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 + 259.547x1 + 322.615x2 + 23823.684x3, where (y = charges, x1 = age, x2 = bmi, x3 = 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

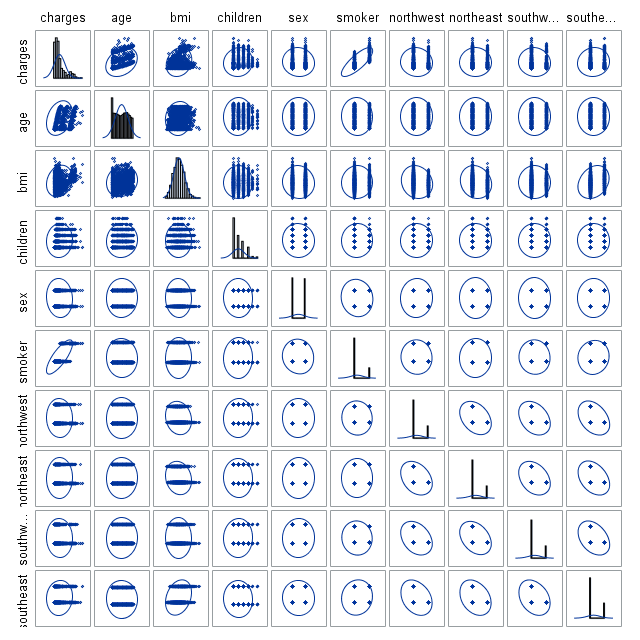


Figure 1.1 Scatter matrix for Charges, age, bmi, children, sex (Female = 1), smoker (yes = 1), and each region (in region = 1) with histogram on diagonal

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.

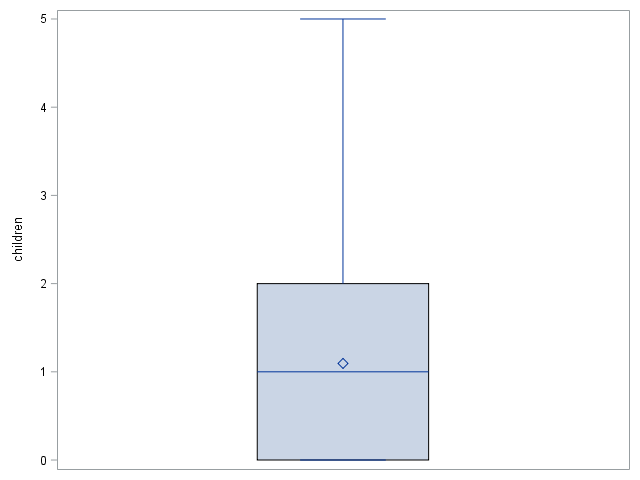
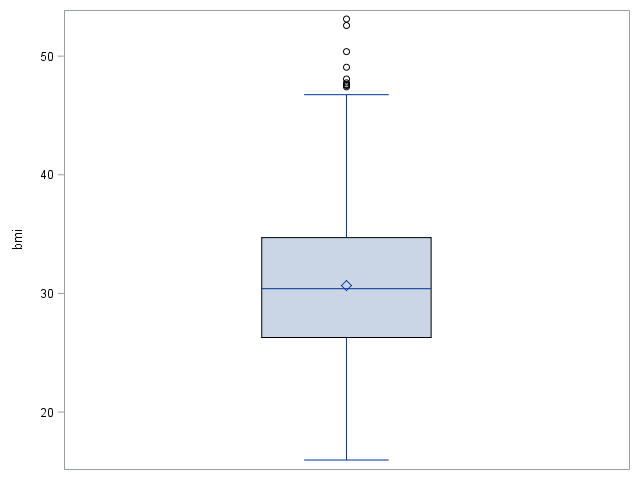
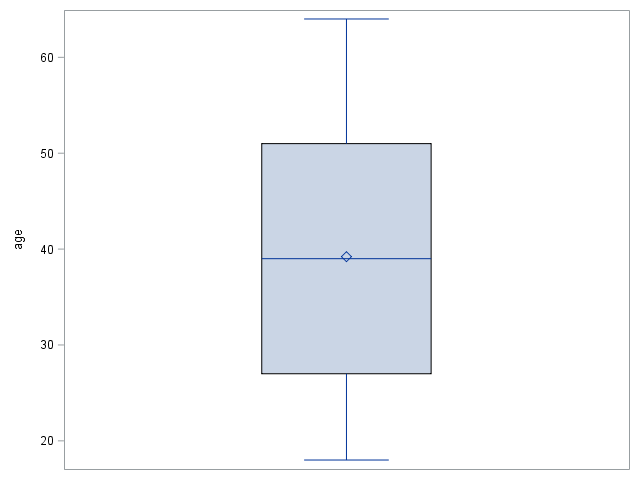
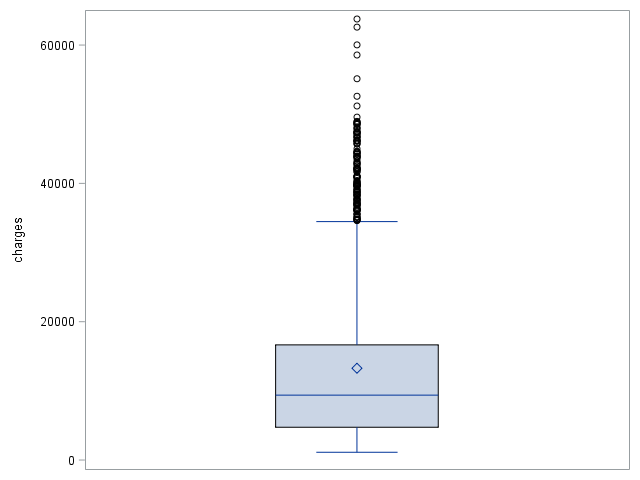


Figure 1.2 Boxplot for Charges (top left), Age (top right), BMI (bottom left), Children (bottom right)

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.

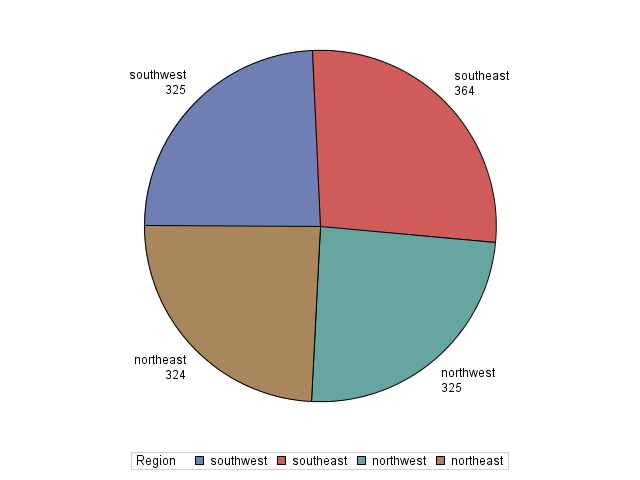
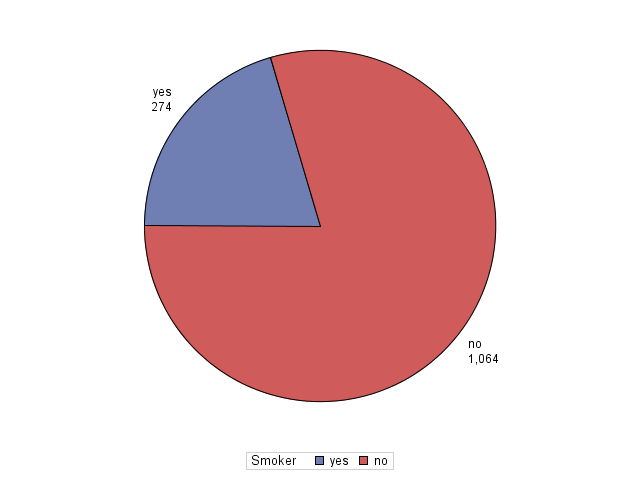
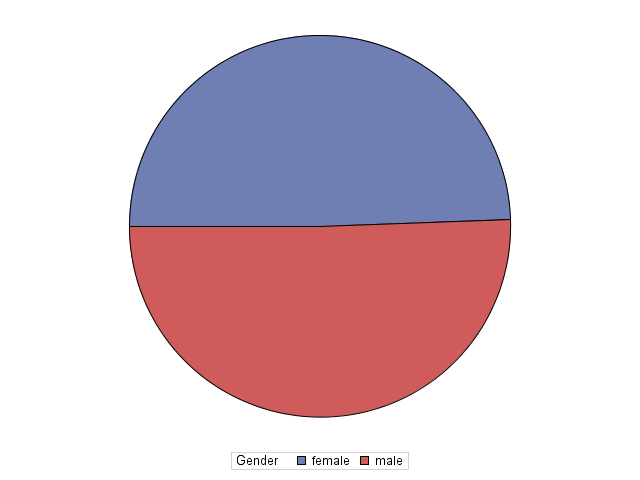


Figure 1.3 Pie Charts depicting the categorical features: Gender (left), Smoker (center), Region (right)

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.

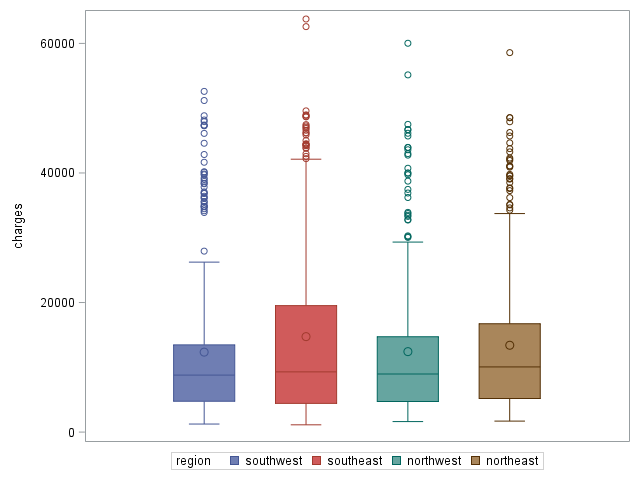
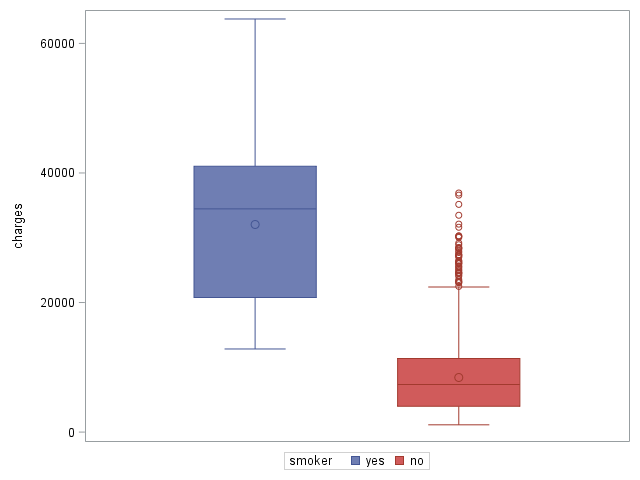
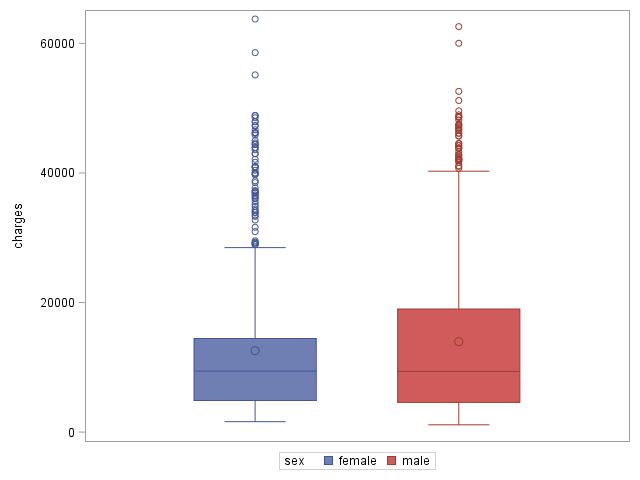


Figure 1.4 Boxplots to compare categorical features: Gender (left), Smoker (center), Region (right)

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

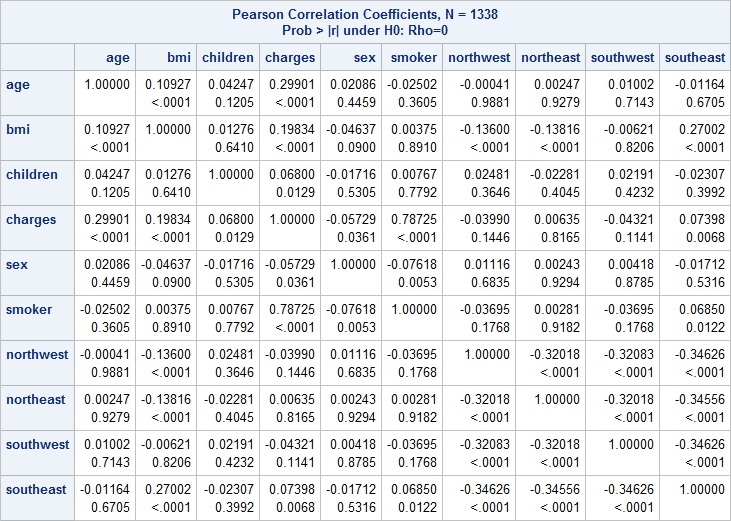


Figure 2.1 Correlation Analysis Data for Charges, Age, BMI, Children, Sex (Female = 1), Smoker (Yes = 1), and each individual Region (In the region = 1)

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

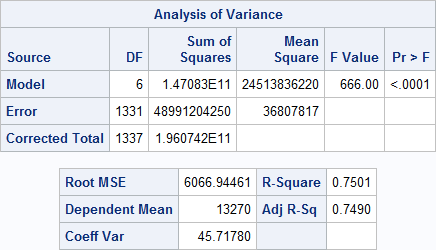


Figure 3.1 ANOVA table for the first model

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.

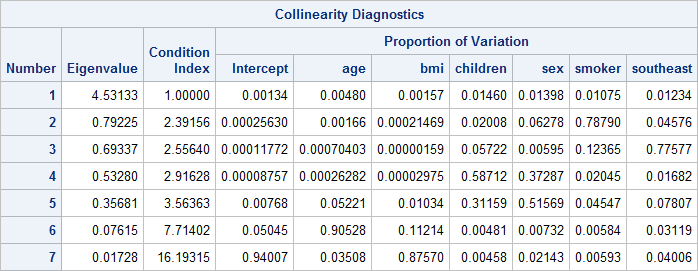
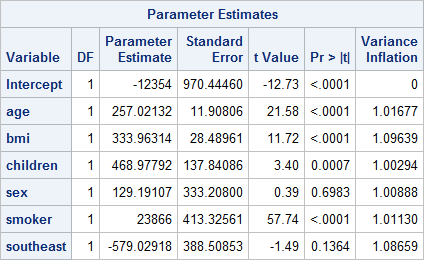


Figure 3.2 Parameter Estimates and Collinearity Diagnostics

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.



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.

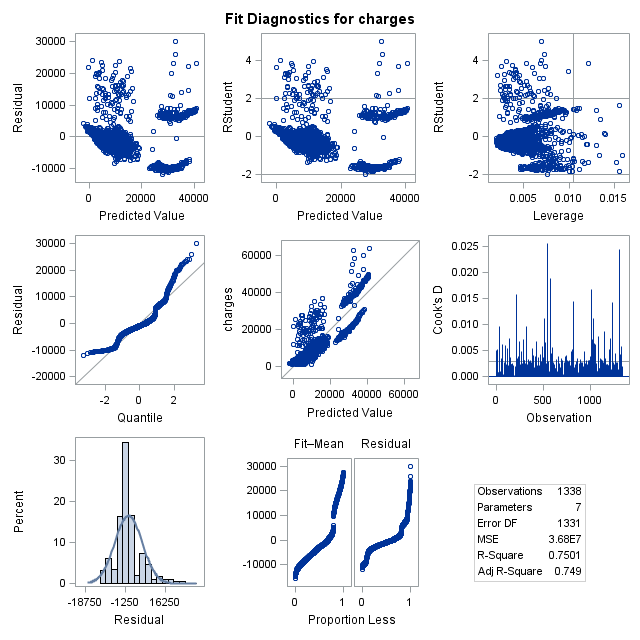


Figure 3.4 Diagnostics Panel for Model

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.

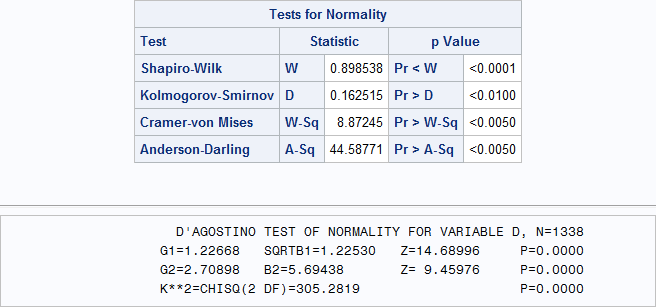


Figure 3.5 Normality Test for the full model

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.

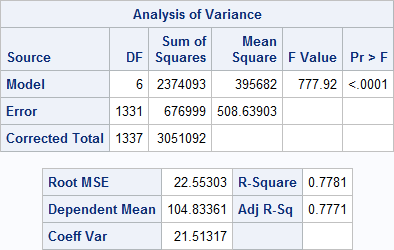
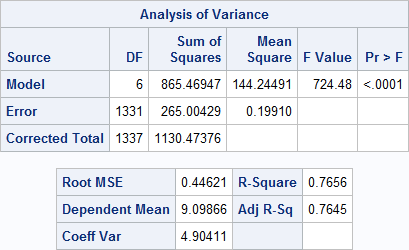


Figure 3.6 ANOVA Table for Log and Sqrt

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.

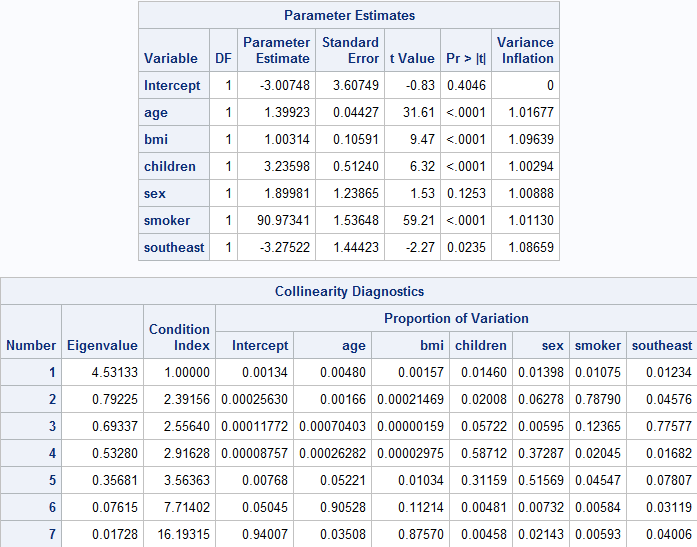
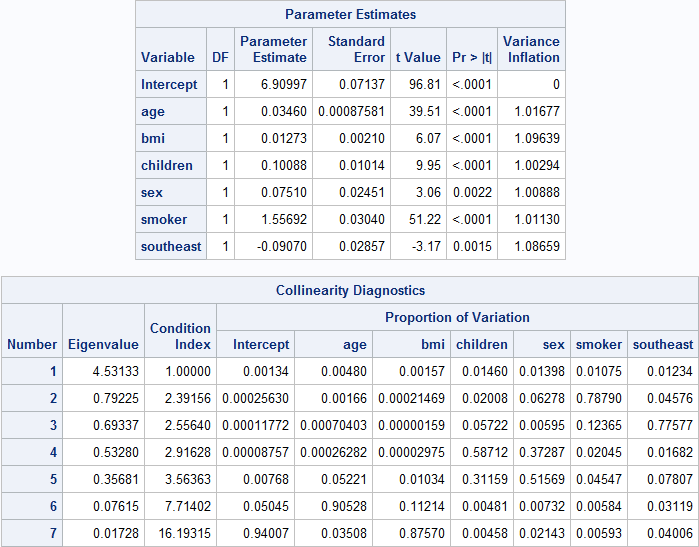


Figure 3.7 Multicollinearity Tests for Log and Sqrt

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

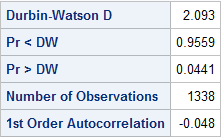


Figure 3.8 Durbin-Watson test for Log and Sqrt

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.

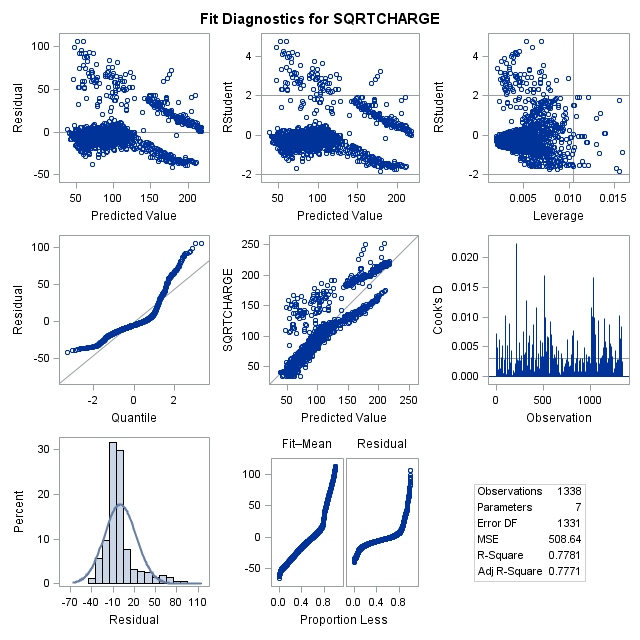
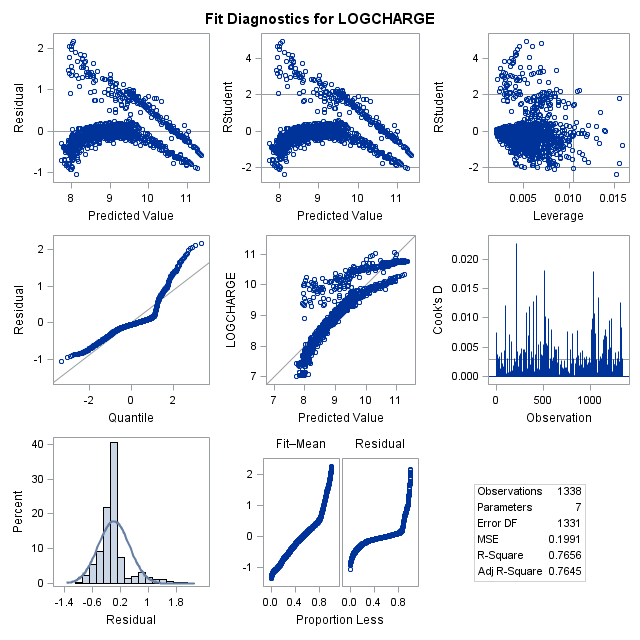


Figure 3.9 Log and Sqrt Model Diagnostics

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.

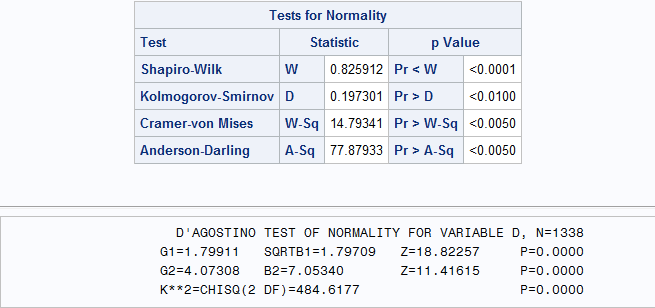
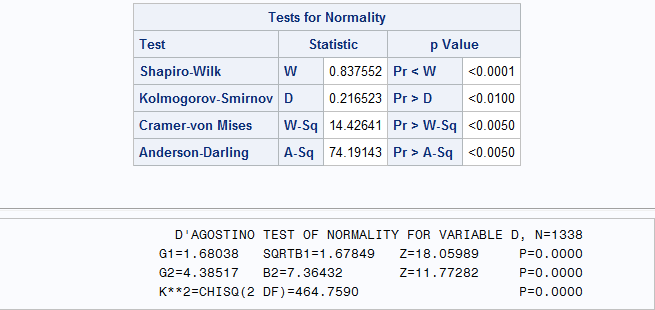


Figure 3.10 Normality tests for Log and Sqrt

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

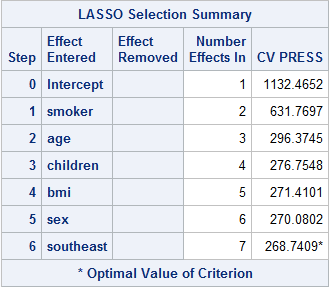
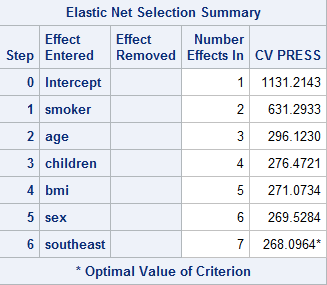
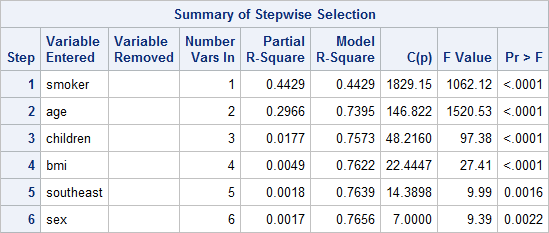
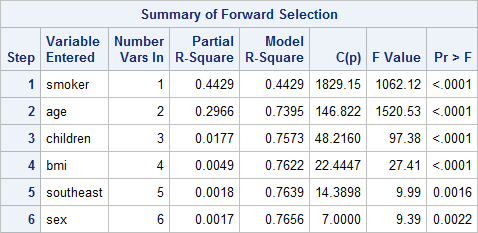
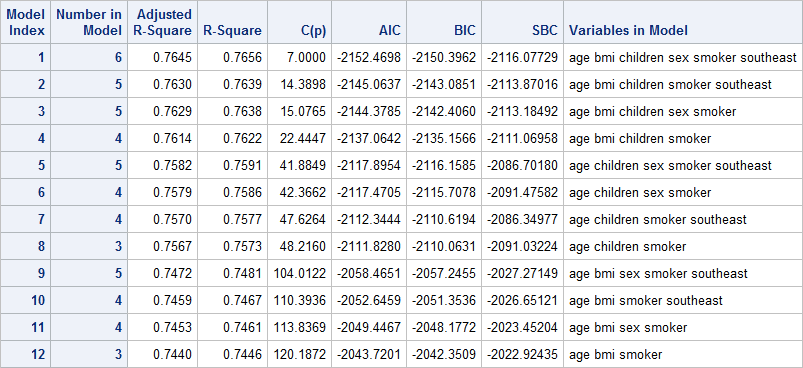


Figure 4.1 Model Selection Summaries: Adj-R-Sq, R-Sq, C(p), AIC, BIC, SBC (top left); Forward Selection (top right); Stepwise Selection (bottom left); Elastic Net (bottom center); LASSO (bottom right)

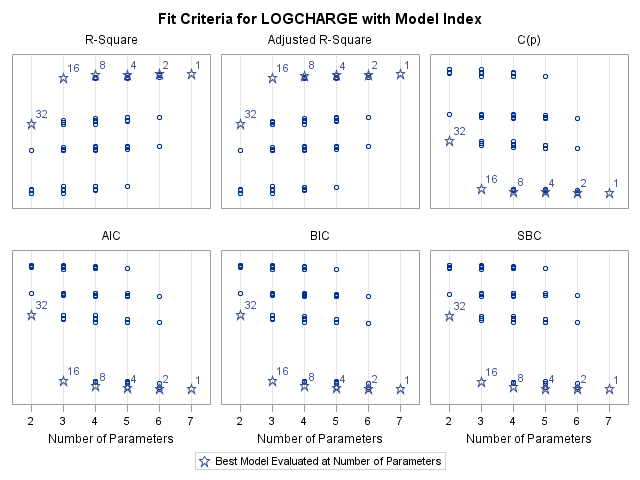


Figure 4.2 Fit Criteria Charts

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](http://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](http://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](http://medium.com/@BAYUGALIH/prediction-of-health-insurance-costs-with-linear-regression-8fd95a905a40).