364 Set1

Süleyman Buğra Gülsoy

2283109

Abstract

The research was conducted to analyze the blood pressure level of individuals. To make an prediction. The aim of the prediction was identifying the high blood pressure risk of an individuals that follows certain characteristics. The main research question was to come up with a model that can predict the blood pressure level in a high efficient way. The generalized linear model was used to investigate the relation between blood pressure and independent variables. Variety of different distribution and with different log link had tested to find the most efficient model. The boxcox and theother common transformation method used to find a better relation between variables. Interaction effect between independent variables were tested.The models were compared mainly with RMSE parameter and their goodness of fit. A stable and good functioning model was accomplished with gamma distribution using log link. The important parameter at predicting blood pressure was identified and inferenced. It has found that the most relevant variables for predicting the Systolic Blood Pressure were: Age, Body Mass Index, Maximum Inflation Levels, Alcohol, Magnesium and 100 cigarettes smoking and all of them except “Magnesium” found out to be statistically increase the blood pressure level of an individual. Thus, it has suggested that to be aware such factors to avoid high blood pressure.

Introduction

The blood pressure of an individual is a critical factor for health. High blood pressure can cause serious crisis and emergency care needed. There can be various factor that effect blood pressure and research is about identifying the factor that effect blood pressure to avoid any health problem caused by high blood pressure. The Blood pressure data has been obtained by “National Health and Nutrition Examination Survey” every year. The data is open source for research and analyzing purposes. We tried to model data with different approaches to come up with the best model to explain the high blood pressure. Details of the methods are available in the methodology. Hence, we aimed to create a model that can detect the blood pressure level of an individual with certain parameters. Thus the model can warn the individuals having certain parameters about their risk ofhigh blood pressure. And without hospitalized people can make an inference about their blood pressure risk. And avoid certain habits that enhances the risk of having high blood pressure with the help of our model.

Literature Review

There are plenty of previously done researched about the same data. Similar findings were obtained throughout year. Yet, more complex statistical analyses method and model was proven to be working better with the data set. Also some past researches used wider range of parameters whichmade their conclusion more genialized.

Research Questions

Which variables are most relevant at predicting the systolic blood pressure Which

model is the best fit at predicting the systolic blood pressure

Is data appropriate for general linear model

Methodology

The data had split into test and training with respect to 70% 30% rule. Multiple liner regression and generalized linear regression was used to investigate relationship between variables. Shapiro Wilk test was used to checking the normality of variables. Than, boxcox method of transformation was used to make variable distributions more linear form. Also most common transformation method had tested to help with normality problem. QQ plot and skewness test were used to understand the distribution of residuals. Gamma and invers gaussian generalized linear models with different link functions were create as candidates for proposal model. Model were trained with stepwise regression with AIC being criteria parameter. Also cross validation control method was used in training. Pearson and log-likelihood goodness of fit test were conductedbetween candidates models. To test how well model fits the data and data meets the assumptions of the model. Also as comparison parameter between different GLM’s Mc Fadens pseudo R squared and Root mean squared error were used. Validation set is used to validate the estimatedmodel parameters.

Data Set

Characteristic	**N = 4,728**[1]
DMDHHSIZ	
1	633 (13%)
2	1,271 (27%)
3	889 (19%)
4	802 (17%)
5	562 (12%)
6	275 (5.8%)
7	296 (6.3%)
INDHHIN2	2,097,152 (78,125, 105,413,504)
RIAGENDR	
1	2,284 (48%)
2	2,444 (52%)
RIDAGEYR	46 (31, 62)
BPXSY1	118 (110, 132)
BMXBMI	28 (24, 32)
BPXPULS	
1	4,645 (98%)
2	83 (1.8%)
BPXML1	140 (140, 160)
DR1TALCO	0 (0, 0)
DR1TSODI	3,186 (2,272, 4,367)
DR1TPOTA	2,392 (1,713, 3,204)
DR1TCALC	808 (521, 1,189)
DR1TMAGN	267 (192, 365)
DR1TPROT	74 (52, 102)
DR1TKCAL	1,944 (1,438, 2,596)
SMQ020	
1	1,985 (42%)
2	2,743 (58%)

[1] n (%); Median (IQR)

The data consist 17 variables: “SEQN” : Unique identification number of observation “DMDHHSIZ” : Total number of people in the Household Discrete variable 1 to 7 “INDHHIN2” : Annual household income divided into categories 1 to 15 “RIAGENDR” : Gender variable 1 is male 2 is female “RIDAGEYR” : Age “BPXSY1” : Systolic blood pressure (mm Hg) “BMXBMI” : Body Mass Index (kg/m**2) “BPXPULS” : Pulse being regular 1 or irregular 2 “BPXML1” : Maximum Inflation Levels (mm Hg) “DR1TALCO” : Alcohol (gm) “DR1TSODI” : Sodium (mg) “DR1TPOTA” : Potassium (mg) “DR1TCALC” : Calcium (mg) “DR1TMAGN” : Magnesium (mg) “DR1TPROT” : Protein (gm) “DR1TKCAL” : Energy (kcal) “SMQ020” : Smokedat least 100 cigarettes in life 1 = yes 2 = no 9= don’t know

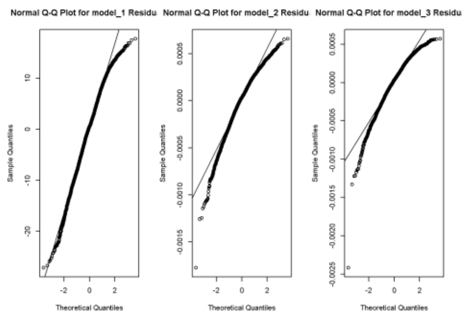
The data has in total 12 numeric 3 categorical and 1 ordered categorical variable, “BPXSY1” being the target variable and 15 regressor. The ordered variable converted into numeric for ease of the computation. The variable “INDHHIN2” indicated the Annual household income divided bygroups from 1 to 15. Thus we transformed it to ordered factor. But it had two problematic factors 77 which means person refused to answer the question and 99 which means person does not know the income. So we decided to remove those observations as they were really a small part of the sample (155 out of 4884). Similarly the variable “SMQ020” which indicated whether the person had smoked 100 cigarettes in his/her life. Had an answer to indicate not knowing (in total only one person). Thus, that observation was removed from the data.

Full data available at : https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey Result and

Findings

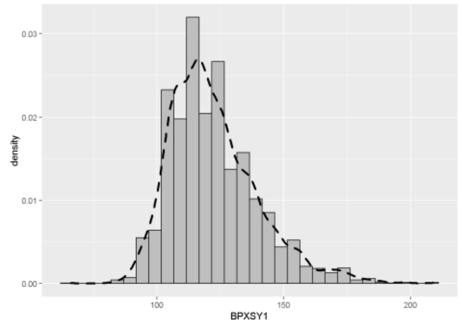
The data did not had any missing values when we started analyzing. To investigate the relation between systolic blood pressure and the predictors, normality of variables was controlled. By the Shapiro wilk test non of the variables were distributed normally($p < .05$). A box cox transformation was applied to the all the non normal variables.

The box cox transformation was successfully transformed only one of the variables(BMXBMI) into normal distribution. Thus, three general linear modelwas created one with all the variables transformed with boxcox, one with only BMXBMI transformed. And, one with using untransformed data.



All three models did not satisfied the normality of error. However, the best model satisfying the normality error was untransformed data, also in terms of R-squared values untransformed data was the better option. To check is there any helpful transformation on dependent variable despite the boxcox (we observed in model_3 that boxcox was not helping with normality error violation). Thus, We tried square root transformation, natural logtransformation, log base 10 transformation and inverse transformation on dependent variable. None of them solved the normality error violation problem. But, in the model_1 graph which was the best result so far was indicating a negative skewness. (Despite dependent variable having a right skewness the residuals had left skewness) Thus, we reflected the entire data by subtracting every data point from its maximum value and adding “1” to it.

After obtaining the reflection of the data, square root, natural log, log base 10, inverse, and boxcox transformations were implemented none of themsolved the non normality of residuals. Thus, we concluded that data is not appropriate for general linear models. We decided to move on to generalized model approach.



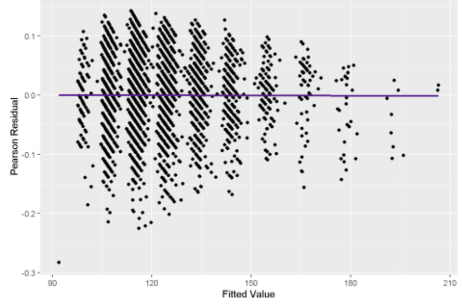
The distribution of the target variable’(“BPXSY1”) structure was more suitable for gamma and inverse gaussian. Since, the variable had a non zerocontinuous distribution with right skewness. Hence, we created different models with different family and link functions.

The inverse gaussian model with canonical link was not appropriate for the data since the mean function for it’s canonical link involves square root and data had some negative values at independent variables so function created NA’s. Yet, we successfully created in total 4 glum models. Gamma with canonical link, gamma with log link, inverse gaussian with log link and gaussian with log link. Hence we compared the four models in termsof their goodness of fit to the data, overdispersion, Mc Fadden’s Pseudo R squared, and RMSE from test data set.

Comparison of GLM Models

Model_Name	Deviance_Fit_p	Pearson_Fit_p	Likelihood_Raito_Overall_Model	Dispersion_Test	McFaddens_Psudo_R_square	RMSE
model_gamma	1	1	0.0000000	0.0044310	0.7698126	0.2034565
model_gamma_log	1	1	0.0000001	0.0042688	0.7783747	0.0635109
model_invg_log	1	1	0.9999451	0.0000361	0.7677684	0.0721631
model_gaus_log	0	0	0.0000000	62.6033997	0.7947873	0.0550715

The Gaussian model with log link did not fitted the data as it’s p value for deviance fits is 0. Also, inverse gaussian model with log link did not fitted as well. Since, null model for it and our model was almost identical according to it’s likelihood ratio test. The best model for out data was obtainedas gamma distribution with log link. It had the better pseudo R squared and RMSE value out of good fitted models. Yet, the model was suffering from under dispersion problems. We attempted to fix it by adding interaction term to the model. However, all possible combinations of two-way interactions did not solve the under dispersion problem nor did a improvement on RMSE. On the contract, the gaussian model does not have any restrictions about it’s variance such as binomial or poison data. Thus, under dispersion shouldn’t be a much of a problem. We examined the error distributions with graphs as follow.



Train Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.6770157	0.0116940	314.4358886	0.0000000
RIDAGEYR	0.0001884	0.0000721	2.6116110	0.0090522
BMXBMI	0.0014988	0.0001613	9.2898468	0.0000000
BPXPULS2	-0.0098881	0.0083657	-1.1819861	0.2372958
BPXML1	0.0072574	0.0000802	90.5169479	0.0000000
DR1TALCO	0.0001073	0.0000412	2.6072841	0.0091672
DR1TPOTA	0.0000016	0.0000017	0.9849757	0.3247077
DR1TCALC	0.0000010	0.0000024	0.4238134	0.6717294
DR1TMAGN	-0.0000107	0.0000136	-0.7852296	0.4323749
SMQ0202	0.0023212	0.0023247	0.9984735	0.3181225

Validation Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.7178748	0.0180503	205.9732643	0.0000000
RIDAGEYR	0.0003345	0.0001151	2.9070489	0.0037066
BMXBMI	0.0010273	0.0002512	4.0899535	0.0000456
BPXPULS2	-0.0173688	0.0143441	-1.2108665	0.2261532
BPXML1	0.0070492	0.0001226	57.4950799	0.0000000
DR1TALCO	0.0000984	0.0000763	1.2890101	0.1976098
DR1TPOTA	0.0000004	0.0000025	0.1487470	0.8817749
DR1TCALC	-0.0000010	0.0000039	-0.2534855	0.7999307
DR1TMAGN	-0.0000095	0.0000197	-0.4810656	0.6305459
SMQ0202	0.0042512	0.0036078	1.1783625	0.2388545

The Pearson residual versus fitted value graph shows no indication of problem. The proposed model for data is concluded as gamma log. Wevalidated the proposed model by, recreating the same model from test data set. We observed that the estimation of variables “BPXPULS”, “DR1TPOTA” and “DR1TCALC” was extremely different between two models. We decided to remove those for finalize the model.

We used all the observation combined to obtain our finalized proposal model for the estimation problem.

Final Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.6891701	0.0097486	378.4323303	0.0000000
RIDAGEYR	0.0002155	0.0000602	3.5812931	0.0003453
BMXBMI	0.0013687	0.0001355	10.1042442	0.0000000
BPXML1	0.0072017	0.0000670	107.5280808	0.0000000
DR1TALCO	0.0001040	0.0000358	2.9083844	0.0036500
DR1TMAGN	-0.0000002	0.0000060	-0.0290846	0.9767983
SMQ0202	0.0028986	0.0019468	1.4889155	0.1365764

Discussion and Conclusion

We tried to fit the data for general linear model despite all the attempts data didn’t fit. we concluded that the data was not eligible for such a fit. Than, we obtained variety of GLM models to compare as a result. We compared those models in terms of their goodness of fit, overdispersion, root mean squared error, pseudo R squared values. In the light of our model testing parameters. We proposed a gamma log model. We tried to use a ordered factorial variable at first but, it’s quadratic forms made predict function hinder from properly working. Thus, we transformed the ordered factorial variable to numerical one. But we observed that the correct form of that variable was significant for our model. We lost a bit of predictionpower from not being able to use ordered factorials. Than we validated our propose model so observe it’s fit to the assumptions and compared the validation model. We decided to remove some variables which has extreme differences between train and validation models. Than we conductedthe final model. With all of the observations included. We concluded that the most relevant variables for predicting the Systolic Blood Pressure were: Age, Body Mass Index, Maximum Inflation Levels, Alcohol, Magnesium and 100 cigarettes smoking. And their relation with blood pressure can be interpreted as exponentiation of corresponding estimate multicity increases/decrease blood pressure level when x is increased by 1 unitand other variables held constant. To specify it, the estimate of coefficient body mass index is 0.0013687 it’s exponential is 1.00137 so every unit increase in body mass index when all the other variables are the same multiples the blood pressure level by 1.00137. Summing up, individuals should be aware of their BMI, Alcohol, Magnesium, Smoking habit and age to not suffer from systolic blood pressure.

Reference

https://community.rstudio.com https://sscc.wisc.edu https://online.stat.psu.edu https://support.minitab.com https://www.researchgate.net https://www.statology.org https://stats.stackexchange.com https://stackoverflow.com

Appendix

The ridge and lasso regression models intentionally left away from the study, their performance could be better than our proposal model.