Insurance Forecasting

STAT 350 Project

Harsha Gamage

Aren Zita (301291488)
Chamodi Basnayake (301309667)
Weibin Huang (301323335)

Introduction

This project aims to analyze a real life data set and use statistical methods to create a predictive model with given data. A data set from Kaggle on insurance forecasting was chosen, which provided certain patient information and their respective insurance costs. R programming language was used to clean, organize and analyze the data to create an accurate model to predict insurance costs based on patient information.

Data Set

The data set consisted of 1338 data points, with one response variable (charges) and six predictive variables (age, sex, bmi, children, smoker and region). The variables were a mix of numerical and categorical. A summary of the variables can be shown in the table below.

Variable	Type	Range
Charges	Numeric	\$1.12k - \$63.8k
Age	Numeric	18 - 64
Sex	Categorical	Female, Male
BMI	Numeric	16 - 50
Children	Numeric	0 - 5
Smoker	Categorical	True, False
Region	Categorical	SouthEast, SouthWest, NorthEast, NorthWest

Table 1: Summary of variables in the data set

To create an accurate predictive model, firstly some descriptive statistics were drawn to better understand the dataset. Secondly various variable selection methods such as forward, backward and stepwise regression along with LASSO regression was performed. Interaction terms were introduced and Box Cox variable transformation was used to transform the model to better fit the data. Finally cross validation was used to pick the best predictive model. R^2 adj and root mean square error were used as measures to select the best predictive model.

Descriptive Statistics

Before diving into model building, some summary data was drawn to better understand the data at hand. While data on ‘Charges’ is distributed over a large range, as shown in Figure 1 (a), the majority of the patients pay a smaller insurance charge, with a few people paying up to \$60k. Looking into the impact that number of dependence a patient has and it’s correlation to insurance cost, it shows on average patients with 2 or 3 dependents have the highest insurance charges (Figure 1 (b)).

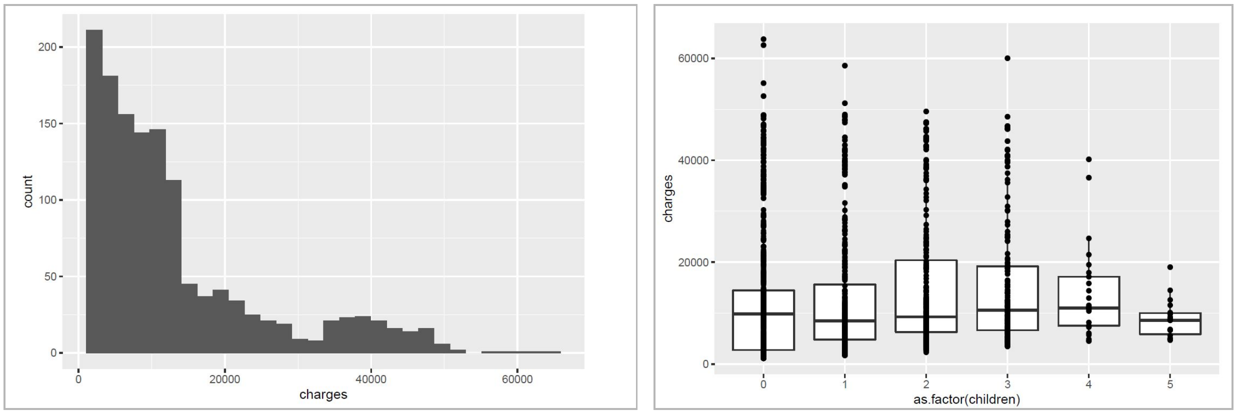


Figure 1: Distribution of charges among patients (a) and number of children vs. charges (b)

A significant variable that can contribute to higher insurance costs is the patient's health status. Variables such as smoker and BMI in the dataset are good indicators of these measures. As shown in Figure 2 (a) and (b), some of our initial assumptions are confirmed. Patients that are smokers on average have a higher insurance charge than those who do not, and secondly as the BMI of a patient increases, so does their insurance charges. Having noted these observations, this could be helpful in making decisions in model building moving forward.

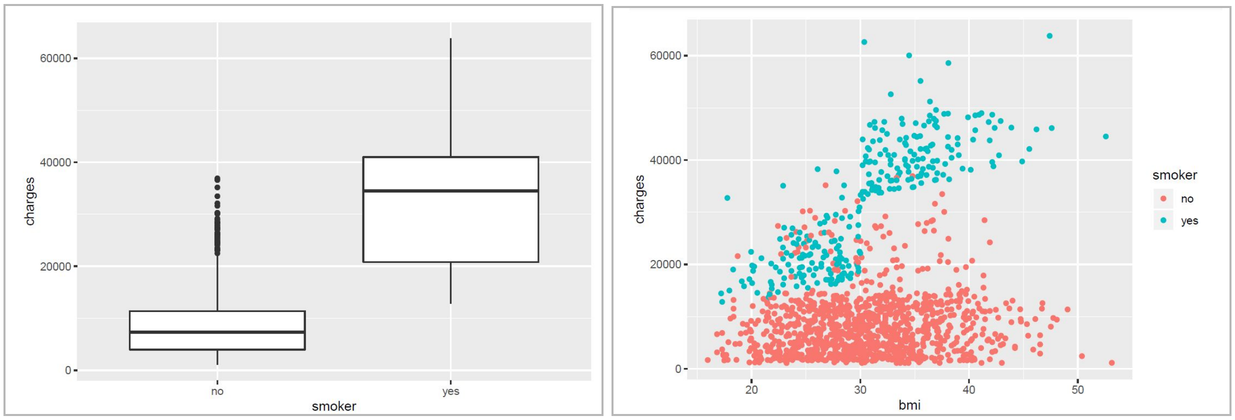


Figure 2: Charges among smokers and non-smokers (a) and charges vs. bmi (b)

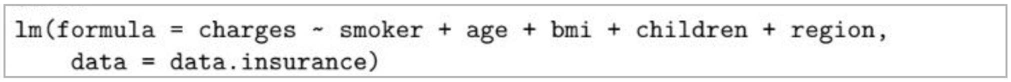
Variable selection

As a starting point of model building, variable selection was implemented to filter out any insignificant variables from the model using forward, backward and stepwise selection methods, along with LASSO regression.

Forward, Backward and Stepwise Regression

These variable selection methods belong to a family of methods which add or remove variables from a model sequentially. In R AIC (Akaike Information Criterion), an estimator of the out-of-sample prediction error, is used to determine the relative quality of the statistical model. Forward stepwise regression starts with the intercept and iteratively add the most contributive predictors until the improvement is no longer statistically significant, signified by the lowest AIC value. With a similar approach, backward stepwise regression starts with the full model and iterative removes predictors until the improvement is no longer significant. The mixed stepwise regression uses a mixed approach where it starts with no predictors then sequentially add the most contributive predictors and then remove any variables that no longer provide any improvement to the model.

All the methods output the same model, with all predictors except ‘Sex’ included in final model as shown in Figure 3. Appendix A, shows the full results of each of these models.



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

Figure 3: Final model with forward, backward and mixed stepwise regression

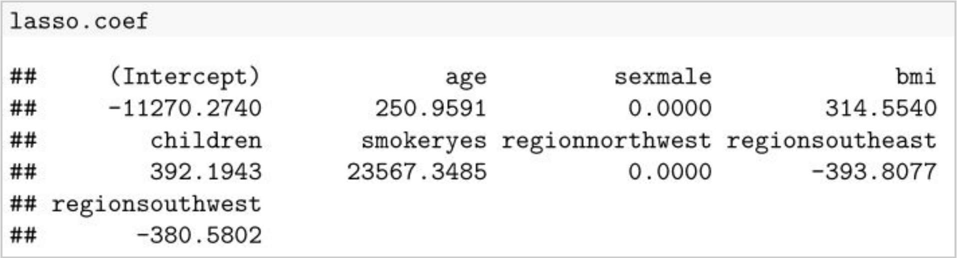
Fitting a multiple linear regression model, the R² adjusted value is at a 74.9% and p-value of 2.2e-16. This model is significant and explains the data fairly well. However the residual standard error (RSE) is 6060, which is the square root of mean square error, a measure of average of squares of errors. This model has large error values, and can definitely be improved further.

LASSO Regression

LASSO, stands for Least Absolute Shrinkage and Selection Operator, is a modern way to select and deselect variables. It deselect some of the variables by shrinking them towards zero. In order to perform LASSO regression, the criterion of LASSO regression needed to be minimized.

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

In this criterion, RSS stands for residual sum of squares and $\lambda$is the shrinkage parameter. In order to choose the best λ, the original data is separated into two groups - training data and testing data. After separating the original data set, cross-validation is used to determine the best λ. Also, as λ increases, several β will be shrunk to zero which means those coefficients are set to zero. As a result, the minimizing criterion contributes to the process of deselecting coefficients. Figure 4 shows the result of LASSO regression



```
lasso.coef

##     (Intercept)             age         sexmale             bmi
##     -11270.2740        250.9591          0.0000        314.5540
##        children       smokeryes regionnorthwest regionsoutheast
##        392.1943      23567.3485          0.0000       -393.8077
## regionsouthwest
##       -380.5802
```

Figure 4: LASSO coefficients of corresponding variables

This figure shows that the predictor ‘Sex’ is excluded from model and it matches the result of stepwise variable selection shown above. In addition, one of the category - region northwest is also considered as insignificant in this model.

Model Adequacy Checking

Having chosen the significant variables in the model, now model adequacy can be checked with the use of residual analysis, residual and normality plots, detection and treatment of multicollinearity. The major assumptions made in this regression analysis is that the relationship between response and predictor variables is at least approximately linear, the error term has zero mean and constant variance, the errors are uncorrelated and normally distributed.

Residual Analysis

Residuals were analyzed using residual and normality plots to check that errors are uncorrelated and normally distributed. The plots are shown in Figure 5 (a) and (b). As observed in (a) the residuals plotted against fitted values, the residuals are not uncorrelated or independent since there’s clustering in the plot rather than randomness. It implies that variance is not constant, and also that there may be some potential outliers in the data. In plot (b), it’s apparent that the normality assumption is being violated since there are very obvious skews in the normality plot.

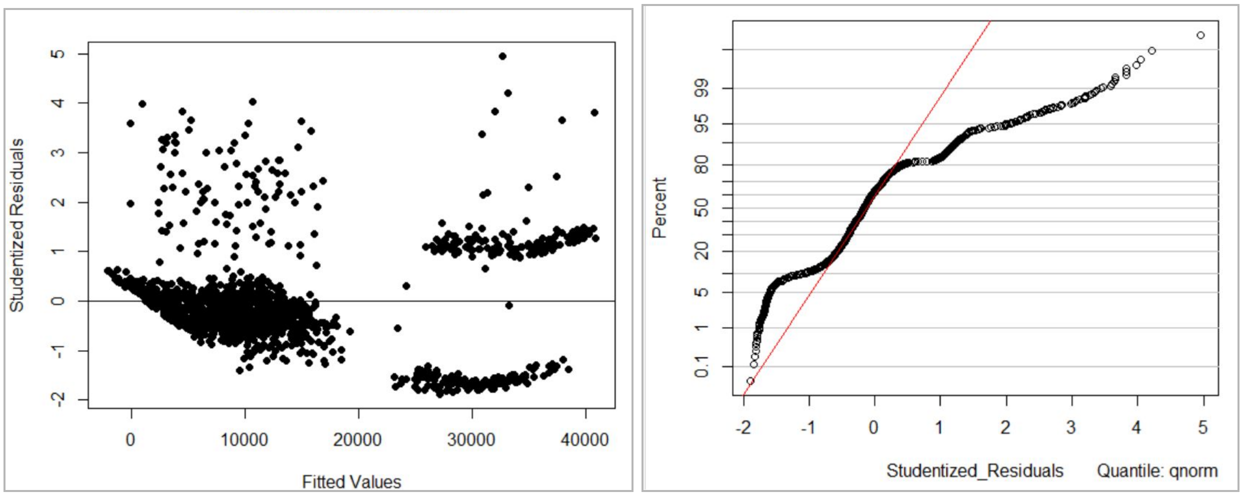
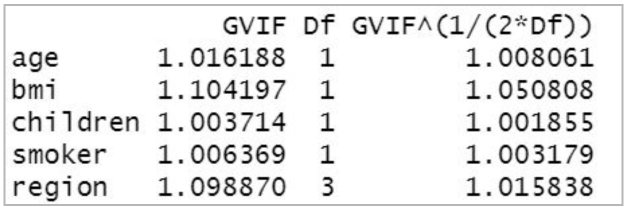


Figure 5: Residuals against fitted values (a) and normality plot of residuals (b)

Multicollinearity

Multicollinearity is the presence of high intercorrelations among the independent variables in a multiple regression model. The presence of it can erratically change the coefficient estimates with small changes in the model or data, often observed with high standard errors. This can impact decisions in model building, and should be treated moving forward. VIF (variance inflation factor) is used as a criterion for collinearity measure, calculated as the inverse of ‘tolerance’ which is (1 - R² Adj). A VIF measure greater than 5 indicates a multicollinearity problem. Figure 6 below shows the output of predictor VIF values. There is no collinearity present, and therefore no adjustments were required.



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

Figure 6: VIF values for predictor variables to detect multicollinearity existence

Interaction Terms

The initial model is able to explain 75% of the data provided. In order to further improve the model, interaction terms were introduced. Interaction terms exist when the effect of an independent variable on a dependent variable changes depending on the values of another independent variable. Interaction plots often reveal the presence or absence of interactions among independent variables. These plots are line graphs with the dependent variable on the y axis and independent variables on x axis. If the slopes of each of the regressors are parallel, then there is no interaction. For the purpose of simplifying the process, pair combinations of the most significant variables in our model (age, BMI, smoker, children) were tested on the model.

The significance of each of the models are shown in Appendix B. One interaction was deemed to be very significant in the model, Smoker::BMI. The interaction plot is shown in Figure 7.

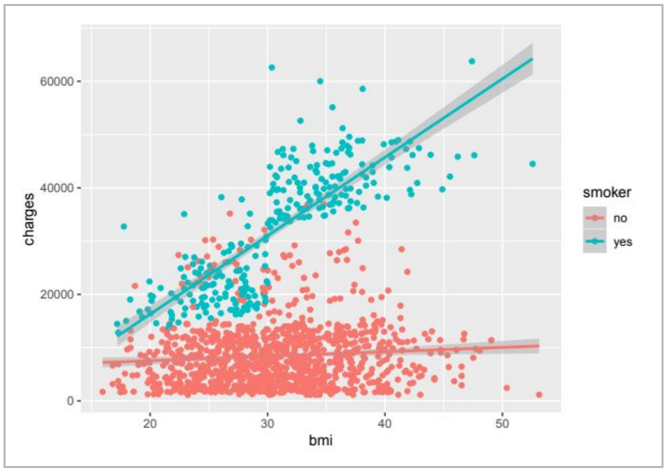


Figure 7: Interaction plot of BMI::Smoker

Due to its high significance in the model, the interaction term was added to the model. This improves the R² Adj to 84%, which is a vast improvement from the 75% previous score. This also decreases the residual standard error to 4851, as opposed to the previous 6060. The addition of the interaction term has improved the model based on both these measures.

<u>Variable transformations</u>

Despite the improvement in the model's fit to the data and the decrease in residual standard error measures, a re-analysis of the residual and normality plots indicate that this model continues to violate the normality assumption. The results are shown in Appendix C. While we have reduced residuals for fitted values, they are not distributed normally and the normality plot is very skewed.

The non-normality nature of the data can potentially be fixed by transforming the dependent variable. A Box Cox transformation is implemented to help meet the normality assumption. At the core, Box Cox has a parameter λ which varies from -5 to 5. The optimal value is which results in the best approximation of a normal distribution curve. This value is calculated using R. The details are shown in Appendix C. The lambda value was found to be λ = 0.262623.

Based on the first equation shown below, the dependent variable is transformed.

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

The residual and normality plots were redrawn, as shown below Figure 8 (a) and (b). The transformation has made the residuals more normally distributed than before, however it still severely violates the normality assumption based on the normality plot. This implies that a more complicated transformation than Box Cox is required to make this model reach linear regression assumptions.

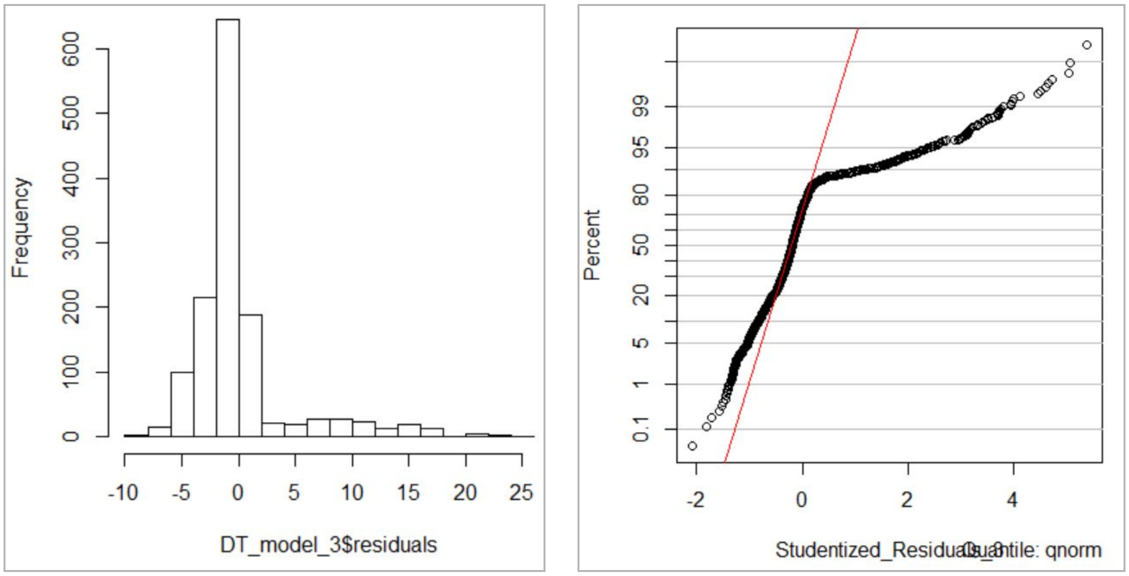


Figure 8: Histogram of residuals (a) and normality plot of residuals (b)

Cross Validation

K-fold cross-validation is a resampling procedure used to estimate the skill of a model, allowing us to compare and select between various models. It is a method in machine learning to check the effectiveness of the method. It generally results in a less biased estimate of the model skill since the model is trained and tested iteratively. It follows the following procedure:

1. Split the whole data randomly to k partitions.
2. Use one partition as test data and other remainings as training data. Fit the model using the training data then validate it by calculating an error using the test data.
3. Repeat it k times picking different partition as a test data each time. Calculate the root mean squared error.

For this project a 10-fold cross-validation was performed with no repeats.

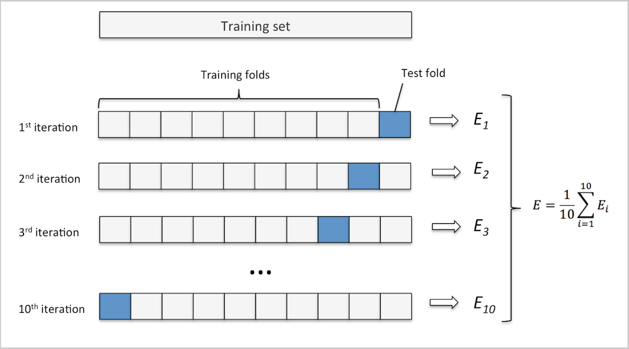
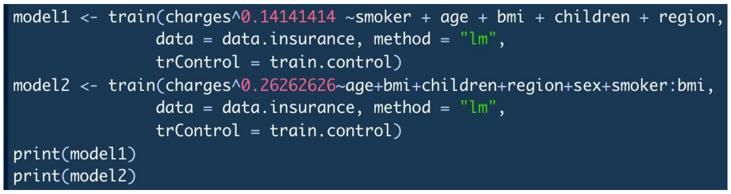


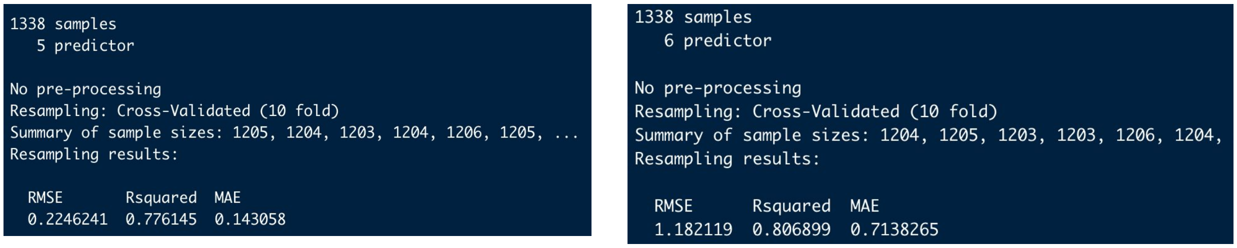
Figure 9: K-fold cross-validation in diagram

Evaluating Final Models



```
model1 <- train(charges^0.14141414 ~smoker + age + bmi + children + region,
                data = data.insurance, method = "lm",
                trControl = train.control)
model2 <- train(charges^0.26262626~age+bmi+children+region+sex+smoker:bmi,
                data = data.insurance, method = "lm",
                trControl = train.control)
print(model1)
print(model2)
```

Figure 10:The models with 10-fold cross-validation in R



```
1338 samples
   5 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1205, 1204, 1203, 1204, 1206, 1205, ...
Resampling results:

  RMSE       Rsquared  MAE
  0.2246241  0.776145  0.143058
```

```
1338 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1204, 1205, 1203, 1203, 1206, 1204,
Resampling results:

  RMSE      Rsquared  MAE
  1.182119  0.806899  0.7138265
```

Figure 11:Outputs of the model without the interaction (a) and with the interaction (b)

The two models were evaluated with the 10-fold cross validation method: the Box Cox transformed model with interaction term Smoker::BMI and Box Cox transformed model without the interaction. The models were evaluated based on their R^2 Adj and root mean squared error (RMSE) measures. Figure 11 shows the outputs of the cross validation for the two models tested. The model without the interaction (Figure 11.a) gives RMSE = 0.2246 and R^2 Adj = 0.7761. The model with the Smoker::BMI interaction term gave RMSE = 1.1821 and R^2 Adj = 0.8069. The first model gives lower errors but smaller R^2 Adj which means it has low variance and less bias in the model. The second model (with interaction term) seems better since the R^2 Adj value got improved while the error stayed small. Thus, we pick the second model with interaction term as our final model.

$$charges = 20.464 + 0.346(age) + 0.034(bmi) + 0.910(children) - 2.410(smokeryes) - 0.717(regionNW) - 1.568(regionSE) - 1.395(regionSW) + 0.684(bmi :: smoker)$$

Regression Trees

Classification and Regression Tree Analysis (CART) is another way of approaching data. This analysis is effective when there are clusters in data and easier to interpret the significance of regressors by visualizing the data into a tree.

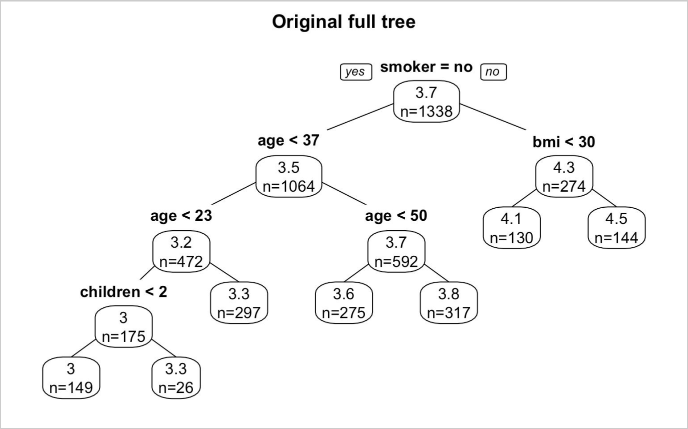


Figure 12: The original decision tree created in R

The model without the interaction term is used here since the analysis cannot handle it. By using R, a large tree is created with the least error, Figure 12. The initial tree has seven leaf nodes and indicates Smoker, BMI, age and children are the most significant variables. However since the error is minimized, the tree might overfit the data, thus, pruning must be applied to find the balance between bias and variance.

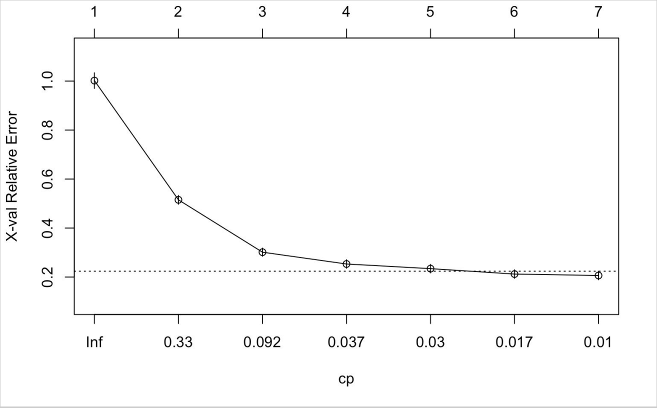


Figure 13: Relative error versus Complexity Parameter (Cp)

The figure shows the relative error versus complexity parameter (Cp). The Cp is used as a stopping criterion of the tree. The calculation shows the residual sum of squared (RSS) at each leaf plus the lambda value which is the parameter of Cp multiplied by S, the number of leaf nodes. This calculation penalizes the error for the addition of nodes to minimize the bias on the tree. The horizontal line in the Figure 13 represents the highest cross-validated error minus the minimum cross-validated error, plus the standard deviation of the error at that tree. It is reasonable to pick the Cp at the leftmost value under the line which is Cp = 0.017.

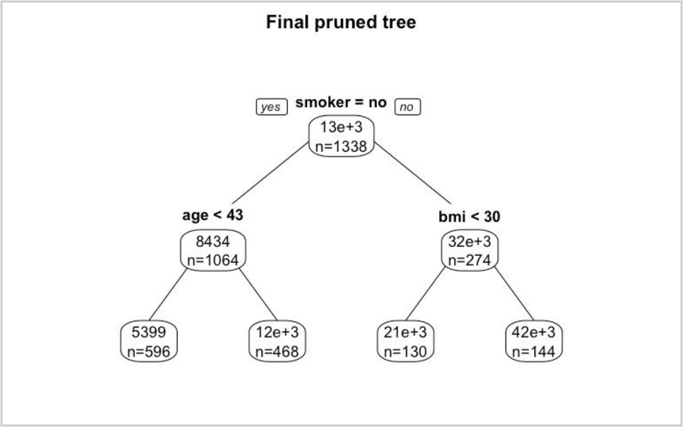


Figure 14: Pruned decision tree

After pruning, the final tree eliminated a terminal node depending on the ‘children’ variable. The tree shows that the smoker variable is the most significant to the charges. Also, the obesity measure of people influences the charges since the right-hand side of the tree splits the data at BMI 30 which is the boundary of obesity.

Conclusion

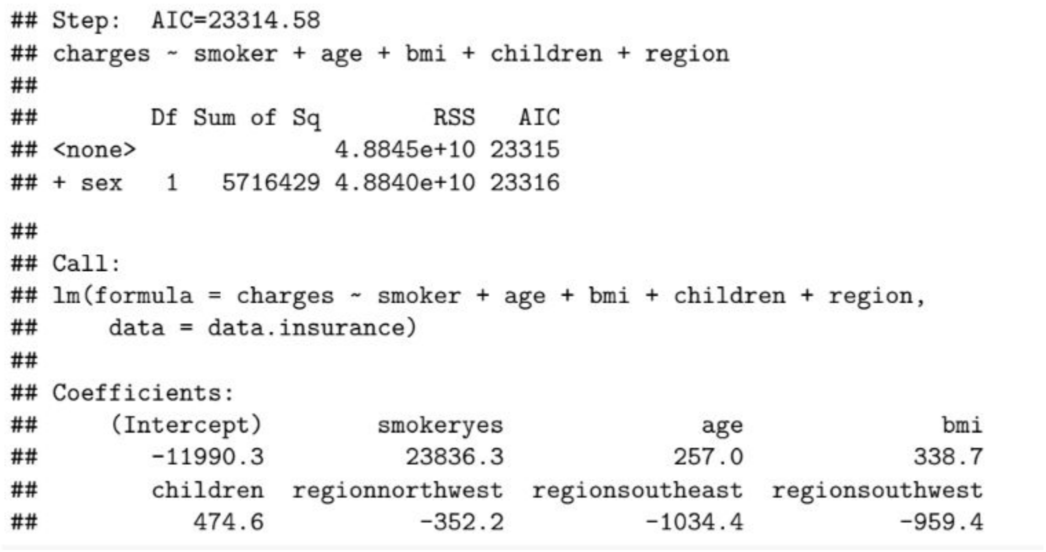
Through this project several regression analysis methods were utilized to come up with a final model that explains the data well while predicting insurance costs with low errors. Stepwise and LASSO regression was used for variable selection, and an interaction term was added to improve the predictive power of the model. In order to help meet normality assumption of linear regression, the response variable was transformed with the use of Box Cox transformation and finally our final models’ skill levels were tested using cross validation. Our final model had root mean square error of 1.1821 and R^2 Adj value of 80.69%. While we were able to greatly improve our model through these techniques, further processing and analysis can be performed to better it.

Some more techniques and methods might help to improve the accuracy of the model with low bias. Based on the normal Q-Q plot, despite the transformation, there is a tailed distribution violating the normality assumption of linear models, thus, more complex transformations may be needed to meet the model assumptions. Secondly, random forest is another way to approach model building. It fits a number of decision tree classifiers on various sub-samples of the data and uses averaging to improve the predictive accuracy while controlling over-fitting. Finally, support vector machine classifier could be used to create a predictive model since different clusters of data points are observed in the scatter plots. This method could help to classify those observations effectively.

References

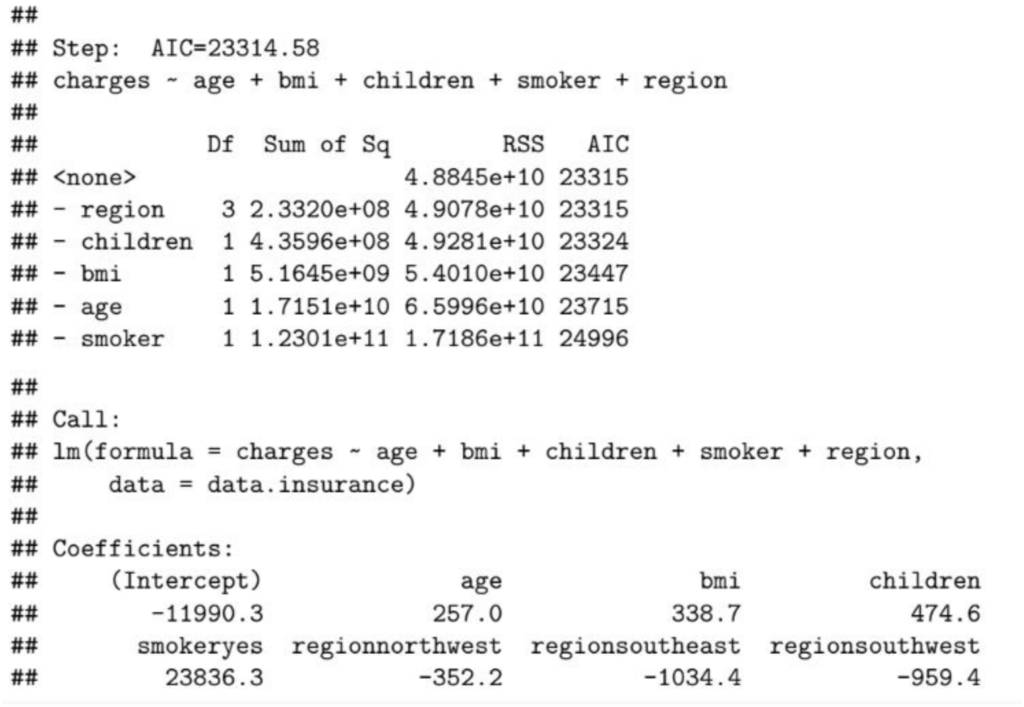
https://www.kaggle.com/mirichoi0218/insurance
http://karlrosaen.com/ml/learning-log/2016-06-20/
https://bookdown.org/max/FES/detecting-interaction-effects.html

Appendix A - Variable selection



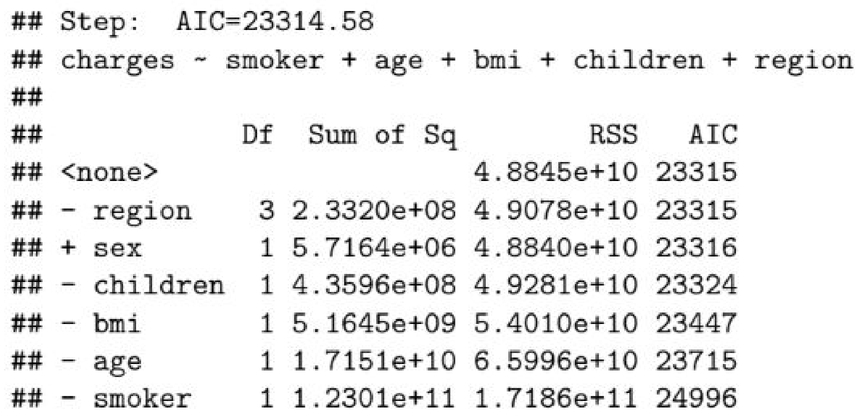
```
## Step:  AIC=23314.58
## charges ~ smoker + age + bmi + children + region
##
##        Df Sum of Sq        RSS   AIC
## <none>              4.8845e+10 23315
## + sex   1   5716429 4.8840e+10 23316

##
## Call:
## lm(formula = charges ~ smoker + age + bmi + children + region,
##     data = data.insurance)
##
## Coefficients:
##     (Intercept)        smokeryes              age              bmi
##        -11990.3          23836.3            257.0            338.7
##        children  regionnorthwest  regionsoutheast  regionsouthwest
##           474.6           -352.2          -1034.4           -959.4
```

Figure 1: Final model from forward stepwise regression



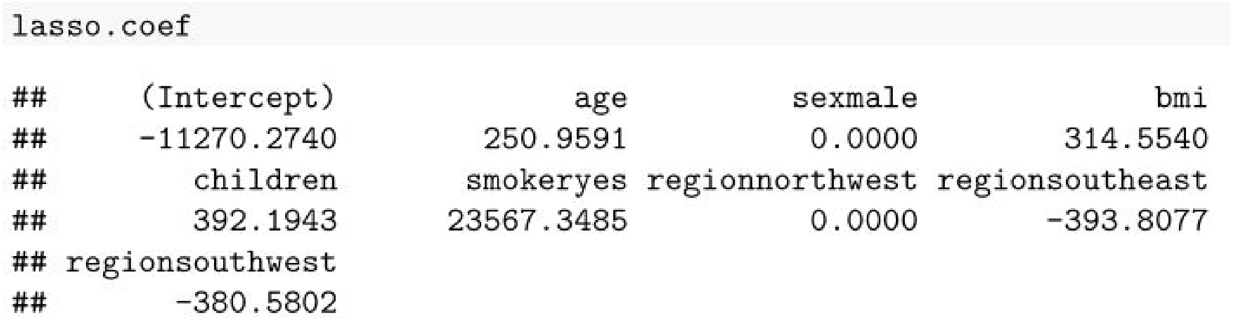
```
##
## Step:  AIC=23314.58
## charges ~ age + bmi + children + smoker + region
##
##            Df  Sum of Sq        RSS   AIC
## <none>                   4.8845e+10 23315
## - region    3 2.3320e+08 4.9078e+10 23315
## - children  1 4.3596e+08 4.9281e+10 23324
## - bmi       1 5.1645e+09 5.4010e+10 23447
## - age       1 1.7151e+10 6.5996e+10 23715
## - smoker    1 1.2301e+11 1.7186e+11 24996

##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + region,
##     data = data.insurance)
##
## Coefficients:
##     (Intercept)              age              bmi         children
##        -11990.3            257.0            338.7            474.6
##       smokeryes  regionnorthwest  regionsoutheast  regionsouthwest
##         23836.3           -352.2          -1034.4           -959.4
```

Figure 2: Final model from backward stepwise regression



```
## Step:  AIC=23314.58
## charges ~ smoker + age + bmi + children + region
##
##            Df  Sum of Sq        RSS   AIC
## <none>                   4.8845e+10 23315
## - region    3 2.3320e+08 4.9078e+10 23315
## + sex       1 5.7164e+06 4.8840e+10 23316
## - children  1 4.3596e+08 4.9281e+10 23324
## - bmi       1 5.1645e+09 5.4010e+10 23447
## - age       1 1.7151e+10 6.5996e+10 23715
## - smoker    1 1.2301e+11 1.7186e+11 24996
```

Figure 3: Final model from mixed stepwise regression



```
lasso.coef
```

```
##     (Intercept)             age         sexmale             bmi
##     -11270.2740        250.9591          0.0000        314.5540
##        children       smokeryes regionnorthwest regionsoutheast
##        392.1943      23567.3485          0.0000       -393.8077
## regionsouthwest
##       -380.5802
```

Figure 4: LASSO coefficients of corresponding variables

Appendix B - Interaction Terms

Interaction Term	Significance
BMI::Smoker	2e-16
Smoker:age:	0.972
Age::BMI	0.227
Age:Children	0.848
Children:BMI	0.767
Smoker::Children	0.170

Figure 5: Interaction terms and their significance in the model

Appendix C - Box Cox Transformation

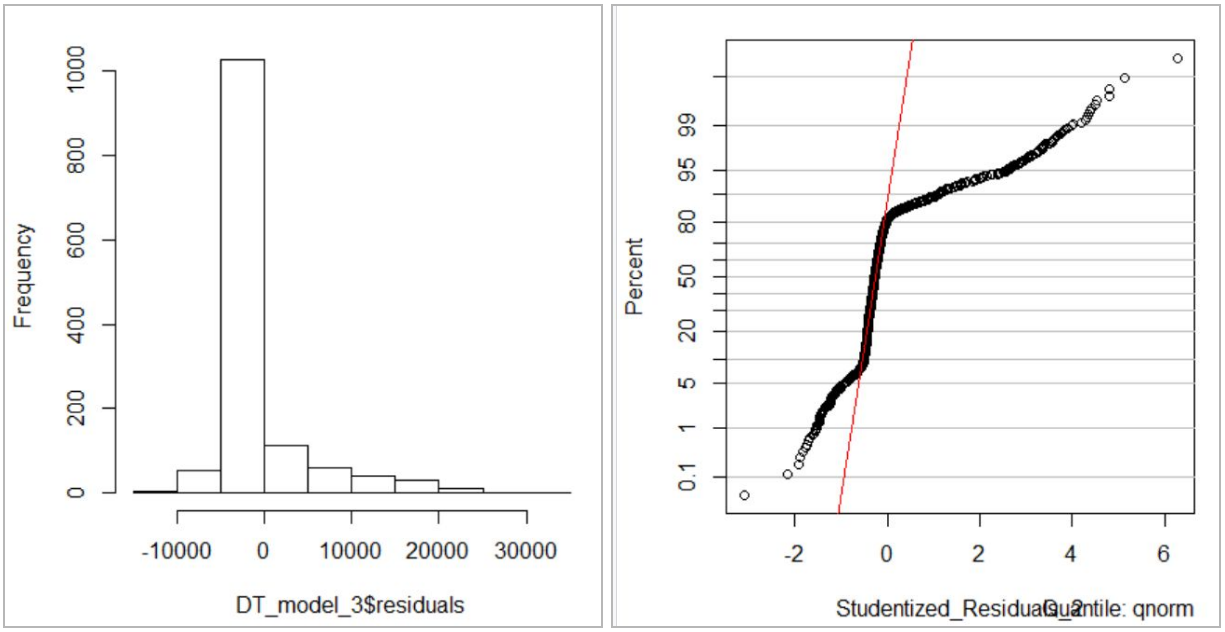


Figure 6: Histogram of residuals for fitted values (a) and normality plot for residuals (b)

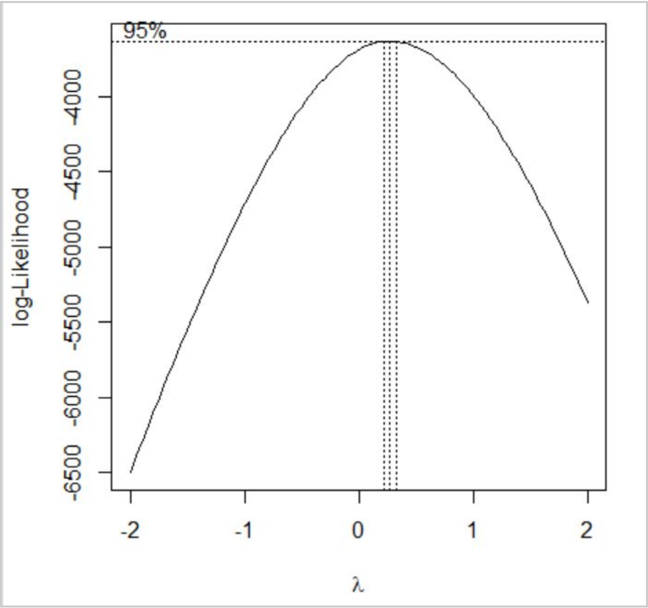
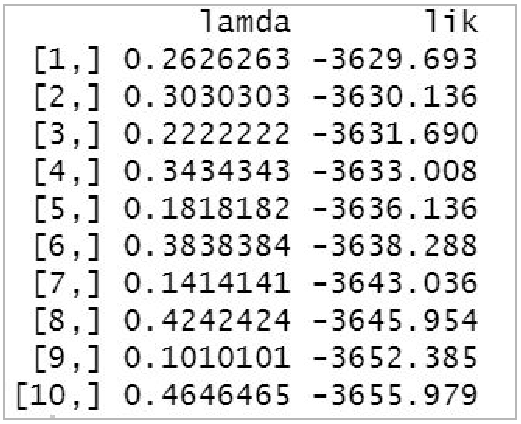
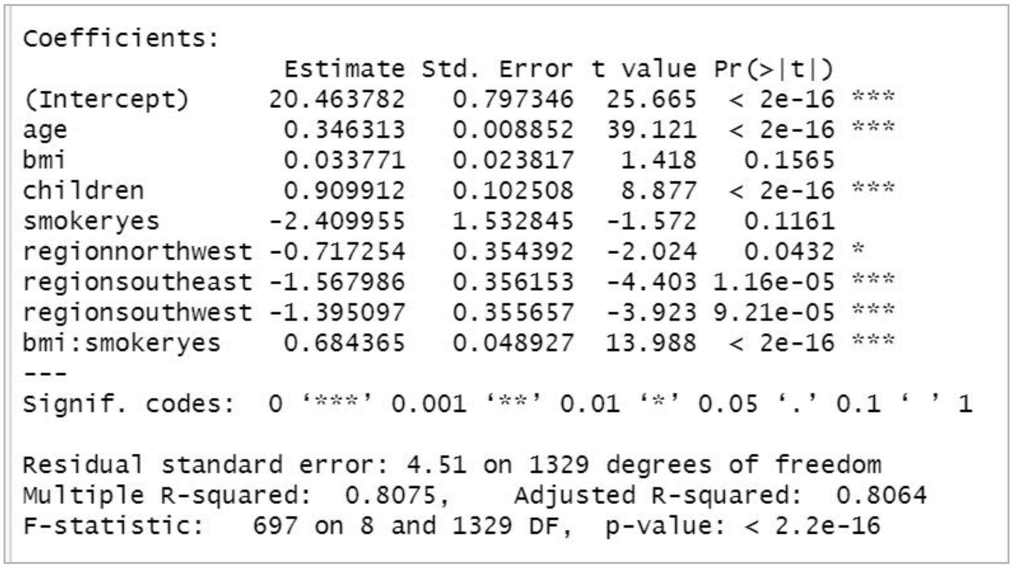


Figure 7: Log Likelihood graph for finding λ



```
            lamda       lik
 [1,] 0.2626263 -3629.693
 [2,] 0.3030303 -3630.136
 [3,] 0.2222222 -3631.690
 [4,] 0.3434343 -3633.008
 [5,] 0.1818182 -3636.136
 [6,] 0.3838384 -3638.288
 [7,] 0.1414141 -3643.036
 [8,] 0.4242424 -3645.954
 [9,] 0.1010101 -3652.385
[10,] 0.4646465 -3655.979
```

Figure 8: λ values



```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     20.463782   0.797346  25.665  < 2e-16 ***
age              0.346313   0.008852  39.121  < 2e-16 ***
bmi              0.033771   0.023817   1.418   0.1565    
children         0.909912   0.102508   8.877  < 2e-16 ***
smokeryes       -2.409955   1.532845  -1.572   0.1161    
regionnorthwest -0.717254   0.354392  -2.024   0.0432 *  
regionsoutheast -1.567986   0.356153  -4.403 1.16e-05 ***
regionsouthwest -1.395097   0.355657  -3.923 9.21e-05 ***
bmi:smokeryes    0.684365   0.048927  13.988  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.51 on 1329 degrees of freedom
Multiple R-squared:  0.8075,	Adjusted R-squared:  0.8064 
F-statistic:   697 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Figure 9: R output after the transformation