PREDICTION OF MEDICAL EXPENSES

A PROJECT REPORT

Presented to the Department of Mathematics and Statistics

California State University, Long Beach

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Mathematics

Option in Statistics

Faculty Reviewer:

Kagba Suaray, Ph.D.

By Ron Lidgi

B.S., University of California, Berkeley

May 2023

**Table of Contents**

Abstract

Introduction

Analysis

Validation

Bibliography

Appendix

**Abstract:**

This project was done to predict medical expenses using regression and machine learning. After finding linear model assumptions were not satisfied, I decided to build a model using the gamma GLM model. After learning that needing to meet the assumptions depends on the size of the data set, and comparing the results of the 2 models, I concluded that the linear regression model was the better model. I also compared the results of the regression models to model results using machine learning models.

**Introduction:**

The dataset that was used for this project is a simulated dataset of hypothetical medical expenses for insureds in the United States. I chose the data because I wanted to learn more about modeling data used to set prices for insurance. It was created using demographic statistics developed by the US Census Bureau, and therefore does approximate real world data. There are 1338 observations, and the response variable is medical expenses. The 6 explanatory variables are age, sex, bmi (body mass index), number of children, smoker/nonsmoker, and region of the United States. The data was randomly split using 80% of the data for training the model and 20% for testing.

**Analysis:**

I started with exploratory data analysis on the dataset. I looked at correlation plots between each of the variables. Also I plotted the histograms for each variable. This is a histogram of the response variable, medical expenses:

Chart, histogram

Description automatically generated

The histogram has a very heavy right tail. Starting at approximately 14500, there is a change, and the distribution of expenses is suddenly lower and almost uniform until the maximum, Then there is a sudden drop at about $30000. I don’t know why this is, but it could be because $15000 and $30000 are very popular limits that insurance companies pay on policies. I also plotted each of the independent predictor variables with the response variable, using a scatter plot. This is the plot with the age variable:

Chart, scatter chart

Description automatically generated

For each age group there appear 3 groups of expenses data points. Holding age constant, I wanted figure what variable moves expenses to the middle and top areas on the plot. I did this is by using the ggplot method in R and color coding the points by a variable. I tried each of the 5 other variables and the also pair interaction terms. The x3\*x5 term (bmi and smoker) had the best result. Then, I created another variable, highbmi, which had values of 1 for bmi above 32, and 2 for bmi above 32. The value for the interaction bmihigh\*smoker would have values 0,1, and 2. Since there were 3 areas in the plot, having 3 values would be a good idea. This result on the plot improved. This is the plot:

Chart, scatter chart

Description automatically generatedThis would mean that the term could be a very important part of the model, and it was added. After fitting the model, I found that the 6 main variables and that interaction variable were significant. However, the normal probability plot was showing the residuals were not normal. A transformation on the dependent variable also didn’t solve the problem.

Since, on the plot of x1 (age) vs. y, I noticed 3 groups of points, which look like regression lines, I thought maybe it was possible to build regression models for each of the 3 groups of data points. By looking at the histogram, I could determine that the first group of points ended at 14500, the second at 30000, and the third started at 30000. Using Excel, I separated the points into 3 datasets, and built a regression model on each. I fit the model and had significant variables each time, but again, when checking for normality, the models failed. Using an unsupervised machine learning method might help to find out more about the 3 areas of points.

Since the histogram of expenses for the complete dataset had a very long right tail, the response variable looks like a Gamma distribution. I decided to try a Gamma GLM regression model, using the log for the link function. This model relaxes the assumption that the residuals be normally distributed and homoscedastic. To check the data for multicollinearity, I examined the VIF (variance inflation factors) values, and did not find they were high. This is the inverse of the matrix of correlation of x variables, the VIF values are highlighted:

x1 x2 x3

x1 **1.01541145** 0.025285125 -0.114352459

x2 0.02528513 **1.008888244** -0.049547843

x3 -0.11435246 -0.049547843 **1.040582971**

x4 -0.04257110 -0.017234153 -0.004718337

x5 0.02429434 -0.075893455 -0.003535126

x6 -0.01648579 -0.003237873 0.163287406

x4 x5 x6

x1 -0.042571096 0.024294344 -0.016485794

x2 -0.017234153 -0.075893455 -0.003237873

x3 -0.004718337 -0.003535126 0.163287406

x4 **1.002481021** -0.007459798 0.015714332

x5 -0.007459798 **1.006467927** -0.003171505

x6 0.015714332 -0.003171505 **1.025925195**

When I fit the model, all the variables except x2 (sex) were significant according to the summary function, but x2 was significant using the Wald test. I therefore decided to include all the variables. I examined the added variable plots, and they indicated that variables should be added to the model, in a linear way.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.957505 0.264408 26.314 < 2e-16 \*\*\*

x1 0.029018 0.002883 10.066 < 2e-16 \*\*\*

x2 -0.061228 0.041048 -1.492 0.13702

x3 0.015319 0.006828 2.244 0.02570 \*

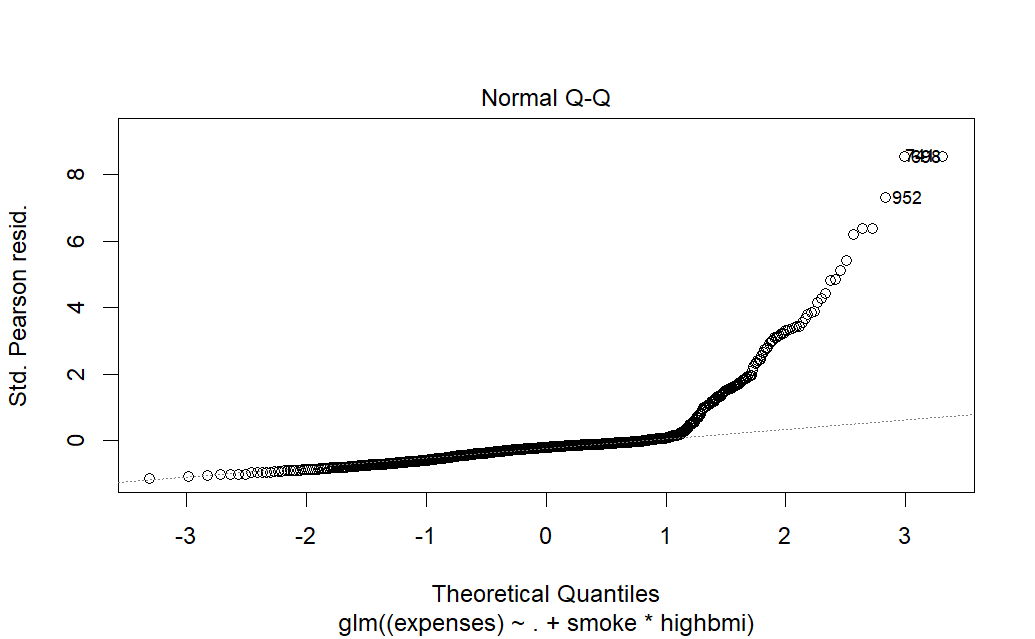
x4 0.078636 0.032974 2.385 0.01781 \*

x5 1.450349 0.102515 14.148 < 2e-16 \*\*\*

x6 0.116108 0.038746 2.997 0.00299 \*\*

smoke:highbmi 0.377279 0.104588 3.607 0.000324 \*\*\*

The normal probability plot showed outliers on the right tail, but, since after removing outliers, new outliers continued to show, removing outliers was not a solution. Using the boxcox method, I had result of 1, so, there was not a possible transformation. The residuals are not normal, but since this is a gamma GLM model, the normality requirement will be ignored. This is the normal probability plot:



This is the anova table for the model:

Terms added sequentially (first to last)

Df Deviance Resid. Df Resid. Dev

NULL 1071 833.98

x1 1 73.64 1070 760.35

x2 1 6.19 1069 754.15

x3 1 21.82 1068 732.33

x4 1 3.63 1067 728.70

x5 1 459.49 1066 269.21

x6 1 1.84 1065 267.37

smoke:highbmi 1 5.97 1064 261.40

To estimate the coefficients of a GLM model, the method of sum of least squares is not possible, and the method of maximum likelihood is used, instead. Numerical procedures are used to find the values that maximize the likelihood. Besides examining the variance of the estimator and its p-value when considering including the variable in the model, it also possible to use Wald tests based on the ANMLE theorem. This second method uses the likelihood ratios of the reduced and full model. The value calculated is known as the deviance, G2. G2 has a chi square distribution with degree of freedom equal to the difference in the number of variables in the reduced and full model. If the test is done for one variable, values above 3.841 are significant. The value for the MSE was 61272749 and the MSPR was 66586634.

After reading a post on the Stack Overflow website, I found that “that the estimators in the linear regression model are not particularly sensitive to heavy tails in the error distribution”, and that “the coefficient estimators……are still usually quite reasonable.” This data set is large enough, that the normality of the response variable is not an issue. Due to the Central Limit Theorom, even when the data comes from a distribution that is not normal, the sample mean is approximately normal. I also learned the downside is there is a reduction in statistical power. I decided I would use the linear regression model. The value for MSE was 25234786and MSPR was 26828202, which was much lower than with the gamma regression model. These are the coefficient estimates:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -7645.67 1087.05 -7.033 3.60e-12 \*\*\*

age 260.27 11.11 23.434 < 2e-16 \*\*\*

sex -148.39 155.10 -0.957 0.33892

bmi 218.05 42.79 5.096 4.11e-07 \*\*\*

children 446.23 129.30 3.451 0.00058 \*\*\*

smoke 792.86 1162.38 0.682 0.49533

region 349.61 141.58 2.469 0.01369 \*

highbmi -1595.40 543.10 -2.938 0.00338 \*\*

smoke:highbmi 16371.78 782.89 20.912 < 2e-16 \*\*\*

This the anova table:

Response: expenses

Df Sum Sq Mean Sq F value Pr(>F)

age 1 1.3501e+10 1.3501e+10 531.5181 < 2.2e-16 \*\*\*

sex 1 1.0058e+09 1.0058e+09 39.5959 4.558e-10 \*\*\*

bmi 1 3.6549e+09 3.6549e+09 143.8892 < 2.2e-16 \*\*\*

children 1 4.8135e+08 4.8135e+08 18.9500 1.471e-05 \*\*\*

smoke 1 9.6473e+10 9.6473e+10 3797.9889 < 2.2e-16 \*\*\*

region 1 6.2667e+07 6.2667e+07 2.4671 0.116549

highbmi 1 1.8607e+08 1.8607e+08 7.3255 0.006907 \*\*

smoke:highbmi 1 1.1108e+10 1.1108e+10 437.3119 < 2.2e-16 \*\*\*

Residuals 1063 2.7001e+10 2.5401e+07

**Machine Learning Methods:**

I decided to compare the results on the data using regression, to what I would get using a machine learning model. I used 3 machine learning methods on the data set Random Forest, Gradient Boosting, and Neural Network. These are the results:

|  |  |  |
| --- | --- | --- |
| Method | MSE (training set) | MSPR (test set) |
| Random Forest | 3456284 | 23186109 |
| Gradient Boosting | 15038456 | 19027354 |
| Neural Network | 14745037 | 22146354 |

I tried different numbers of nodes and layers in the neural network to get the best results. The Random Forest model fit the training data very well, but did not generalize well on the test data. All the models had better results than the regression model, but the Gradient Boosting model was the best.

**Conclusion**:

This project was done to predict medical expenses using regression and machine learning. After finding linear model assumptions were not satisfied, I decided to build a model using the gamma GLM model. But, I concluded that the linear regression model was a possible model. Also, I found it could be possible to use a machine learning model.

Bibliography

Applied Linear Statistical Models, Kutner, 5th Edition

Comparing Gamma and Log-Normal GLMs in R Using Simulation and Real Data Set, Dr. Nagham Mohammad

<http://h2o-release.s3.amazonaws.com/h2o/master/1292/docs-website/datascience/glm.html>

<https://medium.com/swlh/modeling-insurance-claim-severity-b449ac426c23>

https://analyse-it.com/docs/user-guide/101/normality-central-limit-theorem

**Appendix:**

Boxplot of Residuals

Chart

Description automatically generated with medium confidence

Chart, histogram

Description automatically generated