Term Life Insurance

Sean Alldridge, Alvina Putri, Erin Pederson

University of British Columbia

Authors Note

Sean Alldridge, Department of Economics, University of British Columbia

Alvina Putri, Department of Computer Science, University of British Columbia

Erin Pederson, Department of Mathematics, University of British Columbia

Keywords: Statistical learning, Regression analysis, Decision trees, Supervised learning

Table of Contents

Authors Note **1**

I. Introduction **3**

II. Data **3**

III. Methodology and Application **4**

Analysis 1: Multiple Linear Regression 4

Analysis 2: Principal Components Analysis 5

Analysis 3: Decision Trees 7

Analysis 4: K-Nearest Neighbours Regression 8

Analysis 5: Bootstrap Aggregating 9

Analysis 6: Lasso 10

Analysis 7: Neural Networks 11

IV. Conclusion **12**

References **13**

I. Introduction

The American life insurance market is a significant component of the insurance industry and ultimately of the United States economy at large. Nearly 80% of American households owned life insurance in 2004 (He, 2009). Insurance companies are continuously jostling with competitors for market advantage. Through research and innovation, they strive to deliver superior products. Therefore, to gain a competitive advantage, they are most interested in who the consumers of life insurance are and what quantity of life insurance they are interested in purchasing. In other words, they are in effect interested in the demand side economics of the life insurance market. In our study, we examine the attributes of households that contribute to the quantity of life insurance demanded by consumers. Our primary focus is on term life insurance. We used seven statistical learning methods in an attempt to offer meaningful predictions of the quantity of life insurance demanded by households conditional on certain household characteristics. In the following section, we first introduce the data set and its scope. In section III, we discuss the seven statistical learning methods, their methodologies and their application to our data set. We will be using multiple linear regression, principal components analysis, decision trees, K-Nearest Neighbours regression, bootstrap aggregating and random forest, lasso, and neural networks for our analyses. In the last section, we offer our conclusions as to which variables are most important in determining the demand for life insurance by households and interesting outcomes resulting from the analyses.

II. Data

The term life data set studied herein was acquired from the *Modeling with Actuarial and Financial Applications* textbook (Frees, 2009). The Federal Reserve Board conducted a survey of households with positive incomes and documented changes in U.S family finances for the period 2001-2004. The random sample of data was taken from 500 households with positive incomes interviewed in the 2004 Survey of Consumer Finances (SCF) (Frees, 2010). This data set contains 500 observations and 18 variables. The quantity of insurance is measured by the policy FACE, the amount that the company will pay in the event of the death of those insured. Important characteristics in the data set include annual INCOME, the number of years of EDUCATION of the survey responded, and the number of household members, NUMHH. We note here that of the 4522 families surveyed, only 4519 observations were available publicly[1]. In addition, a random sample of 500 from the 4519 observations and 18 variables of the nearly 5000 was selected by Frees (2009). Our dataset includes these 500 observations and 18 variables. Furthermore, as we could not find any meaningful definitions for the variables FACECVLIFEPOLICIES, CASHCVLIFEPOLICIES, BORROWCVLIFE, and NETVALUE, they have been omitted from our study. Below is a simple scatterplot matrix summarizing our economically important variables.

[1] Missing data in the survey of consumer finances was imputed five times using a multiple imputation technique. See federalreserve.gov for further details on the imputation technique used.

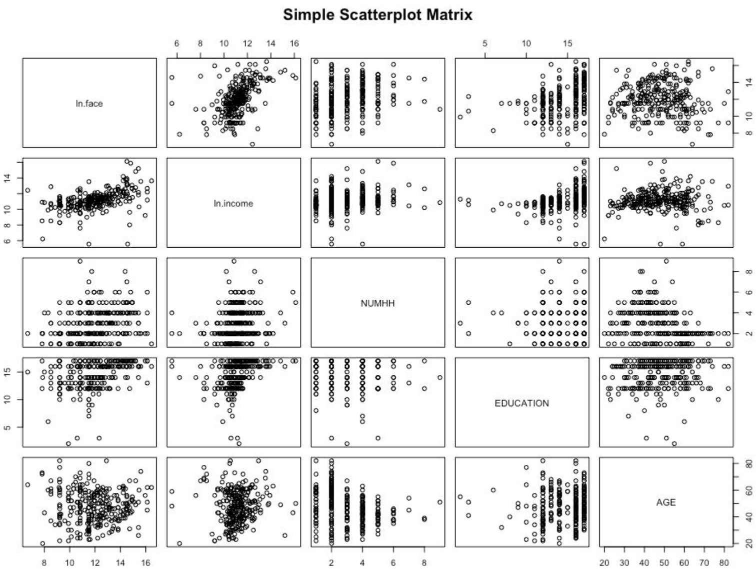


Figure 1. Simple Scatterplot Matrix

III. Methodology and Application

In this section, we provide a detailed description of each statistical learning method used to analyze the term life data set. We do not assume the reader to be an expert in the field of any particular quantitative discipline, however, we do assume the reader to be competent, at an advanced undergraduate level, in mathematics and statistics.

Analysis 1: Multiple Linear Regression

Multiple linear regression is a relatively simple, supervised learning approach. It is particularly good at predicting a quantitative response variable, given that the true underlying functional form of the data is approximately linear. The goal of linear regression is to find a line as close to our *n* observations as possible. This is done through minimizing the residual sum of squares. In effect, ordinary least squares reduces the problem down to estimating *p* beta coefficients. We note here that certain assumptions must hold for the $\widehat{\beta}_j$ ‘s to be unbiased and consistent estimators. These assumptions include that the conditional mean of the error terms be zero, independent and identically distributed observations and non-zero finite fourth moments exist. Moreover, once the line of best fit has been constructed through estimating the parameters, one can predict a response value given a new observed predictor value.

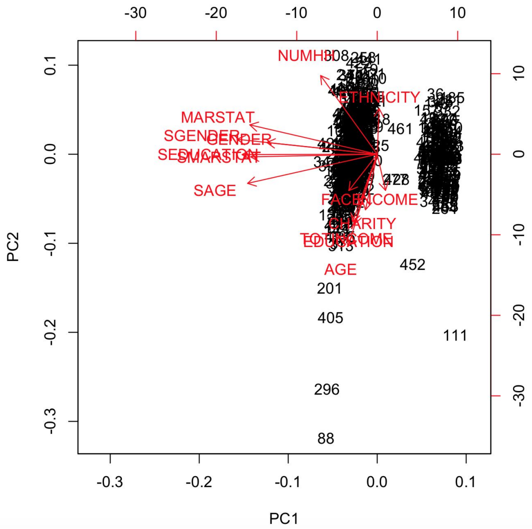
In our multiple linear regression analysis, we regressed the natural logarithm of FACE onto the natural logarithm of INCOME[2], the number of household members (NUMHH), the education of the named insured (EDUCATION), and their age (AGE). We used backward variable selection to determine our significant predictors. After the first regression, we found AGE was not statistically significant (*p*-value = 0.84) and thus removed it from the model. We

[2] A logarithmic transformation was useful here as a simple plot of FACE onto INCOME showed the data clustered in the lower left hand corner. After the logarithmic transformation, the plot appeared more linear.

refit the model excluding AGE and found all remaining variables to be significant. Our final model is a log-log specification:
$\widehat{LN(FACE)} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \widehat{\beta}_3 X_3$ Where $X_1 = EDUCATION$, $X_2 = LN(INCOME)$, $X_3 = NUMHH$ and, $\widehat{\beta}_0 = 2.58$, $\widehat{\beta}_1 = 0.206$, $\widehat{\beta}_2 = 0.494$, $\widehat{\beta}_3 = 0.306$. Our regression results showed an adjusted r-squared of 33%. That is, the above predictors explained approximately 33% of the variation in the response variable. Interestingly, all estimated beta coefficients are positive. In other words, a one-unit increase in any of the predictors in the model results in an increase in demand for term life insurance[3].

Analysis 2: Principal Components Analysis

The term life data set faces a large set of correlated variables and the principal components analysis allows us to summarize this set with a smaller number of representative variables that collectively explain most of the variability in the original set. Principal component analysis is a procedure by which principal components are computed and these components are used as predictors in a regression model in place of the original larger set of variables in order to understand the data. We illustrate the use of PCA on the term life data set. For each of the 500 positive income households observed, the data set measures the quantity of insurance by 14 variables. The rotation matrix constructed from the PCA provides the principal component loadings and each column contains the corresponding principal component loading vector. Note that there are 14 distinct principal components which is expected because there are in general min(n-1, p) informative principal components in a data set with n observations and p variables. The principal component score vectors have length n = 500 and the principal component loading vectors have length p = 14. PCA was performed after standardizing each variable to have mean zero and standard deviation 1. A plot of the first two principal components is shown below:

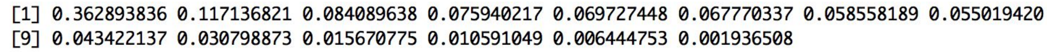


[3] Note that to have a meaningful prediction, we must use exponentiation to LN(FACE) as the resulting predictions are in logarithmic dollars. In addition, a one-unit change in LN(INCOME) is a 1% change, as this variable has been log transformed.

Figure 2. Principal Component Analysis Biplot

The biplot is scaled in order to ensure that the arrows are scaled to represent the loadings. We see that the first loading vector places most of its weight on ETHNICITY and INCOME, while the second loading vector places most of its weight on NUMHH, the number of household members. Furthermore, we observe that the variables (MARSTAT, SGENDER, GENDER, SEDUCATION, SMARSTAT, SAGE), and (FACE, INCOME, CHARITY, TOTINCOME, EDUCATION, AGE) are located close to each other which indicates that they are correlated with one another.

Table 1. Proportion Variance Explained



```
[1] 0.362893836 0.117136821 0.084089638 0.075940217 0.069727448 0.067770337 0.058558189 0.055019420
[9] 0.043422137 0.030798873 0.015670775 0.010591049 0.006444753 0.001936508
```

Based on the proportion variance explained (PVE) values in table 1, we see that the first principal component explains 36.3% of the variance in the data, the next principal component explains 11.7% of the variance, and so forth. Together, the first two principal components explain 48% of the variance in the data and the last two principal components explain only 8% of the variable. We plot the PVE explained by each component as follows:

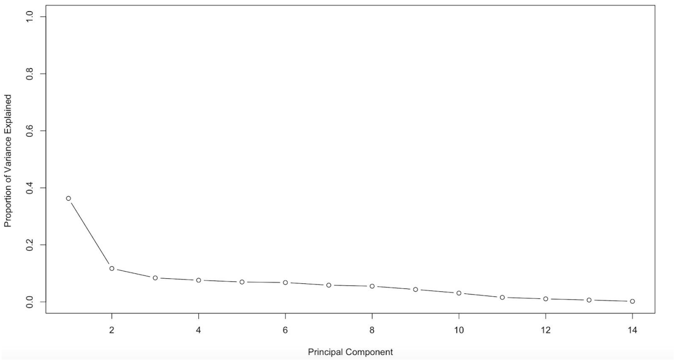


Figure 3. Principal Component Analysis Scree Plot

We typically decide on the number of principal components to visualize the data by examining the scree plot. We choose the smallest number of principal components that are required in order to explain a sizable amount of the variation in the data. This is done by "eyeballing" the scree plot and looking for a point at which the proportion of variance explained by each subsequent principal component drops off. This is often referred to as an *elbow* in the scree plot. By inspecting the plot, we can conclude that a fair amount of variance is explained by the first two principal components as there is an elbow after the second component. The rest of the principal components following that explains less than 10% of the variance in the data, therefore it is essentially worthless.

Analysis 3: Decision Trees

For this analysis, we performed a tree based method using our term life data set to predict the amount an insurance company will payout in the event of a death of the named insured (FACE) based on various family characteristics. The process of building a regression tree is, simply put, a two step process. In the first step, the feature space is partitioned into R_j $(j = 1, \ldots, J)$ regions. The regions of the predictor space are chosen to be divided into boxes for simplifying reasons and interpretational convenience. The aim of the algorithm is to obtain higher-dimensional rectangles $R_1, \ldots, R_J$ that minimize the RSS. However, considering all possible divisions of the predictor space is not feasible, so the algorithm takes a recursive binary splitting path. The second step consists of assigning an identical prediction to every observation that lands in a particular region R_j. This prediction is the expected value of the response values for all the training observations that are contained in the region R_j. Figure 4 shows the decision tree that results from top down greedy splitting. The variables actually used in the tree construction included: charitable contributions, income, age, age of spouse, and number of household members. The number of terminal nodes is eleven.

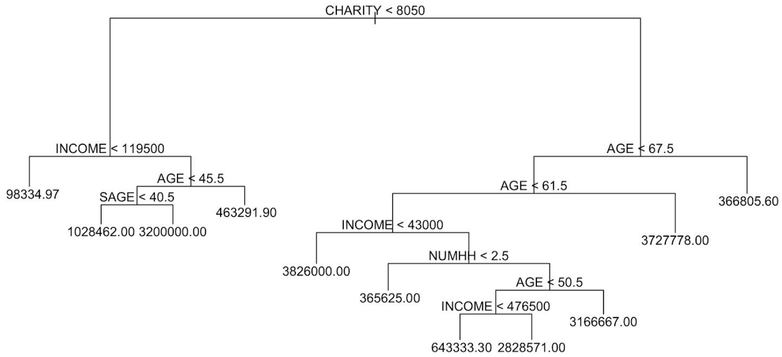


Figure 4. Decision Tree

Our results from the regression tree show that buyers of insurance who had charitable contributions greater than \$8,050, was less than 61.5 years of age and had an annual income less than \$43,000 was expected[4] to purchase \$3,826,000 worth of term life insurance. This is economically significant since we expect those with lower incomes to be unable to purchase such a substantial amount of life insurance. However, a reasonable explanation is that many respondents (observations) that fall within this region are affluent families with low annual incomes and live alone. Hence, affording this quantity of life insurance is feasible given their large net worths.

Analysis 4: K-Nearest Neighbours Regression

Parametric methods, such as multiple linear regression, assume a particular functional form for $f(X)$. This is an obvious disadvantage to the parametric method family as the possibility of this assumption being incorrect exists. If the assumption does not hold for the true underlying function, this may lead to materially inaccurate predictions. KNN regression, on the other hand, is a non-parametric approach and hence does not make an explicit functional form assumption. As a result, the non-parametric approaches such as KNN regression are more

[4] Note that the number in each terminal node is the mean (expected) value of the responses that belong to that particular region.

flexible than parametric approaches. The following is an algorithm of the KNN regression method[5]. First, we must specify a value for K, and have a particular prediction point x_0. Then, the closest training observations to x_0 are identified by KNN regression. This set of closest training observations is represented by $\mathcal{N}$. KNN then estimates the function f by inputting x_0 and averaging over all the training responses that belong to $\mathcal{N}$. The optimal value of K is contingent upon the bias-variance tradeoff and a sampling method such as k-fold cross validation may be used to select the optimal value of K. This is completed next on our data set. The knn.reg function in R studio performs leave-one-out cross-validation implicitly. A loop was run to find the highest r squared and it was determined that $k = 9$ was the optimal k to be used in the KNN regression. We performed KNN regression with $p = 1$. The KNN regression of LN(FACE) onto LN(INCOME) can be found in the figure below:

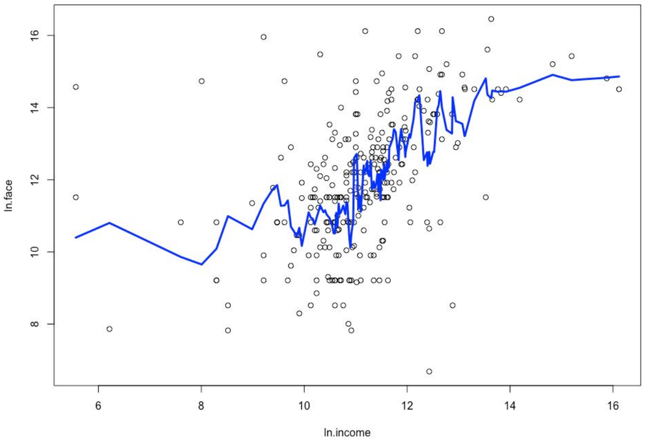


Figure 5. KNN Regression Plot

Analysis 5: Bootstrap Aggregating and Random Forests

To build upon our results from the decision tree we have performed Bagging and Random Forests on the data. Bagging stands for bootstrap aggregating. Since regression trees can result in high variance bagging is a good way to reduce that variance. The method for bagging is taking B regression trees that are made from B groups of bootstrapped training sets. Then you take the average of the B trees, each individual tree has high variance but it's reduced by taking the average of all the trees (James, 2013). The downfall of bagging is that we give up interpretability for accuracy of prediction. Random forests are step up from bagging, with bagging you try to split the tree on all the predictors. Thus, if one of the predictors are most important than the rest the B tree will end up being correlated. With random forests if we have p predictor variables we would only use m of them to split on. For every B tree, we will have m predictors which will change every time. The decision for m is based on $\sqrt{p}$; when m=p this is simply bagging.

[5] The algorithm of the KNN regression method discussed here is as described by James, Witten, Hastie, & Tibshirani (2013).

For this analysis, we started out using bagging. The analysis indicates that between 75 to 100 trees would reduce the error below $1.6\text{x}10^{12}$. The bagging model uses 500 trees and it tries 13 predictor variables at each split of the tree. With the bagging model, we get a training MSE of $1.49\text{x}10^{12}$ and a testing MSE of $2.28\text{x}10^{12}$. The predictors of this model explain 8.01% of the variance is the FACE variable, this seems surprisingly small. From the variable importance plot we found that INCOME is the most important with TOTAL INCOME coming in second. Next, we performed a random forest analysis. The analysis indicates that around 50-75 trees will be enough to produce the lowest error, below $1.5\text{x}10^{12}$. This already shows that random forests is doing a better job than bagging. With random forests, we tried to use 4 variables at each split because $\sqrt{p} \approx 4$. The training MSE for random forests was $1.41\text{x}10^{12}$, which is much lower than the bagging model, and the testing MSE was $2.08\text{x}10^{12}$. The predictors of this model explain 12.91% of the variance in the variable FACE, again this shows that random forests are a better model than bagging. From the variable importance plot we found that INCOME is still the most important variable with CHARITY now coming in second.

Analysis 6: Lasso

Since our data set is great for regression techniques we thought it was fitting to perform Lasso on the data. Lasso is the second shrinkage method of two that we studied, the other being Ridge Regression. Both Lasso and Ridge Regression combine with least squares technique using a shrinkage penalty. The shrinkage penalty uses a tuning parameter, $\lambda \geq 0$, which is combined with a summation term of the absolute values of the estimated coefficients, for lasso. The benefits of using the shrinkage penalty is that the estimated coefficients are being shrunk toward zero so we can determine which coefficients are doing the best job of estimating the desired relationship. Lasso performs better than ridge regression because with ridge regression the final model contains all the predictors since the estimated coefficients can only approach zero they never actually reach zero. The simple change that is the lasso method allows for the estimated coefficients to be equal to zero which is why we chose to use it.

Before we perform the cross-validation we examine a coefficient plot. The coefficients approach zero at different speeds as λ changes but eventually they will all be zero. Next, we performed cross-validation of the training set to find the optimal λ.Comparing the two coefficients plots they seem virtually the same. The first predictor to enter the model is total income and last predictor to enter is education of the named insured. This is an interesting result because in the final model education is still included. Which resulted in the following plot:

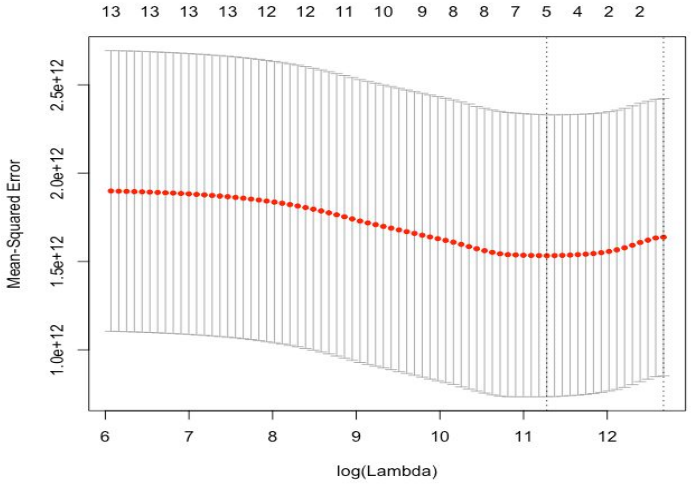


Figure 6. MSE Plot

The Figure 6 shows us how the Mean Squared Error changes with Log(λ) along the bottom, and the number of predictors along the top. As log(λ) increase and the number of predictors decrease we can see the MSE decreases. From this plot we find two important values, the λ that results in the lowest MSE and the λ that results in the simplest model. When producing the final mode we are left with four predictor variables and the intercept. The predictors that are left are education, spousal education, number of household members, and total income. The second model, which is the simplest means only the intercept is left. This model has slightly larger MSE than the first model. Surprisingly when the testing MSE of the two models were calculated the first model (best model) had a higher MSE (2.78x10^{12}) than the second model (simplest) with an MSE of 1.72x10^{12}.

Analysis 7: Neural Networks

Neural network is constructed with an interconnected group of nodes, which involves the input, connected weights, processing element, and output. This analysis can be applied to many areas, such as classification, clustering, and prediction. Using the term life data set, we will fit a simple neural network in order to predict the FACE value of the quantity of insurance using all the other continuous variables available. Before fitting the neural network, we must address data pre-processing. Normalizing the data before training a neural network is very important as avoiding to do so may lead to useless results or a very difficult training process. For this data, we chose to use the min-max method and scale the data in the interval [0,1]. Additionally, we use two hidden layers with the following configuration: 13:5:3:1, where the input layer has 13 inputs, the two hidden layers have 5 and 3 neurons and the output layer has a single output, since we are performing regression. The following shows a graphical representation of the model with the weights on each connection:

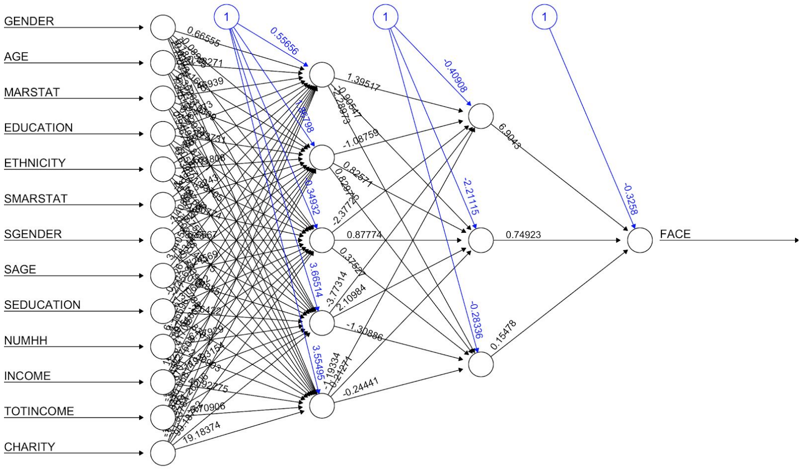


Figure 7. Neural Network Model

The black lines show the connections between each layer and the weights on each connection while the blue lines show the bias term added in each step. The bias can be thought of as the intercept of a linear model. The net is essentially a black box so we cannot say that much about the fitting, the weights and the model. Suffice to say that the training algorithm has converged and therefore the model is ready to be used. From this neural network, we will predict the FACE value for the test set. A first visual approach to the performance of the neural network model on the test set is plotted below:

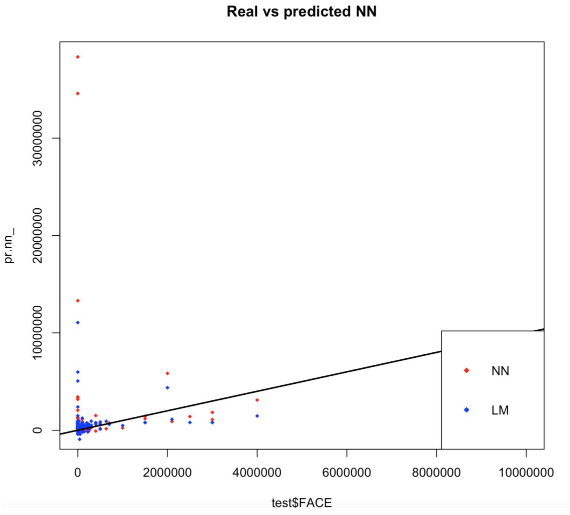


Figure 8. Prediction by Neural Network Plot

By visually inspecting the plot (Figure 8) we can see that the predictions made by the neural network (red) are generally closer around the line than those made by the linear model (blue), thus making it a more ideal prediction.

IV. Conclusion

Our study used many various forms of statistical learning methods to analyze the term life data set. Throughout our analyses, we saw many insightful results. The multiple linear regression analysis showed 33% of the variation of the response variable FACE was explained by the predictors. That is, 33% of the demand for life insurance was explained by our predictors. Using Principal Components Analysis, we identified that there are 14 distinct principal components. Based on the biplot, we observe that the variables (MARSTAT, SGENDER, GENDER, SEDUCATION, SMARSTAT, SAGE), and (FACE, INCOME, CHARITY, TOTINCOME, EDUCATION, AGE) are located close to each other which indicates that they are correlated with one another. The application of the decision tree method to our data also had interesting results. We found those with charitable contributions greater than $8,050, less than 61.5 years of age and had an annual income less than $43,000 was expected to purchase $3,826,000 worth of insurance. This prediction is likely due to a financially well-off spouse or third party (an outlier) purchasing this substantial quantity of insurance for the survey respondent. Furthermore, K-Nearest Neighbours regression resulted in a relatively flexible model that therefore may suffer from high variance when predicting new untrained observations. This may be of particular importance to insurance companies when predicting the quantity of life insurance a prospective household will purchase. When performing the bagging and random forest analysis we find that INCOME, TOTAL INCOME, NUMHH, and CHARITY were the most important variables in determining the demand for life insurance. This result agrees with our initial beliefs of this data set. We used the lasso analysis which resulted in four predictor variables being leftover: education, spousal education, number of household members, and total income. Moreover, Neural Networks shows that predicting the FACE value using the model provides a better and more ideal prediction. We see that through all these analyses that total income, education and number of household members prevail over the rest of the predictor variables when estimating the value of term life insurance demanded.

References

Frees, E. W. (2009). *Regression modeling with actuarial and financial applications*. Cambridge University Press.

Frees, E. W., & Sun, Y. (2010). Household life insurance demand: A multivariate two-part model. *North American Actuarial Journal*, *14*(3), 338-354.

He, D. (2009). The life insurance market: asymmetric information revisited. *Journal of Public Economics*, *93*(9), 1090-1097.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: springer.