Avinash Dindial

STA302

Final Assignment part III

Introduction

In the modern world, especially in the US, health insurance costs have been putting a financial burden on most families, especially with the introduction of the coronavirus in recent times. I believe strongly in universal healthcare, and I believe most insurance companies exploit their customers, using their health as collateral. That is why this topic is of interest to me. It is fairly common to come across articles online that criticize health care giants charging exorbitant charges for their health care services, indeed a google search can yield thousands of results. In my research of this topic, there is mostly political reasons as to how these charges are applied, but I am unable to find substantial results for what physical or personal variables affect cost. I believe the regression analysis I am undertaking may be different from others published. I would have liked to use bigger datasets, but the one chosen was the best I can find. Given the apprehensiveness some have in getting health insurance due to the cost, I think it is relevant to determine what insurance companies are likely to be looking for when determining cost and determining if it is justified. The goal of this project is to find which of the given variables are predictive of insurance costs. In addition, I would want my model to be easy to understand as my audience would be the general public.

Method

In this section, I will describe how I narrowed down the number of predictors to get to my final model. Before I did anything else, I first split my dataset into training and testing data. To try to get the optimum model, I made use of several selection techniques and used each to get a model. So, at the end, I was left with a handful of potential models. For each potential model, assumptions were checked and transformations etc., were performed to assure each potential model can be a proper solution to this project's problem question. After all this was done, I compared each model according to their adj. R^2, AIC and BIC values to see which is the best model.

Firstly, I made use of the backward selection method to come up with the simplest potential model. I would prefer an interpretable model, so less predictors makes it less confusing and can be understood by a wider audience.

Secondly, I used the best selection method which is a function built in R to determine the most significant predictors in a dataset. This model was chosen according to its adjusted R^2 value.

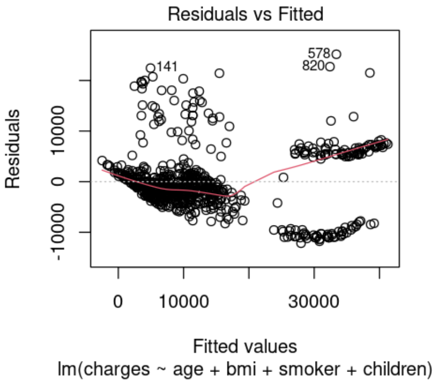
I did not want to rely on just these automated techniques to select models, so I manually decided to build my own. Firstly, using information from the EDA, I noticed that the two biggest predictors of cost were smoking and BMI. So, I decided to create a new variable which categorized BMI into "Obese" and "Healthy" where an obese person is someone whose BMI is higher than 29. This allowed me to pair this with the smoking variable as an interaction term. My main hypothesis for building this model was that obese smokers paid substantially more.

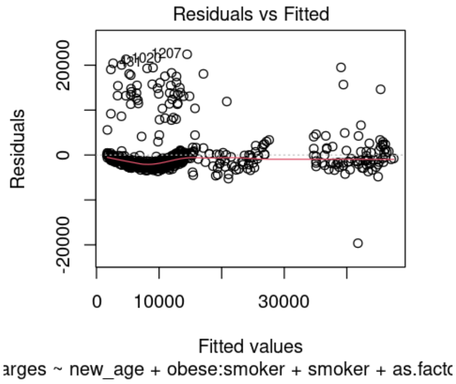
After getting these potential models, I checked the assumptions of each and checked for any evidence of multicollinearity using VIF on each model. In cases where transformations were used, I decided to sacrifice multicollinearity in favor of fixing my model assumptions. Once everything was corrected and assumptions satisfied, then I checked each model's AIC, BIC and adjusted R^2. I also used prediction functions in R to determine how predictive each model is. In theory, the most appropriate model is the one with the lowest AIC, BIC but highest adjusted R^2 values. Using literature as a secondary tool, I concluded due to predictive power and based on empirical evidence that the model with the interaction terms would be the best. Lastly, I validated my model using test data to make sure it can be used in independent datasets.

<u>**Results**</u>

	Adjusted R^2	# of predictors	Validated?	Assumptions violated	Fixes	Multicollinearity?
Full model	76.8	8	Yes	Clumping of residuals	Transformation of response	No
Automated model	76.3	4	Yes	Linearity and patterns in residuals	Transformation of response and predictors	No
Interaction model	85.7	5	Yes	Curved QQ plot, Residuals clumping	Transformation of predictors	No

An example of how I fixed violated assumptions in the interaction model:





Before diagnostics

After diagnostics

Our original dataset contained several variables to predict healthcare insurance charges: age, smoker(yes/no), region, children (number of children), sex, BMI. It was noted that charges were not normally distributed, and the age variable seemed to be slightly curved when plotted against charges. To fix this, I applied a log transformation on charges and raised age to the power of 0.5 to fix this linearity issue.

The mean age of all persons in the dataset is ~39 years old which can influence variables such as BMI and smoking is more prevalent in this age group. Unsurprisingly, smoking and BMI was also correlated with health charges, which is why I decided to link these two variables.

It was found that the model that was built based on interactions between obesity and smoking had the biggest adjusted R^2 of 86% compared to the others which each had an adjusted R^2 value of ~78%. In addition, it was found the AIC and BIC scores of one of the automatically generated models was lower than the full model. These automated models all had similar adjusted R^2 so I ended up choosing the model with the highest adjusted R^2 but lowest AIC and BIC scores. This chosen model was then compared to the interaction model. In the end, I had to choose between a model with very low AIC/BIC but lower interpretability and lower adjusted R^2 and a model with a slightly lower AIC/BIC but high interpretability and high adjusted R^2. I decided to choose the model with interactions that I built manually because of the high predictive power. In addition to this, ANOVA tests proved this model did not leave out any significant predictors. In essence, we conducted a hypothesis test where H0: we can remove the variables that we did. The result of our ANOVA test resulted in a small p-value which means the variables we removed were insignificant. This interaction model was also validated using test data (All comparisons are in the R code).

This allowed me to comfortably conclude my final model:

Name of variable	P-value
√Age	~0 ***
Smoker(yes)	~0 ***
1 Child	0.7
2 Children	0.01 *
3 Children	0.0246 *
4 Children	0.0357 *
5 Children	0.5546
Obese and Smoker	~0 ***
Obese Non-smoker	0.17

Charges = -11632.8 + (3151.9) √ (age) + 13278.3(smoker) + 174.5(1 child) + 1345.3(2 children) + 1323(3 children) + 4476.2(4 children) + 1000.3(5 children) + 556.7(Obese but don't smoke) + 19702.7 (Obese and smokes).

Our new age variable has a minimum value of 4.2 and a max of 8.0. The graph of this vs charges also shows a linear relationship, unlike its untransformed version (in appendix). The insignificant variables are not surprising as they are the ones we expect not to be related to an increase in health charges.

Discussion

This model further elaborates on the fact that BMI and smoking are highly predictive of cost. In this case, it would cost you $19702.70 extra if you are obese and a smoker. It also shows the fact that the more children a person has, the more they generally tend to pay for health insurance. However, an interesting observation is that people with no children is noticed to pay more (see graph2 in appendix). One explanation of this is that those without children may tend to live different lifestyles that probably involve more smoking or other unhealthy activities.

Consider person A who is 19, does not smoke, no kids and a healthy BMI would pay $2106.

Whereas person B who is also 19, smokes, no kids but is obese would pay $35087.

It seems as if insurance companies “punish” those who are obese who smokes. This indicates that you can pay less for health insurance if you take better care of your health. In other words, smoking is very harmful to your health and your wallet.

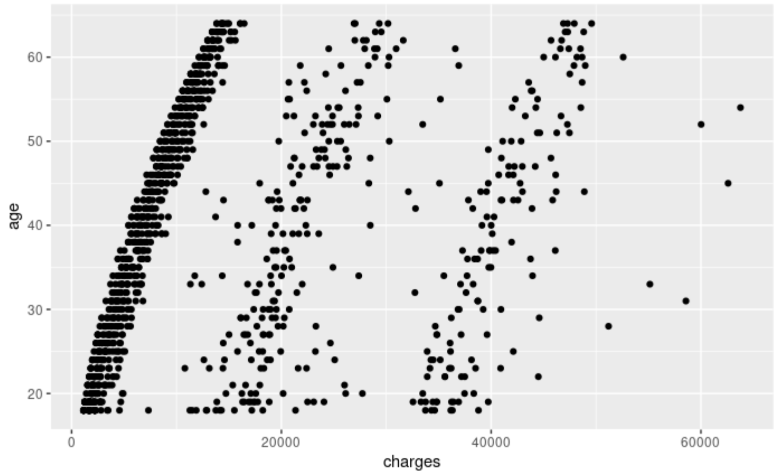
I believe this model is the best one to be used since it encompasses the main predictors of cost, is not overly complicated and is validated. Most of my audience would be the poorer population, so they may not be aware of their BMI etc., so splitting my model into simple categories and interactions I think is suitable for this project.

However, there are some limitations:

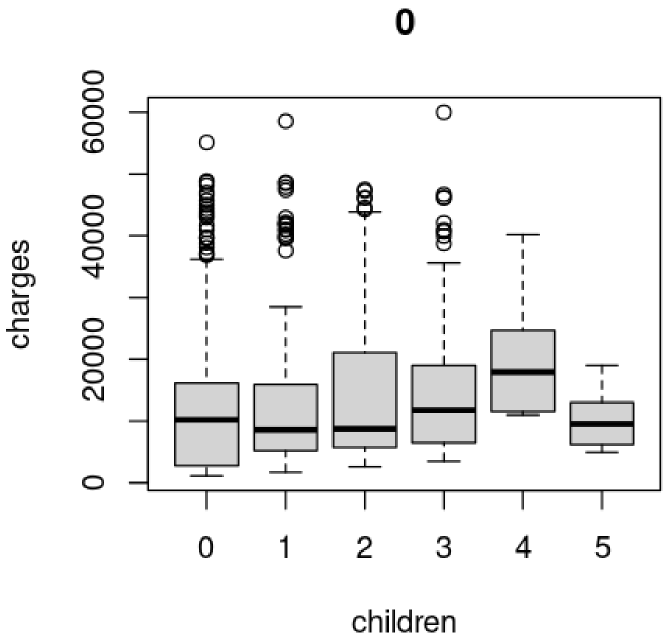
Firstly, in all potential models including my chosen model, the QQ plot is still slightly skewed. Although box cox method made it a little less skewed, there were still deviations from the normal line. This may have impacts on how good this model fits. This may be due to the presence of influential points, and indeed when checking DFFITS etc., it was noted that there were a high number(n>20) of influential points. Despite efforts to try to make the model as specific as possible, there is still some more that can be done. For instance, the smoking variable is vague as it is unclear how we defined a smoker from a non-smoker; making it a binary variable may be too simplistic.

Appendix

NOTE: All graphs are plotted in the R code.



Graph1



Graph2