AMAT 465: Applied Statistics

Multivariable Regression Analysis
Medical Insurance Costs
Instructor: Karin Reinhold
Team: Jahnavi Bonagiri and Nidhi Vadnere
October 28, 2022
University at Albany, SUNY

Section 1: Introduction and Purpose Health care is one of the most costliest expenses when living in the U.S. costing families hundreds of dollars per month. According to the U.S Bureau of Labor, medical insurance is one of the top five spending costs for many middle and low income families. As many families need to budget their daily expenses in an efficient manner, it is important to recognize the factors that may influence a family's rising costs so they can better manage their resources. Economic stress on medical costs can affect long-term monetary dreams of families. The solution to conquer this crisis is to help assist these families in predicting and assisting them with the rising costs. Health insurance is a product that covers fees associated with medicine, surgical procedures or hospital visits of an insured which could be an individual, family, or a collection of people. When people first hear of medical insurance costs, factors such as health history, age, gender, and children first come to mind. The purpose of this project is to gain a deeper understanding of what factors play the most important role in identifying which families have the higher insurance expenses. By understanding the key factors that affect medical costs, we can predict in advance about the health insurance expenses, which could prove to be very beneficial for insurers and patients to manage their assets appropriately. Section 2: Data Source

The dataset we acquired for the project was provided by a verified data scientist on Kaggle on medical costs. The main data we will be using is the charges as the response variable and bmi, age, children, smoker history and region as the explanatory variables. There was prior work done using different models such as Lasso and Random regression, so we will use this opportunity to perform a thorough analysis to conclude a solid result to see if insurance costs are correlated with bmi, age, children, region and medical history using a Linear regression.

We will be using all of these variables to perform regression analysis of various models: Table 1: Numerical Variables Attributes Description age age of primary beneficiary bmi Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ˆ 2) using the ratio of height to weight, ideally 18.5 to 24.9 charges Individual medical costs billed by health insurance

Table 2: Categorical Variables Attributes Description sex insurance contractor gender, female, male 1=Male 0=Female

smoker Smoker or non-smoker 1= yes 0= no region the beneficiary's residential area in the US, northeast, southeast, southwest, northwest. southwest=1 southeast=2 northwest=3 northeast=4

children Number of children covered by health insurance / Number of dependents

Section 3: Data Analysis

Pairs.r When performing a matrix analysis on the quantitative variables, we were able to conclude that there was an interaction present between expenses vs bmi. There was also an interaction between age vs bmi. Moreover, as we see in the scatter plots, the effect of age is additive, producing 3 regression lines that are parallel, with only the intercept changing according to the model. The slope did not change. We will need to perform more analysis of the regression lines to see if the correlation is stronger with or without the interaction effects.

Code: source("pairs.r")
pairs(insurance[c(7, 1,3)],panel=panel.smooth,diag.panel=panel.hist,lower.panel=panel.cor)

Boxplot of Categorical Variables:

Code: par(mfrow=c(2,2)) boxplot(charges~as.factor(sex),data=insurance.2) boxplot(charges~as.factor(region),data=insuranc
boxplot(charges~as.factor(children),data=insurance.2) boxplot(charges~as.factor(smoker),data=insurance.2)

Transformations: Upon analyzing the linear model with all variables without any transformations and interactions, we get an R-Squared value of 0.7519 and an Adjusted R-Squared value of 0.7497; Given this high R-Squared value, we can infer that the variance in Charges is accounted for by the explanatory variables. We then ran a model by applying a poly of 2 to bmi and age, with remaining variables being categorical. This gave us a R-Squared value of 0.7561 and an Adjusted R-Squared value of 0.7536, slightly more than the values in the first linear model.

We updated this model and applied a logarithmic transformation to the predictor variable, which gave us an R-Squared value of 0.7727 and an Adjusted R-Squared value of 0.7703, even better than the previous model with only polynomial terms. Alternatively, the model with only a logarithmic transformation to the predictor variable and non polynomial terms had an R-Squared value of 0.7703 and an Adjusted R-Squared value of 0.7682. After looking at all of these models, it appears that the model with the logarithmic transformation and polynomial terms has the best fit. Table 3: Models

Model R-Squared Adjusted R-Squared Test 1: lm(charges~as.factor(sex)+bmi+as.factor(children)+as.factor(smoker)+as.fact
0.7519 0.7497 Test 2: lm(charges~as.factor(sex)+poly(bmi,2)+as.factor(children)+as.factor(smoker)+as.factor(region)+poly(
0.7561 0.7536 Test 3: lm(log(charges)~as.factor(sex)+poly(bmi,2)+as.factor(children)+as.factor(smoker)+as.factor(region)+p
0.7727 0.7703 Test 4: lm(log(charges)~as.factor(sex)+bmi+as.factor(children)+as.factor(smoker)+as.factor(region)+age,data
0.7703 0.7682

Significance of Preliminary Model: Using bmi, age, sex, children, region, smoker and the response variable of expenses we were able to get a good linear trend going between the variables for the preliminary model. The regression line we used was lm(charges~as.factor(sex)+bmi+as.factor(children)+as.factor(smoker)+as.factor(region)+age,data and there was no interaction included for the initial model. The summary of the model gave us an Adjusted R-Squared value of .7497 and R-Squared of .7519 indicating that the model fits fairly well. In addition, all the categorical and numerical variables included in the summary produced low p-values showing that they are quite significant. However, when we performed various transformations on the variables using log and poly, those models had slightly higher R-Squared and Adjusted R-Squared values. This shows that although the preliminary model was fairly fit and significant, there were other models that fit the data better and accounted for the variability. We will be analyzing those models more thoroughly to get accurate results.

Interaction Terms: Upon examining the plot matrix developed using pairs., it appears that bmi could potentially be an interaction term. By adding an interaction term of bmi to the model with logarithmic transformation and polynomial terms, we got a R-Squared value of 0.7888 and an Adjusted R-Squared value of 0.7848.

Literature Review:

`There were many international papers that analyzed the medical costs using these datasets. In the paper`

0.380189, an MAE value of 0.17448, and an accuracy of 85.82. They concluded that Stochastic gradient boosting can be used in the estimation of insurance costs with better performance than other regression models.

Conclusion: We will be analyzing more literature papers and other works more thoroughly to understand the concepts at a deeper level and also perform more transformations using regression models. For the preliminary data analysis though, we were able to get the best model to be lm(log(charges)~as.factor(sex)+poly(bmi,2)+as.factor(children)+as.factor(smoker)+as.factor(region)+poly(age,2),data=insu which showed to have high summary statistics, and low p-values. It is vital to better forecast insurance costs based on certain factors to help insurance providers and families to better manage their assets ahead of time.

Works Cited: https://www.kaggle.com/datasets/mirichoi0218/insurance?select=insurance.csv https://www.researchgate.net/publication/348559741_Predict_Health_Insurance_Cost_by_using_Machine_Learning_and_DNN_Regression_Models https://ijcrt.org/papers/IJCRT2203462.pdf https://healthpayerintelligence.

com/news/health-insurance-costs-placing-stress-on-majority-of-americans#:~:text=Seventy%2Dtwo%20percent%20of%20m
https://www.bls.gov/news.release/cesan.nr0.htm

an introduction which motivates the problem. A description of where you obtained the data. A description of the variables in the data set. Nidhi A preliminary data analysis of the variables involved: do pairs.r on all the quantitative variables so that we see the plots and the correlations between variables. If you have categorical variables, do boxplots of the response variable ~ categorical variable. Nidhi Do your variables need to be transformed? Should you transform the predictor variable(s)? Should you transform the response variable? Polynomial terms? Nidhi A preliminary model for the study. Are all variables significant? Should you include interaction terms? Nidhi A literature review of other people's work that may be relevant to your study. NEED TO DO