Determinants of Individual Health Insurance Coverage:

A Statistical and Visual Analysis

Louis Booth

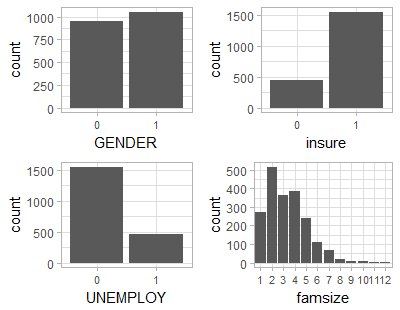
Determinants of Individual Health Insurance Coverage:

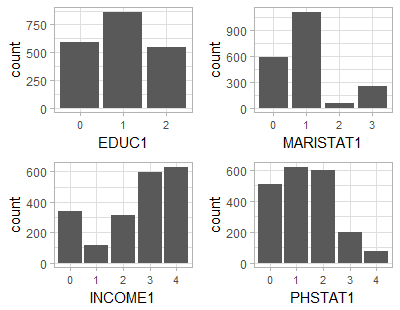
A Statistical and Visual Analysis

Data for this project was acquired from the Wisconsin School of Business and consists of information obtained from the Medical Expenditure Panel Survey. I began by importing the data and creating a data frame to be subset. The head and structure of the data frame were examined to locate variables of interest. Education level (less than high school, high school, college), marriage status (never married, married, widowed, divorced or separated), income (poor, near poor, low income, middle income, high income), and physical health status (excellent, very good, good, fair, poor) are factors. Gender (female=1, male=0), unemployment status (unemployed=1, employed=0), and insured status (insured=1, uninsured=0) are binary variables with two levels. Family size (1 to 12) is a continuous variable representing number of persons.

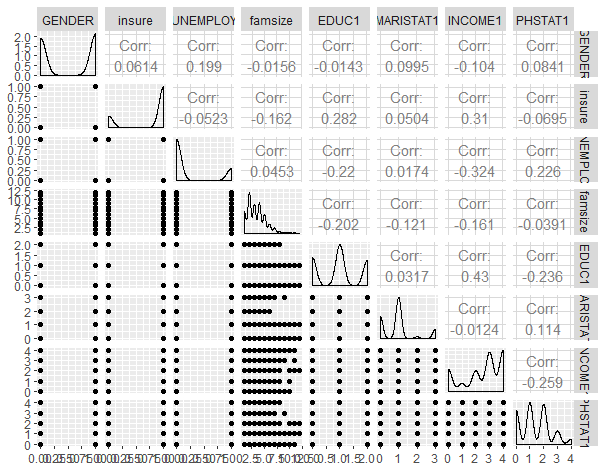
The subset data frame has 2000 rows (observations) and 8 columns (variables). No missing or NA values were found in the subset data frame. There were 730 duplicated rows, I chose not to remove them because each row represents an individual, and it is not unreasonable to expect that many individuals would share common traits in respect to the metrics being considered. Correlations between variables were calculated and rounded, confirming that multicollinearity was not an issue. This satisfies one of the assumptions underlying logistic regression, as none of the predictor variables included are heavily correlated.

Frequency tables were created for each predictor and the response variable to illustrate relative frequencies of individual insurance coverage for each level of each predictor. I utilized the ggplot2 package to plot the frequencies of each individual variable.

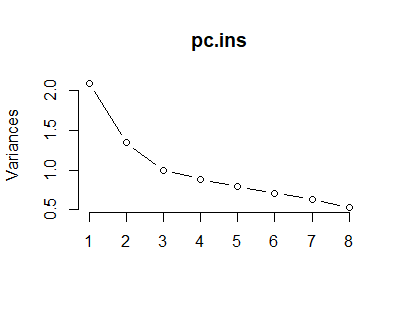


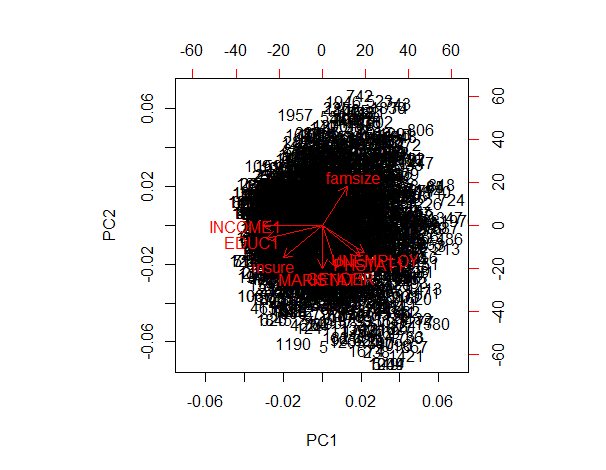


The package GGally was employed to visualize relationships between variables.

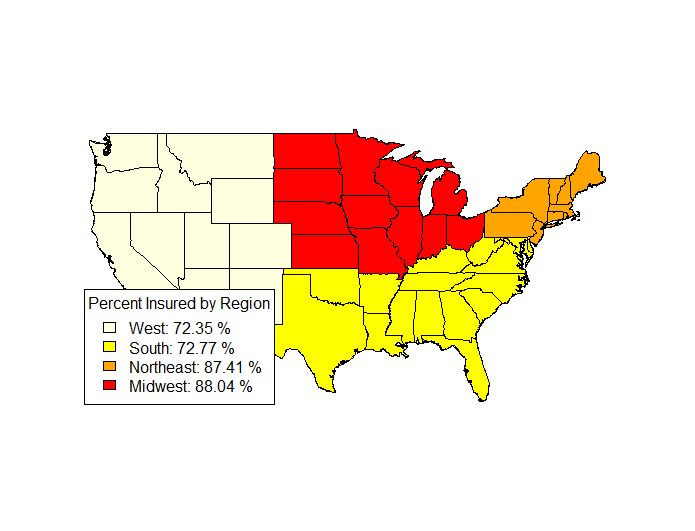


Principal component analysis was performed, and the principal components were plotted in a line graph against variances. A biplot was utilized to show the relationships between principal components, as well as the spread of the individual observations. The biplot tells us that income and education are strongly related the second principal component, while unemployed status, physical health status, married status, and gender are all strongly related to the first principal component. Insured status is related to both principal components, and family size seems to not be related to either principal component.





The subset data frame was then indexed and tallied based on region of the United States, as well as based on region and individual insurance status. The goal was to calculate the proportion of insured observations, to all observations, by region. These calculations were used to create a heat map of the four regions of the United States by percentage of insured individuals. The maps package was utilized to create the visualization.



When it was time to perform the logistic regression, I created a new subset data frame with variables in the form of (Excellent, Very Good, Good, Fair, Poor) rather than (0, 1, 2, 3, 4), for physical health status, as well as other factors, to avoid the implicit rankings of integer values. From the results, we find that the intercept, gender, family size, education: high school, education: less than high school, married status: married, income: low income, income: middle income, income: near poor, and income: poor are all statistically significant within the model. However, unemployed status, married: never married, married: widowed, physical health status: fair, physical health status: good, physical health status: poor, and physical health status: very good, are not statistically significant within the model.

The coefficients of the regression model can be interpreted as follows, while holding all other predictors constant: Females have an increase of .482 in the log odds of being insured, over males. An increase of 1 family member decreases the log odds of being insured by .134. Being unemployed increases the log odds of being insured by .22. Having a high school education rather than a college degree leads to a .791 decrease in the log odds of being insured. Having less than a high school education rather than a college degree leads to a decrease of 1.334 in the log odds of being insured. Being married rather than divorced or separated leads to an increase of .404 in the log odds of being insured, while never being married leads to a decrease of .139 and being widowed leads to a decrease of .254. With having a high income as the base category, the log odds of being insured decrease by 1.244 if you have a low income, decrease by .447 if you have a middle income, decrease by 1.614 if you are near poor, and decrease by 1.566 if you are poor. The base category for physical health status is excellent, as such, having fair health leads to an increase in the log odds of carrying health insurance by .027, good health leads to a decrease of .255, poor health leads to an increase of .625, and very good health leads to a decrease of .008.

Finally, fitted values were predicted from the logistic regression, they were rounded, and displayed in a table. Actual values of insurance status were presented in a similar table. Both tables were then combined to obtain the number of correctly, and incorrectly, predicted cases. This information was used to display the percentage of successful, and unsuccessful, predictions of individual insurance coverage status:

Successful prediction percentage for insured individuals: 95.43 %

Successful prediction percentage for uninsured individuals: 22.25 %

Unsuccessful prediction percentage for insured individuals: 4.57 %

Unsuccessful prediction percentage for uninsured individuals: 77.75 %

The model is much more accurate when it comes to predicting whether an individual will carry insurance, rather than whether they will not. This is likely because only about 25% of the individuals in the sample are uninsured.

References

Wisconsin School of Business, Regression Modeling with Actuarial and Financial Applications. (2017). Medical Expenditure Panel Survey Health Expenditures [Data set]. Retrieved from <http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html>