Final Assignment

Group: Potato

GRA4157

1 Introduction

In recent years, there has been an increasing focus on leveraging data to make decisions. This has pushed the medical insurance industry to collect vast amounts of data on their customers and use it to set prices according to the firms perceived risk of the customer. Much research has been done on medical insurance, such as increasing costs and the social issues related to decreasing coverage in the public[[DMC05], [Had07]]. We have therefore chosen to focus our report on the ethical issues regarding using data to guide insurance prices. There are several regulations for the use of data, most prominently GDPR, but due to the technological landscapes rapid development, the current regulations leave many gray-areas.

2 Methodology

Our approach to uncover the underlying driving factors of medical insurance charges is to fit an appropriate model on the data and use the results to analyze its ability to predict insurance charges and which variables have a high impact. Before training a model, we needed to pre-process our data. The data contains a mixture of quantitative and qualitative variables. For us to be able to draw information from our qualitative variables we used the Dummy coding method[VSMS16] to have the information represented numerically in our dataset. This method replaces categorical variables with binary variables. Continuing, split the data into training and test sets in order to run in-sample and out-of-sample on our models. With the training data we start off with a OLS model including all variables in the dataset. The multiple linear regression model takes the form:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 age_i + \hat{\beta}_2 BMI_i + \hat{\beta}_3 1\{gender_i = "Male"\} + ... + \hat{\beta}_p 1\{smoker_i = "yes"\}$$

Running this regression resulted in an exceptional in-sample fit represented by an adjusted R^2 of 0,996. Additionally, according to the p-values of each estimate, all the variables seem to be significant. By making predictions on our test data and comparing these predictions to the real charges for the test data allows us to test the out-of-sample accuracy of our model as well. This resulted in a Mean Squared Error of 83 157, or 1,62% Mean Absolute Percentage Error. Our model exhibits exceptional out-of-sample prediction power as well, insinuating that the high R^2 is not due to overfitting. For further analysis of the model, we decided to look at the residuals of the model, which we calculate in the following way:

$$e = Y - X\hat{\beta}$$

Working with the OLS model, we have assumed that the residuals are normally distributed, but for our dataset it is uniformly distributed.

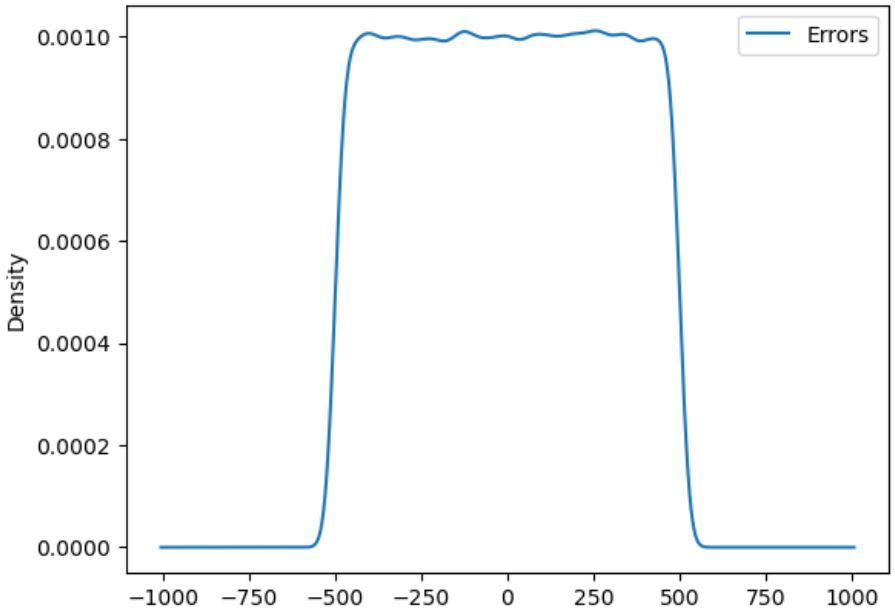

Figure 1: Distribution of the errors

From the figure above we clearly see that the distribution of the errors resemble a uniform distribution. The uniform distribution is symmetric just as the normal distribution, thus the line that best fits the data will be the same. The problem however, is that the estimators are not t-distributed in this case because of the uniformly distributed residuals, thus their associated p-values are unreliable. With unreliable test statistics we can't easily perform feature selection. To curcumvent this issue, we decide to run a Lasso regression to act as a sort of feature selection by pulling estimates deamed less important towards zero[Tib18].

3 Results

4 Conclusions

References

[DMC05] Michael Chernew Patricia Seliger Keenan David M. Cutler. Increasing health insurance costs and the decline in insurance coverage. *Health Services Research*, 40(4):1021–1039, 2005.

[Had07] Jack Hadley. Insurance coverage, medical care use, and short-term health changes following an unintentional injury or the onset of a chronic condition. *JAMA*, 297(10):1073–1084, 03 2007.

[Tib18] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 12 2018.

[VSMS16] M Venkataramana, M Subbarayudu, Rajani Meejuru, and K Sreenivasulu. Regression analysis with categorical variables. *International Journal of Statistics and Systems*, 11:135–143, 01 2016.