STATISTICAL ANALYSIS OF SYSTOLIC BLOOD PRESSURE

A FINAL PROJECT REPORT SUBMITTED
IN FULFILMENT OF THE REQUIREMENTS FOR COURSE STAT 364 –
LINEAR MODELS II
DEPARTMENT OF STATISTICS OF
METU

BY

DİLAY GÜMÜŞ 2361301
İLAYDA YILMAZ 2361657
OKAN ÖZHAYAT 2361459

June 2022

TABLE OF CONTENT

1. Introduction

Blood pressure is vital. Blood pressure is affected by many environmental and genetic factors. It is necessary to examine these factors in order to reach accurate blood pressure values. In this study, more than 4000 participants were examined. During the study process, systolic blood pressure and some factors that may affect it were analyzed with statistical tests, regression models, and graphics. To examine the data in detail, descriptive statistics are used, and some graphs are reviewed. Also, for further analysis, the gamma regression model, wilcoxon rank sum test, and hypothesis tests are used. Since the response variable of the data did not provide normality despite all transformation attempts, it was decided to apply the Generalized Linear Model.

2. Literature Review

Blood pressure is the force of circulating blood on the walls of the vascular system. The heart pushes blood into blood arteries, which transport it throughout the body. High blood pressure can be fatal. As a result, people should be aware of the elements that might affect blood pressure. Lower socioeconomic status (SES) was related with higher mean BPs in virtually all research in industrialized nations, according to one study. Women had a greater and more stable inverse gradient than males. The extent of the link varied but was typically fairly minor, with age mean scores of systolic BP variations between the lower and upper SES groups of only 2–3 mm Hg (Colhoun et al.,1998). At Northern, people's blood pressure of average systolic (SBP) and diastolic (DBP) were significantly greater than Southern.SBP and DBP were substantially and separately linked to age, BMI, heart rate, antihypertensive medication usage, sera triglyceride level, alcoholic beverage consumption, and having a smoke (Huan et al.,1994).

3. Data

The data includes 4884 participants and their some test results. Gender, and pulse-type are used as categorical variables. In the graphics below, the counts of each of these variables based on systolic blood pressure used are given.

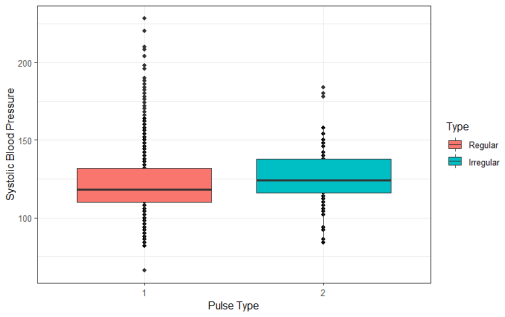

Figure 1

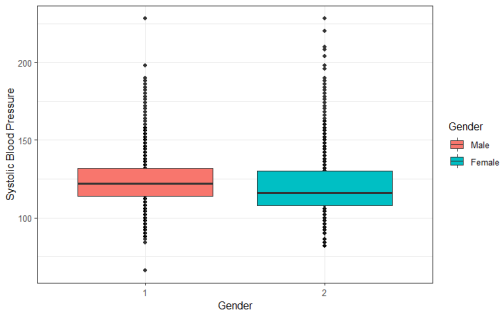

Figure 2

Systolic blood pressure, age in years at screening, participant's body mass index, and alcohol level, maximum inflation levels are used as numeric variables. In the graphics below, the counts of each of these variables based on systolic blood pressure used are given.

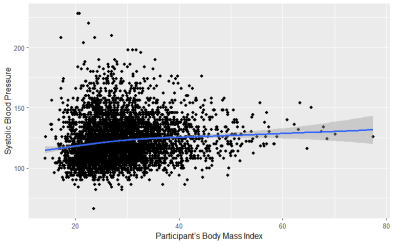

Figure 3

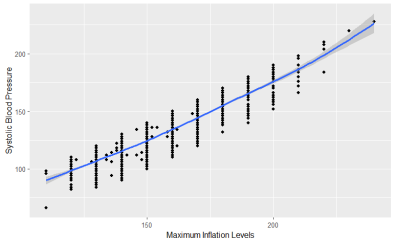

Figure 4

4. Methods

Shapiro Wilk Test: It is used to test whether the distribution of a random sample is normal.

Gamma Regression Model: It is used for a regression model with a skewed dependent which has a gamma distribution.

Wilcoxon Rank Sum Test: It is a non-parametric test to find out the two samples are derivative from the same population.

Deviance Goodness of Fit Test: It is used to test the goodness of fit of the model.

Plots: Normal QQ, Scale-Location plots are used for checking the adequacy of the model. Also, the correlation matrix, scatter plot and box-plots are used.

5. Statistical Results

5.1 Which variables are most related with systolic blood pressure?

Model	Coefficients	Point Estimate (se)	P value	Deviance Goodness of Fit (df)	Train Set RMSE	Test Set RMSE
Gamma model with a log-link	RIDAGEYR	2.155e-04 (6.664e-05)	0.00123	0.004141762 (3660)	118.6261	118.7523
	BMXBMI	1.434e-03 (1.532e-04)	< 2e-16			
	BPXML1	7.202e-03 (7.639e-05)	< 2e-16			
	DR1TALCO	1.089e-04 (3.932e-05)	0.00563			
Robust Model	RIDAGEYR	-3.002e-06 (5.906e-07)	3.71e-07		123.3821	123.508
	BMXBMI	-1.437e-05 (1.323e-06)	< 2e-16			
	BPXML1	-5.373e-05 (5.879e-07)	< 2e-16			
	DR1TALCO	-1.098e-06 (3.328e-07)	0.000967			
Inverse Gamma Model	RIDAGEYR	-1.054e-05 (2.959e-06)	0.000371	0.000184753 (3660)	123.1843	123.3105
	BMXBMI	-6.625e-05 (6.767e-06)	< 2e-16			
	BPXML1	-3.043e-04 (3.310e-06)	< 2e-16			
	DR1TALCO	-5.174e-06 (1.730e-06)	0.002797			
Gamma Model Identity	RIDAGEYR	2.026e-04 (6.753e-05)	0.00271	0.0001831269 (3660)	118.6281	118.7543
	BMXBMI	1.430e-03 (1.560e-04)	<2e-16			
	BPXML1	7.223e-03 (7.919e-05)	<2e-16			
	DR1TALCO	1.099e-04 (4.020e-05)	0.00628			

Figure 5

The Figure 5 illustrates the estimated regression coefficients and p-values of the model which was obtained by Gamma Regression. The stepwise regression and LRT variable selection techniques were used to determine the best subset regression model to predict the systolic blood pressure. Both methods provided the same model. From the figure 5, it can be seen that the variables Age in years at screening (RIDAGEYR), Body Mass Index (BMXBMI), Maximum inflation levels (BPXML1), Alcohol (DR1TALCO) are the most related variables to predict the systolic blood pressure.

5.2 Is the average systolic blood pressure significantly lower in regular pulse compared to irregular pulse?

From the Figure 1, it seems that people with an irregular pulse have a higher systolic blood pressure compared to regular pulse. To confirm this, the Wilcoxon-Rank Sum Test was used, and it demonstrated that the difference was significant with a p-value 0,00334. So, people with a regular pulse have a significantly lower average systolic blood pressure than the irregular pulse.

5.3 Is the average systolic blood pressure significantly higher for males than females?

From the figure 2, it was observed that males have higher systolic blood pressure compared to females. From the Wilcoxon Rank-Sum Test, it is obtained that the p-value < 0.05. Therefore, males have significantly higher average systolic blood pressure than females.

Model Adequacy Check

Outliers and influential observations: From the influence measure function, 228 points seem to be influential by cov.r. Also, the function gives the DFFITS value as 36. So, these points may be investigated in detail to see they are highly influential points.

The Constant-Variance Check: By checking the Scale-Location plot, it was observed that there is no problem with constant variance. So, variance seems to stabilized by using Gamma Regression Model.

Normality of The Residuals: From the Normal QQ-Plot, it was observed that deviance residuals have a normality problem. To confirm this, the Shapiro-Wilk test was used. It provided a p-value (2.232e-15) which is less than 0.05. So, it was concluded that deviance residuals are not normally distributed.

The Goodness of Fit Test: The deviance goodness of fit test is used to see whether the model has a lack of fit or not. The model has a lack of fit problem since residual deviance over degrees of freedom is far away from unity.

Multicollinearity Check:

From the correlation matrix, multicollinearity is observed in some variables. However, all VIF values are less than 10. This suggests that the model does not have a multicollinearity problem.

Model Validation

		Test Set	Train Set
Coefficients	Intercept	3.691e+00	3.690e+00
	RIDAGEYR	1.660e-04	2.155e-04
	BMXBMI	1.412e-03	1.434e-03
	BPXML1	7.193e-03	7.202e-03
	DR1TALCO	2.489e-05	1.089e-04
MSE		66.64062	61.17263
RMSE		118.7523	118.6261

Figure 6

From Figure 6, RMSE, MSE values, and estimated coefficients are close to each other in test and train sets. This can suggest that the Gamma model with log link will behave as it was intended.

6. Conclusion

In the first research question, according to both variable selection techniques age in years at screening, body mass index, maximum inflation levels, and alcohol were found to be the most related variables with the systolic blood pressure. Models from different distributions and different link functions were compared with various graphs and tests to decide which model to use in the GLM concept. At the end, it was decided to use the Gamma Regression Model with log link since it provides the lowest root mean square error compared to other models. Also, it behaves better in QQ-plot compared to other models. It was tried to check whether the response came from the gamma distribution through the goodness of fit tests. Many different models were formed, but it was encountered a lack of fit problem in all of them. As a result, the lack of fit problem could not be solved. In the second research question, it is found that average systolic blood pressure of people with regular pulse is significantly lower than irregular pulse by the Wilcoxon-rank sum test since p-value was lower than 0.05. In the third research question, it is concluded that males have significantly higher average systolic blood pressure than females by the Wilcoxon-rank sum test. It is concluded that model is valid since from Figure 6 it is observed that estimated coefficients, MSE and RMSE values are close to each other in train and test sets. It was observed that residuals are not from normal distribution. By influence measure function, 228 points are influential. Also, there is no multicollinearity problem since VIF values are less than 10. The constant variance assumption is satisfied according to the Scale-Location plot. After this research, new factors that may be related to systolic blood pressure can be investigated and analyzed on the new data set to solve the lack of fit problem. By investigating whether the models used in the analysis are successful on the newly collected data set, the best factors for the estimation of systolic blood pressure can be obtained. As a result of the first research question, one of the most related variables to systolic blood pressure is alcohol. In the world, high blood pressure (hypertension) is a frequent disease, and alcohol consumption is extensive. So, this positive relationship between alcohol and systolic blood pressure can be examined in more detail to prevent high blood pressure.

REFERENCES

Colzhounz, H., Hemingway, H. & Poulter N. (1998). Socio-economic status and blood pressure: an overview analysis.https://www.nature.com/articles/1000558

Huang, Z., Wu, X. & Cen, R. (1994). A north-south comparison of blood pressure and factors related to blood pressure in the People's Republic of China.shttps://www.scholars.northwestern.edu/en/publications/a-north-south-comparison-of-blood-pressure-and-factors-related-to.