Multiple Linear Regression Model to Predict Insurance Charges

By Jake Schinto
5/13/2020

In this report, I plan to uncover the secret of health insurance prices. Everyone needing insurance has a secret price that is determined through several hidden variables. I will look at a few of these variables and, using my knowledge of multiple linear regression and SAS, find the variables that are most important. To achieve this goal, I will use Exploratory Data Analysis to find all correlations in the data. Then I will do a correlation analysis to figure out for certain which attributes are worth pursuing. Lastly, I will assess a few plausible models and perform a model selection to ultimately find the best model. After going through these steps, I will determine that a log transformation is suitable and result in an equation of Log(Charges) = 6.91 + 0.035(age) + 0.013(bmi) + 0.1(children) + 0.75(sex) + 1.56(smoker) - 0.09(southeast). Although it appears to be a sound model, the analysis ends with a normality violation in the residuals that could make the model more prone to error.

Introduction

In the world's current state, health has become a major concern for people. The coronavirus has been especially costly on the average person because of the effects it has on people but also the limited testing can be expensive if someone is uninsured. According to the last count in 2018, 27.5 million Americans are uninsured; this makes the cost of testing a large burden on the uninsured. Costs to visit the doctor alone could be as much as $1,151 and adding the testing could make it as much as $3,270 [1]. These costs and public crises are all part of the complex equation that insurance companies need to take into account when deciding costs to cover an individual. In addition to my curiosity about the current pandemic, my family also has a long running background in healthcare. For many generations before me, my family has practiced dentistry and are therefore influenced by people who have insurance coverage. Lastly, I will soon graduate and be in need of health insurance, so knowing the factors that go into deciding costs are important for minimizing expenses. For these reasons, I have decided to take a closer look at the insurance dataset [2].

The dataset is split up into 7 different columns: age, sex, bmi (Body Mass Index, the ratio of weight to height), children, smoker (binary yes/no), region (quadrant of US), and charges (costs billed by insurance). Using these features, I hope to create a model using my knowledge of statistics and multiple linear regression (multiple independent variables) that will allow me to accurately predict a new given person's insurance costs based only on the features of that individual.

This ability to predict medical costs is important because it is the deciding factor of how much someone will have to spend on having health insurance. This information is therefore both valuable to the individual who will end up having to pay the bills as well as to the insurance companies themselves so that they can ensure that they are not losing money to a particularly pricy individual to cover.

When looking for others' work with this kind of data, I found a Medium article by Bayu Galih Prianda who documented his attempt at finding a linear regression model to use with health insurance data. He used Python libraries and multiple linear regression in order to find a best fit line in order to accurately predict the medical costs. His results came out to $y = -11676.830 + 259.547x_1 + 322.615x_2 + 23823.684x_3$, where ($y$ = charges, x_1 = age, x_2 = bmi, x_3 = smoking(0/1)). Immediately what strikes me as interesting about these findings is that smoking is equivalent to about 90 years of age, which seems wild and possibly could lead to alarm for any smokers looking to get insurance [3].

In conclusion, I look forward to experimenting myself with this data and finding the true factors behind what can cause a certain person to have a higher bill than others.

1. Exploratory Data Analysis

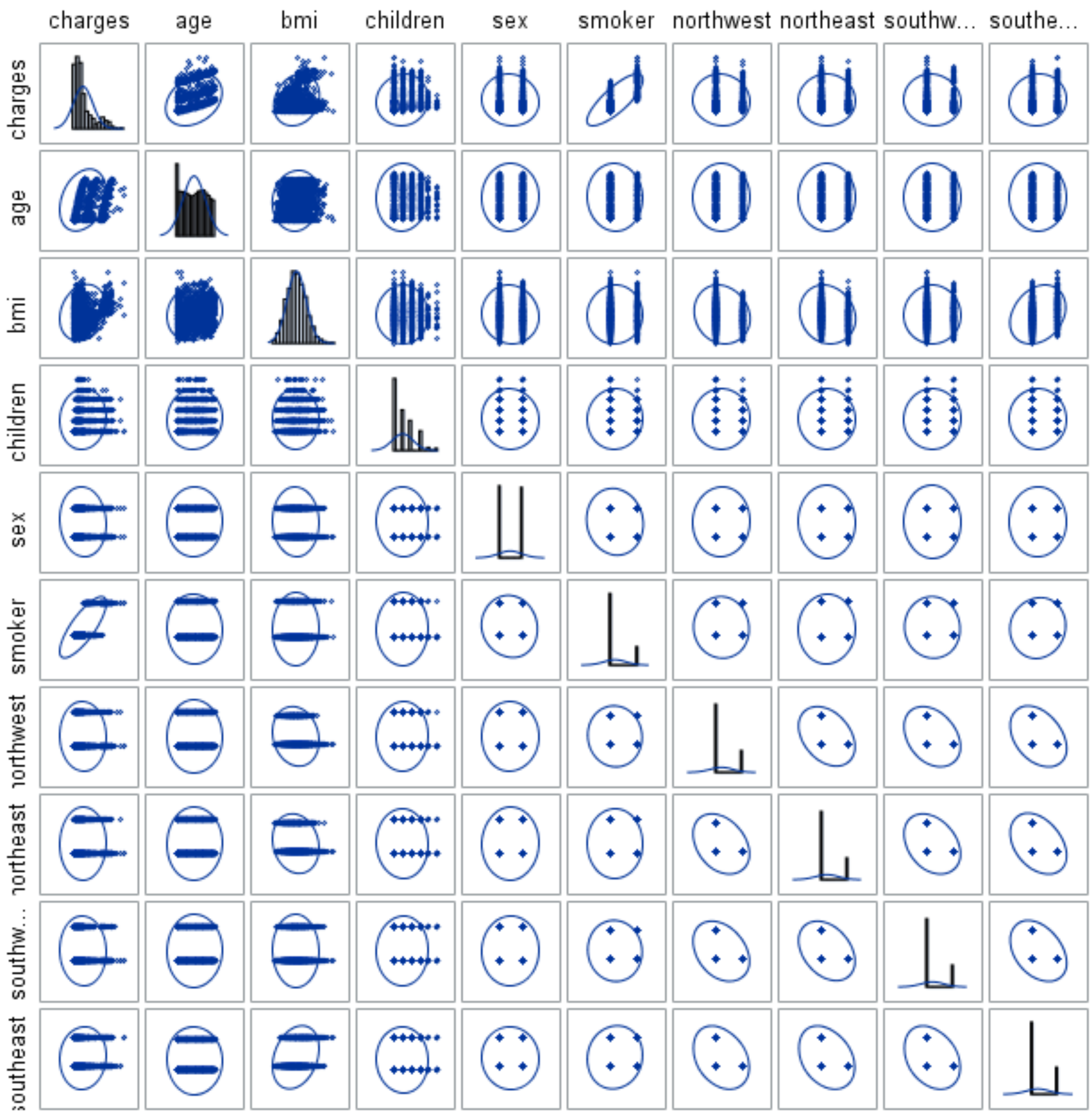

Figure 1.1 Scatter matrix for Charges, age, bmi, children, sex (Female = 1), smoker (yes = 1), and each region (in region = 1) with histogram on diagonal

In order to create this scatterplot, I first transformed each of the categorical attributes into a binary (0,1) field so that correlations can be spotted easier. By looking at the resulting scatterplot data, it becomes easily apparent that there are multiple correlations within this dataset. Specifically, charges appears to correlate with age, bmi, children, and smoker status. Looking at the histograms, it looks like there is a right skew in charges, bmi, and children.

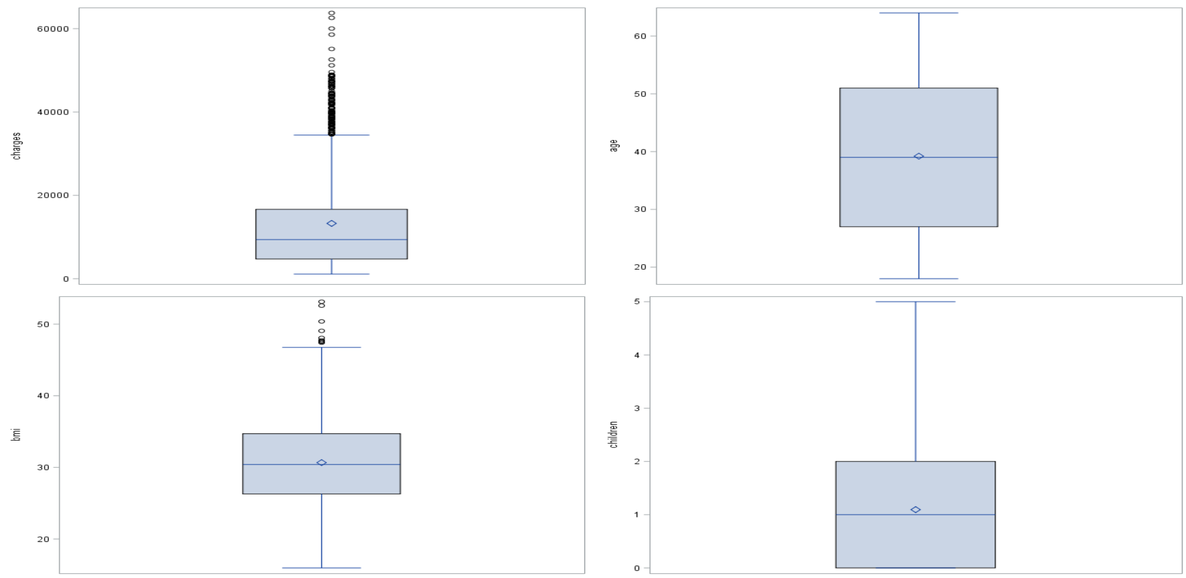

Figure 1.2 Boxplot for Charges (top left), Age (top right), BMI (bottom left), Children (bottom right)

The skewness predicted above seems fairly consistent with what the boxplots show. Charges and Children seem heavily skew, while BMI appears less skewed in the boxplots than it did in the histogram. All three skews are still apparent however.

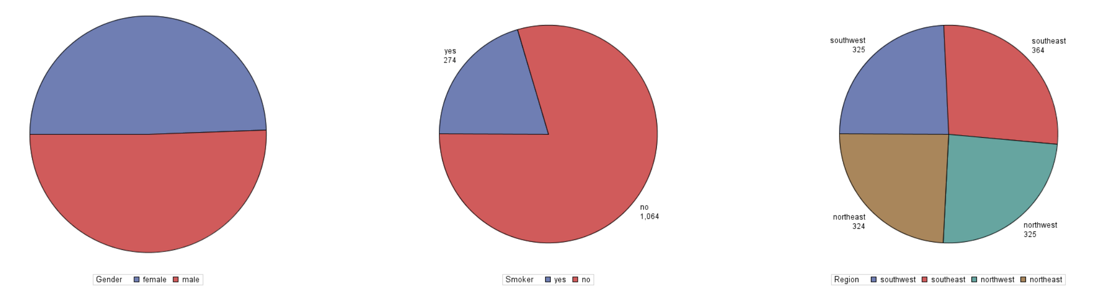

Figure 1.3 Pie Charts depicting the categorical features: Gender (left), Smoker (center), Region (right)

Above shows the distribution of the categorical features. The only surprising feature is smoker where only 20% is "yes." Every other feature seems to be very close to a perfect balance.

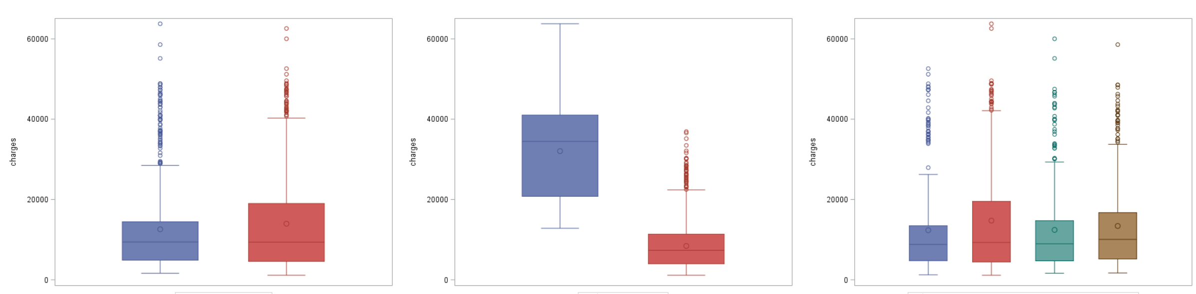

Figure 1.4 Boxplots to compare categorical features: Gender (left), Smoker (center), Region (right)

Looking at the boxplots for the categorical features, gender and region appear to have right skews and no real differentiation between individual categories. On the other hand, smoking appears to have a rather large difference in charge between the categories, which is unlikely to be caused solely from the smaller amount of test data for “yes.” This boxplot appears to back up the predictions made in the scatterplot about smoking’s rather large correlation to charge.

2. Correlation Analysis

Pearson Correlation Coefficients, N = 1338
Prob > |r| under H0: Rho=0

	age	bmi	children	charges	sex	smoker	northwest	northeast	southwest	southeast
age	1.00000	0.10927 <.0001	0.04247 0.1205	0.29901 <.0001	0.02086 0.4459	-0.02502 0.3605	-0.00041 0.9881	0.00247 0.9279	0.01002 0.7143	-0.01164 0.6705
bmi	0.10927 <.0001	1.00000	0.01276 0.6410	0.19834 <.0001	-0.04637 0.0900	0.00375 0.8910	-0.13600 <.0001	-0.13816 <.0001	-0.00621 0.8206	0.27002 <.0001
children	0.04247 0.1205	0.01276 0.6410	1.00000	0.06800 0.0129	-0.01716 0.5305	0.00767 0.7792	0.02481 0.3646	-0.02281 0.4045	0.02191 0.4232	-0.02307 0.3992
charges	0.29901 <.0001	0.19834 <.0001	0.06800 0.0129	1.00000	-0.05729 0.0361	0.78725 <.0001	-0.03990 0.1446	0.00635 0.8165	-0.04321 0.1141	0.07398 0.0068
sex	0.02086 0.4459	-0.04637 0.0900	-0.01716 0.5305	-0.05729 0.0361	1.00000	-0.07618 0.0053	0.01116 0.6835	0.00243 0.9294	0.00418 0.8785	-0.01712 0.5316
smoker	-0.02502 0.3605	0.00375 0.8910	0.00767 0.7792	0.78725 <.0001	-0.07618 0.0053	1.00000	-0.03695 0.1768	0.00281 0.9182	-0.03695 0.1768	0.06850 0.0122
northwest	-0.00041 0.9881	-0.13600 <.0001	0.02481 0.3646	-0.03990 0.1446	0.01116 0.6835	-0.03695 0.1768	1.00000	-0.32018 <.0001	-0.32083 <.0001	-0.34626 <.0001
northeast	0.00247 0.9279	-0.13816 <.0001	-0.02281 0.4045	0.00635 0.8165	0.00243 0.9294	0.00281 0.9182	-0.32018 <.0001	1.00000	-0.32018 <.0001	-0.34556 <.0001
southwest	0.01002 0.7143	-0.00621 0.8206	0.02191 0.4232	-0.04321 0.1141	0.00418 0.8785	-0.03695 0.1768	-0.32083 <.0001	-0.32018 <.0001	1.00000	-0.34626 <.0001
southeast	-0.01164 0.6705	0.27002 <.0001	-0.02307 0.3992	0.07398 0.0068	-0.01712 0.5316	0.06850 0.0122	-0.34626 <.0001	-0.34556 <.0001	-0.34626 <.0001	1.00000

Figure 2.1 Correlation Analysis Data for Charges, Age, BMI, Children, Sex (Female = 1), Smoker (Yes = 1), and each individual Region (In the region = 1)

Looking at the Correlation Data reported by SAS, it shows multiple possibly correlated values. All values <.05 have I high chance of correlation. Charges seems to correlate with age, bmi, children, smoker, and the southeast region. There are many other correlations as well, but I will not be focusing very heavily on them going forward. The most interesting of the extraneous correlations is the southeast region. It seems to correlate well with bmi and smoker, which might help explain the loose correlation it had with charges. Likewise, sex seems to correlate with smoking, so that is a likely explanation for the weak correlation.

3. Regression Analysis

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1.47083E11	24513836220	666.00	<.0001
Error	1331	48991204250	36807817		
Corrected Total	1337	1.960742E11			

Root MSE	6066.94461	R-Square	0.7501
Dependent Mean	13270	Adj R-Sq	0.7490
Coeff Var	45.71780		

Figure 3.1 ANOVA table for the first model

The F-test in the table shows a significant value of <.0001 returned. This value is below .05 which suggests that the full model should be considered. The Adjusted R-Sq value is 0.749.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-12354	970.44460	-12.73	<.0001	0
age	1	257.02132	11.90806	21.58	<.0001	1.01677
bmi	1	333.96314	28.48961	11.72	<.0001	1.09639
children	1	468.97792	137.84086	3.40	0.0007	1.00294
sex	1	129.19107	333.20800	0.39	0.6983	1.00888
smoker	1	23866	413.32561	57.74	<.0001	1.01130
southeast	1	-579.02918	388.50853	-1.49	0.1364	1.08659

Collinearity Diagnostics									
			Proportion of Variation						
Number	Eigenvalue	Condition Index	Intercept	age	bmi	children	sex	smoker	southeast
1	4.53133	1.00000	0.00134	0.00480	0.00157	0.01460	0.01398	0.01075	0.01234
2	0.79225	2.39156	0.00025630	0.00166	0.00021469	0.02008	0.06278	0.78790	0.04576
3	0.69337	2.55640	0.00011772	0.00070403	0.00000159	0.05722	0.00595	0.12365	0.77577
4	0.53280	2.91628	0.00008757	0.00026282	0.00002975	0.58712	0.37287	0.02045	0.01682
5	0.35681	3.56363	0.00768	0.05221	0.01034	0.31159	0.51569	0.04547	0.07807
6	0.07615	7.71402	0.05045	0.90528	0.11214	0.00481	0.00732	0.00584	0.03119
7	0.01728	16.19315	0.94007	0.03508	0.87570	0.00458	0.02143	0.00593	0.04006

Figure 3.2 Parameter Estimates and Collinearity Diagnostics

Above are SAS predictions for the first model. Importantly, the Variance Inflation column reports all values less than the major threshold of 10. In addition, the Condition Index does not get unusually large. For these reasons, we do not have to worry about any serious multicollinearity for now.

Durbin-Watson D	2.089
Pr < DW	0.9476
Pr > DW	0.0524
Number of Observations	1338
1st Order Autocorrelation	-0.046

Figure 3.3 Durbin-Watson test

The Pr<DW and Pr>DW values appear very close to 0.05 which is worrying, but given our threshold it still passes and we can continue without considering any significant autoregressive effects.

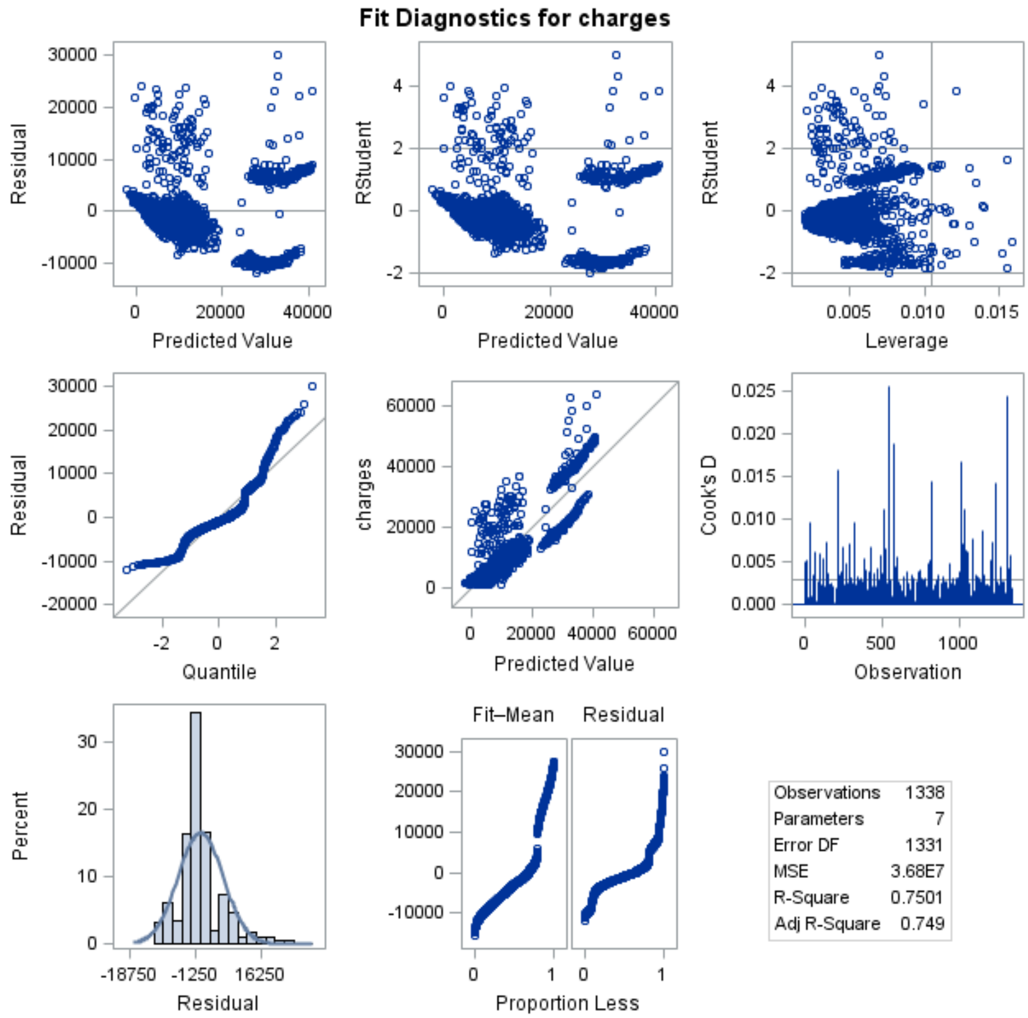

Figure 3.4 Diagnostics Panel for Model

The residuals seem to be moderately distributed however, you can still see an upward sloping curve in the residuals suggesting that there might be some evidence of heteroscedasticity violations.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.898538	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.162515	Pr > D	<0.0100
Cramer-von Mises	W-Sq	8.87245	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	44.58771	Pr > A-Sq	<0.0050

```
   D'AGOSTINO TEST OF NORMALITY FOR VARIABLE D, N=1338
G1=1.22668   SQRTB1=1.22530   Z=14.68996      P=0.0000
G2=2.70898   B2=5.69438       Z= 9.45976      P=0.0000
K**2=CHISQ(2 DF)=305.2819                     P=0.0000
```

Figure 3.5 Normality Test for the full model

Finally, looking at the normality information given it would appear that the residuals do not abide by a normal distribution. This is given by every p-value being <0.05, so the model fails in normality.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	865.46947	144.24491	724.48	<.0001
Error	1331	265.00429	0.19910		
Corrected Total	1337	1130.47376			

Root MSE	0.44621	R-Square	0.7656
Dependent Mean	9.09866	Adj R-Sq	0.7645
Coeff Var	4.90411		

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	2374093	395682	777.92	<.0001
Error	1331	676999	508.63903		
Corrected Total	1337	3051092			

Root MSE	22.55303	R-Square	0.7781
Dependent Mean	104.83361	Adj R-Sq	0.7771
Coeff Var	21.51317		

Figure 3.6 ANOVA Table for Log and Sqrt

Trying out Log and Square Root Transformations to see if there is an improvement. In the ANOVA Table both have an acceptable F-test value and a similar R-Sq value which are greater than the original non-transformed.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	6.90997	0.07137	96.81	<.0001	0
age	1	0.03460	0.00087581	39.51	<.0001	1.01677
bmi	1	0.01273	0.00210	6.07	<.0001	1.09639
children	1	0.10088	0.01014	9.95	<.0001	1.00294
sex	1	0.07510	0.02451	3.06	0.0022	1.00888
smoker	1	1.55692	0.03040	51.22	<.0001	1.01130
southeast	1	-0.09070	0.02857	-3.17	0.0015	1.08659

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-3.00748	3.60749	-0.83	0.4046	0
age	1	1.39923	0.04427	31.61	<.0001	1.01677
bmi	1	1.00314	0.10591	9.47	<.0001	1.09639
children	1	3.23598	0.51240	6.32	<.0001	1.00294
sex	1	1.89981	1.23865	1.53	0.1253	1.00888
smoker	1	90.97341	1.53648	59.21	<.0001	1.01130
southeast	1	-3.27522	1.44423	-2.27	0.0235	1.08659

Collinearity Diagnostics										
			Proportion of Variation							
Number	Eigenvalue	Condition Index	Intercept	age	bmi	children	sex	smoker	southeast
1	4.53133	1.00000	0.00134	0.00480	0.00157	0.01460	0.01398	0.01075	0.01234
2	0.79225	2.39156	0.00025630	0.00166	0.00021469	0.02008	0.06278	0.78790	0.04576
3	0.69337	2.55640	0.00011772	0.00070403	0.00000159	0.05722	0.00595	0.12365	0.77577
4	0.53280	2.91628	0.00008757	0.00026282	0.00002975	0.58712	0.37287	0.02045	0.01682
5	0.35681	3.56363	0.00768	0.05221	0.01034	0.31159	0.51569	0.04547	0.07807
6	0.07615	7.71402	0.05045	0.90528	0.11214	0.00481	0.00732	0.00584	0.03119
7	0.01728	16.19315	0.94007	0.03508	0.87570	0.00458	0.02143	0.00593	0.04006

Collinearity Diagnostics									
			Proportion of Variation						
Number	Eigenvalue	Condition Index	Intercept	age	bmi	children	sex	smoker	southeast
1	4.53133	1.00000	0.00134	0.00480	0.00157	0.01460	0.01398	0.01075	0.01234
2	0.79225	2.39156	0.00025630	0.00166	0.00021469	0.02008	0.06278	0.78790	0.04576
3	0.69337	2.55640	0.00011772	0.00070403	0.00000159	0.05722	0.00595	0.12365	0.77577
4	0.53280	2.91628	0.00008757	0.00026282	0.00002975	0.58712	0.37287	0.02045	0.01682
5	0.35681	3.56363	0.00768	0.05221	0.01034	0.31159	0.51569	0.04547	0.07807
6	0.07615	7.71402	0.05045	0.90528	0.11214	0.00481	0.00732	0.00584	0.03119
7	0.01728	16.19315	0.94007	0.03508	0.87570	0.00458	0.02143	0.00593	0.04006

Figure 3.7 Multicollinearity Tests for Log and Sqrt

Looking at the Variance Inflation and Condition Index there is still no multicollinearity to be aware of.

Durbin-Watson D	2.054
Pr < DW	0.8366
Pr > DW	0.1634
Number of Observations	1338
1st Order Autocorrelation	-0.028

Durbin-Watson D	2.093
Pr < DW	0.9559
Pr > DW	0.0441
Number of Observations	1338
1st Order Autocorrelation	-0.048

Figure 3.8 Durbin-Watson test for Log and Sqrt

Unfortunately it would appear that the promising Sqrt transformation is failing to pass the Durbin-Watson test. The Sqrt transformation shows significant likelihood of autoregressive effect. The Log transformation on the other hand is showing significant improvement towards removing all evidence of any autoregressive effect.

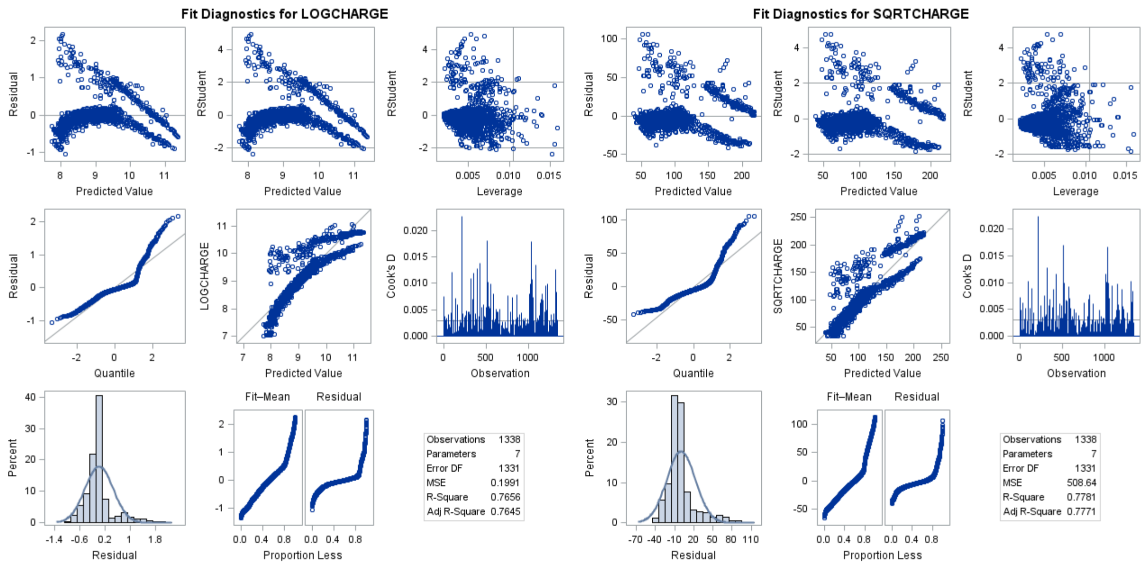

Figure 3.9 Log and Sqrt Model Diagnostics

Taking a look at the model diagnostics, the log transformation appears to be showing a significantly decreasing variance in the residuals. This appears like it could be a violation against heteroscedasticity. The Sqrt transformation, on the other hand, appears to improve significantly towards being random in its distribution within the band. However, like the Log, a possible violation is visible.

Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.837552	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.216523	Pr > D	<0.0100
Cramer-von Mises	W-Sq	14.42641	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	74.19143	Pr > A-Sq	<0.0050

Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.825912	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.197301	Pr > D	<0.0100
Cramer-von Mises	W-Sq	14.79341	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	77.87933	Pr > A-Sq	<0.0050

```
D'AGOSTINO TEST OF NORMALITY FOR VARIABLE D, N=1338
G1=1.68038   SQRTB1=1.67849   Z=18.05989   P=0.0000
G2=4.38517   B2=7.36432       Z=11.77282   P=0.0000
K**2=CHISQ(2 DF)=464.7590                  P=0.0000
```

```
D'AGOSTINO TEST OF NORMALITY FOR VARIABLE D, N=1338
G1=1.79911   SQRTB1=1.79709   Z=18.82257   P=0.0000
G2=4.07308   B2=7.05340       Z=11.41615   P=0.0000
K**2=CHISQ(2 DF)=484.6177                  P=0.0000
```

Figure 3.10 Normality tests for Log and Sqrt

It would appear that neither model significantly made any sort of improvement toward normality, so I will ultimately choose the log as the best transformation and move on.

4. Model Selection

Model Index	Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	BIC	SBC	Variables in Model
1	6	0.7645	0.7656	7.0000	-2152.4698	-2150.3962	-2116.07729	age bmi children sex smoker southeast
2	5	0.7630	0.7639	14.3898	-2145.0637	-2143.0851	-2113.87016	age bmi children smoker southeast
3	5	0.7629	0.7638	15.0765	-2144.3785	-2142.4060	-2113.18492	age bmi children sex smoker
4	4	0.7614	0.7622	22.4447	-2137.0642	-2135.1566	-2111.06958	age bmi children smoker
5	5	0.7582	0.7591	41.8849	-2117.8954	-2116.1585	-2086.70180	age children sex smoker southeast
6	4	0.7579	0.7586	42.3662	-2117.4705	-2115.7078	-2091.47582	age children sex smoker
7	4	0.7570	0.7577	47.6264	-2112.3444	-2110.6194	-2086.34977	age children smoker southeast
8	3	0.7567	0.7573	48.2160	-2111.8280	-2110.0631	-2091.03224	age children smoker
9	5	0.7472	0.7481	104.0122	-2058.4651	-2057.2455	-2027.27149	age bmi sex smoker southeast
10	4	0.7459	0.7467	110.3936	-2052.6459	-2051.3536	-2026.65121	age bmi smoker southeast
11	4	0.7453	0.7461	113.8369	-2049.4467	-2048.1772	-2023.45204	age bmi sex smoker
12	3	0.7440	0.7446	120.1872	-2043.7201	-2042.3509	-2022.92435	age bmi smoker

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	smoker	1	0.4429	0.4429	1829.15	1062.12	<.0001
2	age	2	0.2966	0.7395	146.822	1520.53	<.0001
3	children	3	0.0177	0.7573	48.2160	97.38	<.0001
4	bmi	4	0.0049	0.7622	22.4447	27.41	<.0001
5	southeast	5	0.0018	0.7639	14.3898	9.99	0.0016
6	sex	6	0.0017	0.7656	7.0000	9.39	0.0022

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	smoker		1	0.4429	0.4429	1829.15	1062.12	<.0001
2	age		2	0.2966	0.7395	146.822	1520.53	<.0001
3	children		3	0.0177	0.7573	48.2160	97.38	<.0001
4	bmi		4	0.0049	0.7622	22.4447	27.41	<.0001
5	southeast		5	0.0018	0.7639	14.3898	9.99	0.0016
6	sex		6	0.0017	0.7656	7.0000	9.39	0.0022

Elastic Net Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	CV PRESS
0	Intercept		1	1131.2143
1	smoker		2	631.2933
2	age		3	296.1230
3	children		4	276.4721
4	bmi		5	271.0734
5	sex		6	269.5284
6	southeast		7	268.0964*

* Optimal Value of Criterion

LASSO Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	CV PRESS
0	Intercept		1	1132.4652
1	smoker		2	631.7697
2	age		3	296.3745
3	children		4	276.7548
4	bmi		5	271.4101
5	sex		6	270.0802
6	southeast		7	268.7409*

* Optimal Value of Criterion

Figure 4.1 Model Selection Summaries: Adj-R-Sq, R-Sq, C(p), AIC, BIC, SBC (top left); Forward Selection (top right); Stepwise Selection (bottom left); Elastic Net (bottom center); LASSO (bottom right)

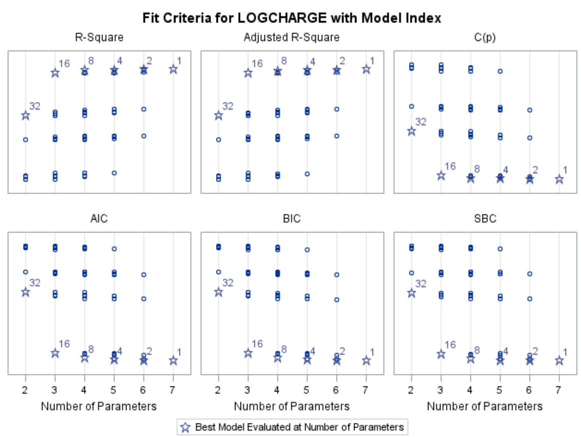

Figure 4.2 Fit Criteria Charts

Ultimately every model selection type selected Model 1 as the most accurate model. This means that Smoker, Age, Children, BMI, Southeast Region, and Sex were the most influential aspects and what should be used in the final model.

Since I have already displayed the final model earlier on, I will recap the final parameters described in Figure 3.7:

Log(Charges) = 6.91 + 0.035(age) + 0.013(bmi) + 0.1(children) + 0.75(sex) + 1.56(smoker) - 0.09(southeast)

Or:

Charges = exp(6.91 + 0.035(age) + 0.013(bmi) + 0.1(children) + 0.75(sex) + 1.56(smoker) - 0.09(southeast))

Conclusion

In conclusion, the model above seems like the most accurate that could be made through the transformations that I have made to the data. Unfortunately, the normality violation still remains and no amount of my own experimentation with transformations seemed to relieve it. Perhaps with further transformations, it could be done, but in the end it looks like non-parametric tests are what would be required. Despite all of that, I still believe that the final model is a very good resource for understanding what aspects most contribute towards health insurance charges. Especially in our current environment where the fear of getting sick is a huge influence on society. To summarize, smoking is the largest contributor to charges followed by age and bmi. In addition to those, children, sex, and being from the southeast seem to have a minor impact on price. All together they form the above log transformed linear equation.

References:

1. Srikanth, Anagha. "How Much Will Getting Coronavirus Cost You?" *TheHill*, 3 Mar. 2020, thehill.com/changing-america/respect/poverty/485666-how-much-will-getting-coronavirus-cost-you.
2. Stedy. "Stedy/Machine-Learning-with-R-Datasets." *GitHub*, 28 Mar. 2017, github.com/stedy/Machine-Learning-with-R-datasets.
3. Prianda, Bayu Galih. "Prediction of Health Insurance Costs with Linear Regression." *Medium*, Medium, 24 Dec. 2018, medium.com/@BAYUGALIH/prediction-of-health-insurance-costs-with-linear-regression-8fd95a905a40.