsmoker Status

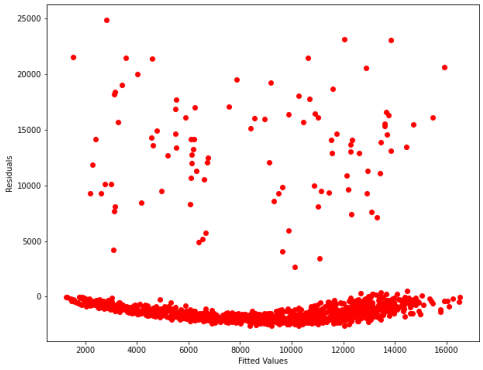

Figure 16: Residuals plotted against Fitted Values for Nonsmoker Status

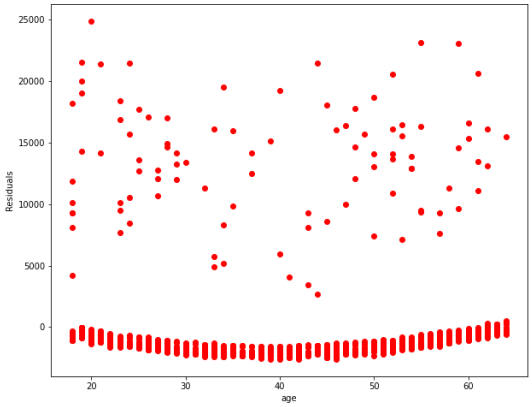

Figure 17: Residuals plotted against age for Nonsmoker Status

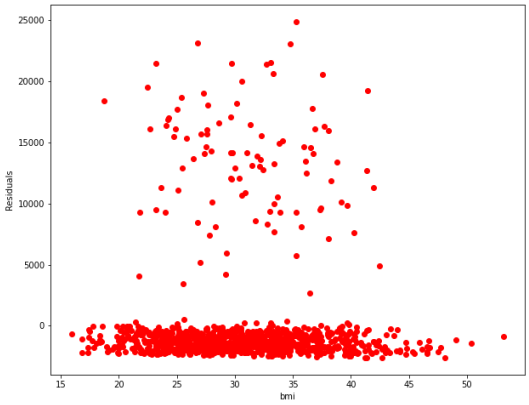

Figure 18: Residuals plotted against BMI for Nonsmoker Status

The distribution of residuals now looks more like a normal distribution, although it has a very long tail. When plotted against the fitted values, age, and BMI the residuals exhibit behavior that is more expected given the assumptions of normality of the errors. The points in the top half of each of Figures 16-18 is due to the long tail seen in the Figure 15.

Smoker Status

Next, the dataset was filtered for the smoking status being smoker. This reduced the dataset to 274 individuals. The same analysis was performed, and the results are shown in Figures 19-22.

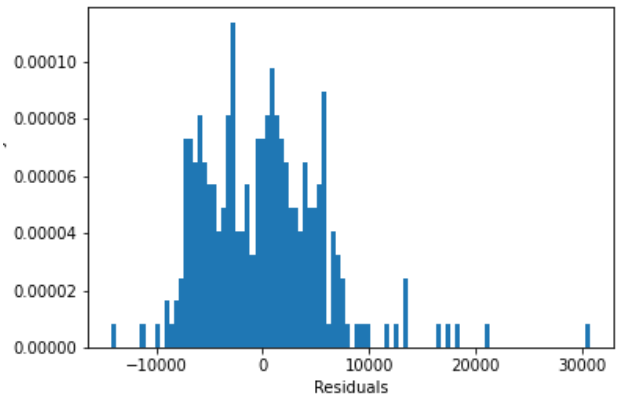

Figure 19: Distribution of Residuals for Smoker Status

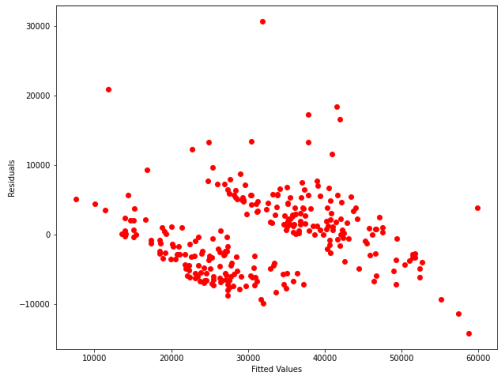

Figure 20: Residuals plotted against Fitted Values for Smoker Status

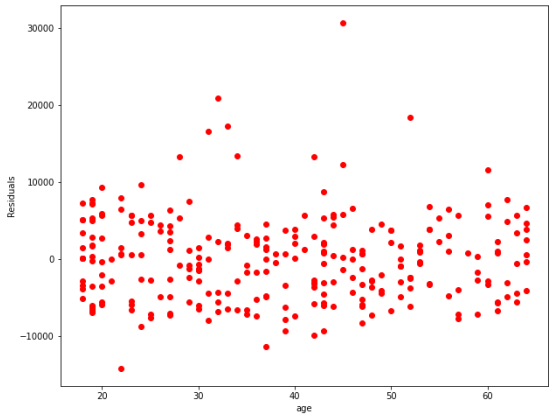

Figure 21: Residuals plotted against age for Smoker Status

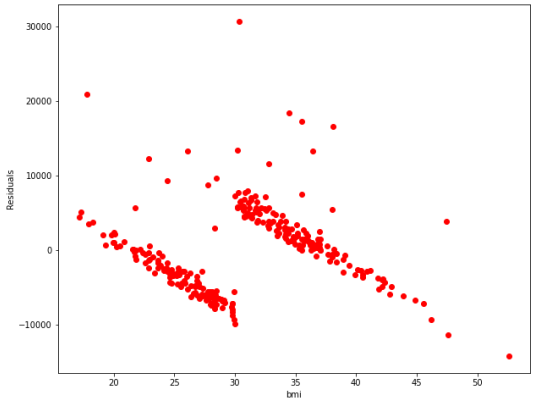

Figure 22: Residuals plotted against BMI for Smoker Status

In Figure 19 the residuals appear to have a bimodal distribution. This is also reflected in Figs 20 and 21. Further examination of Figure 22 shows a distinct difference in the residuals at a BMI of 30. Therefore, an additional parameter was added to the smoker dataset. This additional parameter is an indicator variable where a 0 value indicates a BMI level less than 30 and a value of 1 indicates a BMI value greater than or equal to 30. With this additional parameter, the same analysis was performed, and the results are plotted in Figures 23-26.

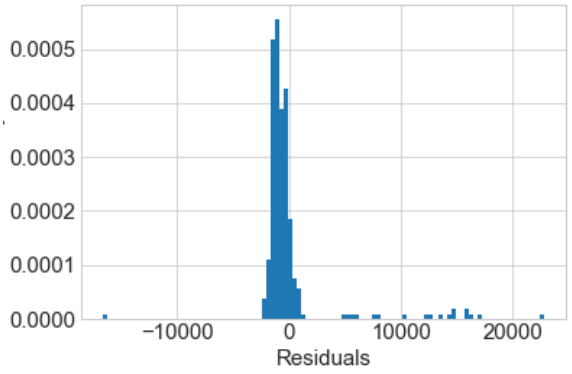

Figure 23: Distribution of Residuals for Smoker Status and Additional BMI parameter

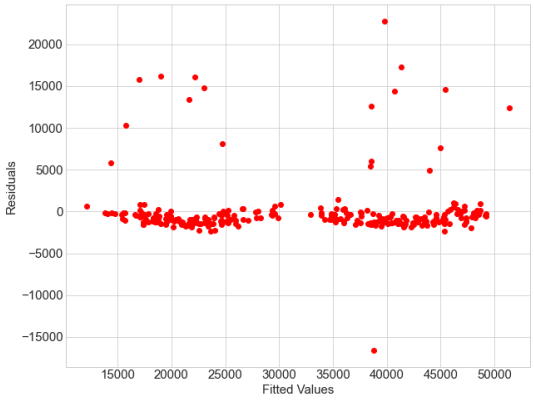

Figure 24: Residuals Plotted against Fitted Values

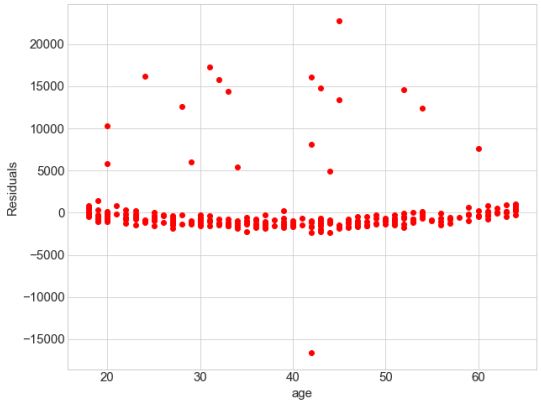

Figure 25: Residuals plotted against Age

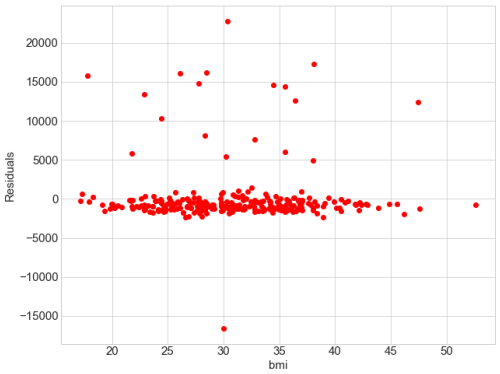

Figure 26: Residuals plotted against BMI

The distribution of the residuals looks more normally distributed, albeit with long tails. The other residual plots also display more expected behavior for normality assumptions. Therefore, given these results, the dataset was split into two separate datasets based on smoker status, and an extra regressor was added to the smoker dataset based on BMI level.

Outlier Detection

The long tails in the distributions suggest that there may be outliers in the dataset. An outlier is a point with a large residual. An influential point is a point that has a large impact on the regression. Outliers that are not high-leverage points do not have a very strong influence on the fitted regression plane unless they are very large. The situation is different when outliers are also high-leverage points. To detect outliers, a studentized residual was used. A studentized

residual is the quotient resulting from the division of a residual by an estimate of its standard deviation. The following formula was used to calculate the studentized residuals:

$$t_i = r_i \left(\frac{n-p-2}{n-p-1-r_i^2} \right)^{1/2}$$

where r_i is the *i*th standardized residual, n is the number of observations, p is the number of parameters, and t_i is a student's t distribution with $n-p$ degrees of freedom. The standardized residuals are calculated as follows:

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

Where e_i is the *i*th residual and h_{ii} is the *i*th diagonal of the hat matrix, $X(X^\top X)^{-1}X^\top$. Any point with a studentized residual greater than the critical value using a very conservative value for $\alpha = 0.001$ was flagged as an outlier. Using this metric, 47 and 12 outliers were respectively found in the nonsmoker and smoker datasets and were removed.

Cook's distance is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis. Cook's distance is calculated as follows:

$$D_i = \left(\frac{r_i^2}{p+1} \right) \left(\frac{h_{ii}}{1-h_{ii}} \right)$$

If a point has a Cook's distance > 1, it is typically considered a high-leverage point. Using this metric, none of the points in the three datasets were found to be high-leverage points.

Hypothesis Testing

Once the outliers were identified and removed, the next step in the regression analysis was to recalculate the parameters and test each of them individually for significance. The beta values for the two datasets and the standard errors for each beta are listed in Tables 3-5.

Nonsmoker			
Regressor Name	**Beta**	**Beta Value**	**Standard Error**
Intercept	$\widehat{\beta_0}$	-4,370	504
Age	$\widehat{\beta_1}$	266	6
Sex	$\widehat{\beta_2}$	-377	164
BMI	$\widehat{\beta_3}$	24	14
Number of children	$\widehat{\beta_4}$	499	67
Northeast	$\widehat{\beta_5}$	1,042	232
Northwest	$\widehat{\beta_6}$	355	232
Southeast	$\widehat{\beta_7}$	462	233

Table 3: Estimated beta values and standard errors for regressors

Smoker			
Regressor Name	**Beta**	**Beta Value**	**Standard Error**
Intercept	$\widehat{\beta_0}$	-1,420	698
Age	$\widehat{\beta_1}$	264	6
Sex	$\widehat{\beta_2}$	-625	182
BMI	$\widehat{\beta_3}$	473	25
BMI Level	$\widehat{\beta_4}$	15,262	304
Number of children	$\widehat{\beta_5}$	388	77
Northeast	$\widehat{\beta_6}$	83	265
Northwest	$\widehat{\beta_7}$	400	276
Southeast	$\widehat{\beta_8}$	-212	250

Table 4: Estimated beta values and standard errors for regressors

Next, hypothesis testing was conducted on each of the parameters, β_i. Each individual parameter was tested using the null hypothesis for each β_i,

$$H_0: \beta_i = 0$$

and the f score was calculated as:

$$f = \frac{\left(\widehat{\beta_i}\right)^2}{S^2 d_{ii}}$$

where d_{ii} is the i^{th} element of the diagonal of $(X^\top X)^{-1}$ and f is a $f(1, n-p)$ distribution. This test was used because the normality conditions appeared to be met based on the plots of the residuals. Using a value of $\alpha = 0.05$, the results of the f-tests are shown in Tables 5 and 6 for each dataset respectively.

Nonsmoker				
Regressor Name	**Beta**	**Beta Value**	**p-value**	**Reject Null?**
Intercept	$\widehat{\beta_0}$	-4,370	0.000	Yes
Age	$\widehat{\beta_1}$	266	0.000	Yes
Sex	$\widehat{\beta_2}$	-377	0.022	Yes
BMI	$\widehat{\beta_3}$	24	0.100	No
Number of children	$\widehat{\beta_4}$	499	0.000	Yes
Northeast	$\widehat{\beta_5}$	1,042	0.000	Yes
Northwest	$\widehat{\beta_6}$	355	0.127	No
Southeast	$\widehat{\beta_7}$	462	0.047	Yes

Table 5: p-values for Null Hypothesis of Individual Parameters

Smoker				
Regressor Name	**Beta**	**Beta Value**	**p-value**	**Reject Null?**
Intercept	$\widehat{\beta_0}$	-1,420	0.043	Yes
Age	$\widehat{\beta_1}$	264	0.000	Yes
Sex	$\widehat{\beta_2}$	-625	0.001	Yes
BMI	$\widehat{\beta_3}$	473	0.000	Yes
BMI Level	$\widehat{\beta_4}$	15,262	0.000	Yes
Number of children	$\widehat{\beta_5}$	388	0.000	Yes
Northeast	$\widehat{\beta_6}$	83	0.754	No
Northwest	$\widehat{\beta_7}$	499	0.148	No
Southeast	$\widehat{\beta_8}$	-212	0.397	No

Table 6: p-values for Null Hypothesis of Individual Parameters

For the smoker dataset, the null hypothesis was rejected for all the regressors except BMI and Northwest, which means these regressors were not deemed significant. Dropping them from the model is something to consider. For the smoker model, all the regressor variables were deemed significant except the regional variables. This implies that the region of the individual doesn't have a significant effect on the insurance costs or that the effect is overwhelmed by the strong influences of the other regressor variables for smokers.

Confidence Intervals for Parameter Estimation

The last step in the regression analysis was to calculate confidence intervals for all of the parameters. Again, the normality of the error terms was assumed, so the following formula was used for calculating the confidence intervals:

$$\beta_i \pm t_{n-p}\big(\alpha/(2 \cdot \mathrm{p})\big) S d_{ii}^{1/2}$$

where d_{ii} is the i^{th} element of the diagonal of $(X^\top X)^{-1}$ and t is a $t(n-p)$ distribution. Note that the Bonferroni correction was used, and $1-\alpha$ is a lower bound for all the confidence intervals. Using a value of $\alpha = 0.05$, the following intervals were calculated for each β_i:

Nonsmoker				
Regressor Name	**Beta**	**Beta Value**	**Lower C.I.**	**Upper C.I.**
Intercept	$\widehat{\beta_0}$	-4,370	-5,750	-2,991
Age	$\widehat{\beta_1}$	266	249	282
Sex	$\widehat{\beta_2}$	-377	-826	72
BMI	$\widehat{\beta_3}$	24	-16	62
Number of children	$\widehat{\beta_4}$	499	315	683
Northeast	$\widehat{\beta_5}$	1,042	400	1,684
Northwest	$\widehat{\beta_6}$	355	-282	992
Southeast	$\widehat{\beta_7}$	462	-174	1,101

Table 7: Confidence Intervals for the Regression Coefficients

Smoker				
Regressor Name	**Beta**	**Beta Value**	**Lower C.I.**	**Upper C.I.**
Intercept	$\widehat{\beta_0}$	-1,420	-3,371	532
Age	$\widehat{\beta_1}$	264	246	282
Sex	$\widehat{\beta_2}$	-625	-1,133	-116
BMI	$\widehat{\beta_3}$	473	403	543
BMI Level	$\widehat{\beta_4}$	15,262	14,410	16,114
Number of Children	$\widehat{\beta_5}$	388	74	602
Northeast	$\widehat{\beta_6}$	83	-658	825
Northwest	$\widehat{\beta_7}$	499	-371	1,171
Southeast	$\widehat{\beta_8}$	-212	-911	487

Table 8: Confidence Intervals for the Regression Coefficients

Although normality assumptions were used for the confidence interval estimation for the regression coefficients, there are sampling methods that do not rely on the normality assumptions. The method of bootstrapping was explored to see how the confidence intervals generated by this method compared to those calculated above.

Bootstrapping

Bootstrapping is a method of random sampling with replacement. The bootstrapping approach does not violate or bypass the normality assumptions, but rather it relies on the Central Limit Theorem. This means that the distribution of the bootstrapped samples is asymptotically normal.

The residual bootstrapping method was used to sample distributions of the variances for the estimated beta values. These variances were then used to compute confidence intervals for each estimated beta value.

Recall that the residuals $\hat{\varepsilon}$ are calculated as $Y - X\hat{\beta}$. The residual bootstrap first generates IID

$$\widehat{\varepsilon_1}^*, \ldots, \widehat{\varepsilon *_n}$$

such that for each $\widehat{\varepsilon_i^*}$

$$P(\widehat{\varepsilon_i^*} = \widehat{\varepsilon_i^*}) = \frac{1}{n}, \qquad \forall n = 1, \ldots, n$$

And then a new bootstrap sample

$$(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$$

Is generated via

$$x_i^* = x_i, y_i^* = \widehat{\beta_0} + \widehat{\beta_1} x_i + , \ldots, + \widehat{\beta_n} x_n + \widehat{\varepsilon_i^*}$$

Namely, the regressor x_i is fixed but a new value of y_i is generated using the fitted regression function and the sampled residual. This process is repeated M times.

For each bootstrap sample $\left(x_i^{*(l)}, y_1^{*(l)}\right), \ldots, \left(x_n^{*(l)}, y_n^{*(l)}\right)$ a linear regression model is fitted, *l*eading to a bootstrap estimate of the fitted coefficients $\widehat{\beta_0^{*(l)}}, \widehat{\beta_1^{*(l)}}, \ldots, \widehat{\beta_n^{*(l)}}$. The M bootstrap samples generate M estimates of β_i:

$$\left(\widehat{\beta_0^{*(1)}}, \widehat{\beta_1^{*(1)}}, \dots, \widehat{\beta_n^{*(1)}}\right), \dots, \left(\widehat{\beta_0^{*(M)}}, \widehat{\beta_1^{*(M)}}, \dots, \widehat{\beta_n^{*(M)}}\right)$$

The variance is then estimated by

$$\widehat{Var_M}(\widehat{\beta_i}) = \frac{1}{M}\sum_{l=1}^{M}\left(\widehat{\beta_i^{*(l)}} - \overline{\beta_i^*}\right)^2, \qquad \overline{\beta_i^*} = \sum_{l=1}^{M}\widehat{\beta_i^{*(l)}}$$

The confidence intervals can be constructed as follows:

$$\text{C. I. }(\beta_i) = \widehat{B_i} \pm z_{1-\alpha/(2*\text{p})}\sqrt{\widehat{Var_M}(\widehat{\beta_i^*})}$$

This follows from the fact that the fitted coefficients β_i are roughly normally distributed around the true values β_i.

With the calculations described above, sampling 10,000 times, using a value for $\alpha = 0.05$ and a Bonferroni correction, confidence intervals were derived for each estimated β_i value and are shown in Tables 9 & 10.

Nonsmoker				
Regressor Name	**Beta**	**Beta Value**	**Lower C.I.**	**Upper C.I.**
Intercept	$\widehat{\beta_0}$	-4,370	-5,744	-2,996
Age	$\widehat{\beta_1}$	266	249	282
Sex	$\widehat{\beta_2}$	-377	-819	65
BMI	$\widehat{\beta_3}$	24	-15	62
Number of children	$\widehat{\beta_4}$	499	317	681
Northeast	$\widehat{\beta_5}$	1,042	405	1,679
Northwest	$\widehat{\beta_6}$	355	-278	989
Southeast	$\widehat{\beta_7}$	462	-171	1,097

Table 9: Confidence Intervals for the Regression Coefficients using Bootstrapping Method

Smoker				
Regressor Name	**Beta**	**Beta Value**	**Lower C.I.**	**Upper C.I.**
Intercept	$\widehat{\beta_0}$	-1,420	-3,338	498
Age	$\widehat{\beta_1}$	264	247	282
Sex	$\widehat{\beta_2}$	-625	-1,116	-132
BMI	$\widehat{\beta_3}$	473	405	541
BMI Level	$\widehat{\beta_4}$	15,262	14,439	16,085
Number of Children	$\widehat{\beta_5}$	388	178	597
Northeast	$\widehat{\beta_6}$	83	-638	805
Northwest	$\widehat{\beta_7}$	499	-360	1,161
Southeast	$\widehat{\beta_8}$	-212	-892	468

Table 10: Confidence Intervals for the Regression Coefficients using Bootstrapping Method

While the confidence intervals generated by the bootstrapping method were usually a bit smaller, in general they were very similar to those calculated by the t-distribution. This is reassuring!

Conclusion

Based on the regression analysis described in this paper, several conclusions can be drawn from the dataset. The most important factors that affect insurance costs appear to be smoking status and BMI. Based on these models, smoking status adds an estimated $2,950 to the average annual insurance cost. BMI doesn't have a significant affect on costs for the nonsmoker dataset. But that is a different story for smokers. For these individuals, the BMI level makes a significant difference in costs. Smokers who have a BMI greater than 30 incur an estimated $15,262 in costs per year over smokers with a lower BMI. That is a large increase! Also, for each unit increase in BMI, smokers have an additional $473 in insurance costs. This can be seen in Figure 10, which displays a linear relationship between BMI and smokers and a lack of one for nonsmokers.

The other regressors had some affect, but none of them were nearly as significant as smoking status and BMI. Age affects both smokers and nonsmokers similarly as can be seen by the regression coefficient that is almost identical for each dataset (266 vs. 264). This makes sense by looking at Figure 3 since all the lines appear to have a similar slope. On average for every yearly increase in age, an additional $266 is incurred in insurance costs. The number of children appears to make a difference in costs, between $388 and $499 per child on average. Intuitively this makes sense as more children will increase the odds for more health needs and hence costs. However, this isn't a perfect model because in Figure 4 we can see that the average insurance costs for individuals with 5 children is less than those with any other number of children. Sex had some affect on costs, with women incurring higher costs ($377 to $625) than men. This makes sense as pregnancy costs for women most likely increase the average insurance costs when compared to men. Regional variations appeared more in the nonsmoker dataset. They were insignificant in the smoker dataset. This could be due to the fact that other factors, i.e. smoking status and BMI levels overshadowed any affects region might have had on the insurance costs.