Prevent Patients Heart Failure from Ejection Fraction

Xinxue Guo
1005025697

1.Introduction:

The originality of the research comes from the fact that heart failure is one of the significant health problems that affect more than 5.7 million American adults and around 26 million worldwide. It is also associated with mortality and huge cost for healthcare (Hypertension, 2017). The main reason that leads the heart failure is that the heart cannot support enough blood to the body. However, the pumping quality directly is related to the ejection fraction (Chicco, 2020). More specifically, the normal ejection fraction should be 50% to 70%. However, heart failure patients, are categorized by the ejection fraction. If the ejection fraction is equal to or greater than 50%, they are preserved (HFpEF). If the ejection fraction is between 41% to 49%, they are borderline (HFbEF). If the ejection fraction is equal to or less than 40%, they are reduced (HFrEF) (Shah, K. S,2017). Thus, ejection fraction is one of the most important reasons leading to heart failure. The purpose of this paper is to figure out the relevant factors that may affect the ejection fraction. After that people can prevent heart failure indirectly by keeping their body features of these factors on the balance level. In addition, our exploring, based on the patients' current body features and quantifying symptoms, efficiently helps the doctors to predict and estimate the patient's illness that patients receive timely medical treatment.

The data we used records the 299 heart failure patients' clinical features from 13 different aspects in 2015. Our paper will follow the typography. The research method will be shown in the method section, the paper result will be discussed in the result section. And the limitation will be analyzed in the discussion section.

2.Method Section:

2.1Variable Selection:

Once we load the data, we will separate the data into two parts. One is for training, and another is for testing. Based on the training data, we will draw the ggplots for the numerical variables. In addition, we will also draw the bar graph for the categorical variables. According to the explanations for each predictor and Chicco (2020) paper, we will have a brief understanding. Then choose the predictors that are related to the ejection fraction to form the original full model.

2.2Model Violations and Diagnostics

After that, we will use condition 1 to check whether the ejection fraction can be shown as a linear regression model. And condition 2 to check whether each predictor is a linear function with others. If the model satisfies two conditions, we will use the model to

draw a residual plot (residuals versus predictor plots; residuals versus fitted values plots) and QQ-plot. Otherwise, we will choose one predictor left and delete the others which have the relation then recheck the conduction. Based on the residual plots, if the points are disordered distribution, then we can know that the assumptions are held, such as linearly of the relationship, uncorrelated errors, constant variance. And for the QQ-plot, once most of the points flow the lines to distribute, we know that the model follows a normal distribution. Otherwise, we need to modify our model by adding or deleting the predictors even doing the transformation to ejection fraction or the predictors then check the two conditions again. The outliers, leverage points, and influential points need to be checked. If there are outliers, leverage points, and influential points, then we need to identify whether they are valid. If so, we must change the model to maintain them. Otherwise, delete those points by modifying the model then recheck the two conditions. To identify and avoid severe multicollinearity, we will check the VIF for the full model, the predictors whose VIF is greater than 5, will be removed from the model.

After all these steps, we can try to figure out the reduced model. We prefer to use the AIC stepwise variable selection method to get the reduced model randomly. Then, we will check the two conditions, and draw the residual plots, QQ-plots. And check the outliers, leverage points, influential points. If all the checking is satisfied, we will use the VIF to avoid severe multicollinearity.

Then we will use the partial F-test to check whether the reduced model is better than the full model by checking whether the p-value is greater than 0.05. We also will compare the AIC, BIC for both models, the smaller number for AIC, BIC the better model we get. Then compare the adjusted R square, the large number we calculate, the better model we get. If AIC and BIC for reducing model are less than full model and adjust R square for reducing model is greater than full model, we can choose the reduce model as the final model otherwise, we can use full model as the final model.

2.3 Model Validation

To verify the accuracy of the final model, we will use the same predictors but test data to create the checking model. For this model, we also do similar jobs like before, check two conditions, and draw the residual plots with QQ-plots to check whether the model satisfies the assumption. After that check the outliers, leverage points, and influential points. Then check the VIF for all the predictors to avoid multicollinearity. After that, we will compare the summary table between the final model in training data and the final model in testing data. If they have a similar estimate and p-value, we can conclude that the model we find is corresponding. Otherwise, the final model maybe not be perfect.

3.Result Section:

3.1 Description of Data:

Table1: Variables

Numerical Variable:	ejection_fraction	platelets	serum_creatinine	creatinine_phosphokinase	age
Min.	14.00	25100	0.50	23.0	40.00
1st Qu.	30.00	214000	0.90	116.5	51.00
Median	38.00	263358	1.10	248.0	60.00
Mean	37.96	263643	1.42	574.4	60.86
3rd Qu.	45.00	305000	1.45	582.0	70.00
Max.	80.00	742000	9.00	7861.0	95.00
VIF:		1.054281	1.041615	1.090146	1.078746
Categorical Variable:	anaemia	high_blood_pressure	diabetes	sex	smoking
Yes / Male	101	84	97	152	76
No / Female	138	155	142	87	163
VIF:	1.044059	1.037794	1.041787	1.350802	1.299945

Table1 shows the numerical variables and categorical variable information. From the numerical variable, we find that most patients have a lower ejection fraction which is under 40%. We also notice that the age of patients who have heart failure is between 40 to 95 years old. The serum creatinine and creatinine phosphokinase for most patients are quite lower. From the categorical variable, we find that there is more probability to get heart failure than females. And non-smoking people have a lower probability to get heart failure than smoking people. And we try to figure out the relation between ejection fraction and other variables.

3.2 Process of Obtaining Final Model

Table2: Full Model Explore

Full Model: ejection_fraction_hat ~ platelets + serum_creatinine + creatinine_phosphokinase + age + Anaemia + High_blood_pressure + Sex + Smoke + Diabetes (data = train)

Outliers	Leverage points	Influential point
11	9	None

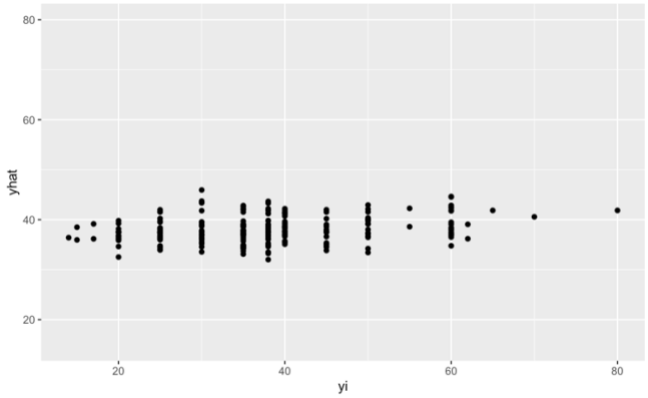
Numerical Variable	platelets	serum_creatinine	creatinine_phosphokinase	age

VIF	1.054281	1.041615	1.090146		1.078746
Categorical Variable	Smoke	High_blood_pressure	Anaemia	Sex	Diabetes
VIF	1.299945	1.037794	1.044059	1.350802	1.041787

From table2, for the full model, condition 1 and condition 2 are satisfied. And the residual plots are disordered distribution. For the qq-plot, the points on the graph always follow the distribution line. We also checked the outliers and leverage points, VIF.

And we use the AIC stepwise variable selection method to figure out the reduce model.

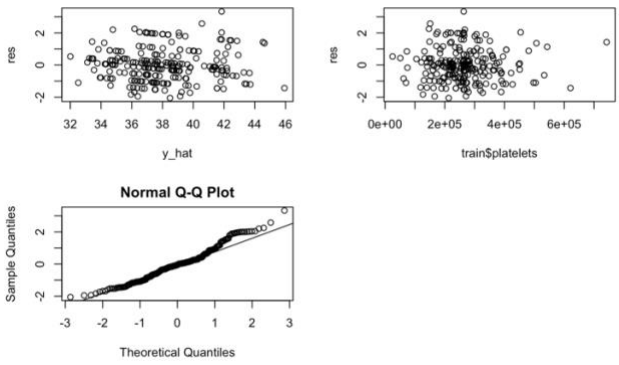
Reduce Model: ejection_fraction_hat ~ platelets + Sex + Diabetes (data = train)



Graph1: Condiction1 for Reduce Model

From the graph1, we conclude that the function of this model is the linear combination of the predictors which satisfy the condition 1.

Since there is only one numerical variable, condiction2 is satisfied generally.



Graph2: Residual Plot & QQ-Plot for Reduce Model

Based on graph2, residual plots, we find that there is no pattern in the graph. This means that this model holds the three assumptions, which are linearity of relationship, uncorrelated error, and constant variance. And based on the QQ-plot, we also find that almost all the points follow the line, and it means that this model holds normality.

Table3: Reduce Model Explore

Reduce Model: ejection_fraction_hat ~ platelets + Sex + Diabetes (data = train)

Outliers	Levarage Points	Influential point
10	11	None

Variable:	platelets	Sex	Diabetes
VIF	1.030301	1.055004	1.033132
Estimate	1.142e-05	-4.335e+00	-2.782e+00

Anova (P-value):	0.9263

	AIC	BIC	adjusted R square
full_model	1863.589	1901.83	0.02098507
reduce_model	1853.583	1870.965	0.03798936

From Table3, we notice that there are fewer outliers and leverage points that can be ignored. And for the VIF of all the predictors are small which means the possible multicollinearity in the predictors is quite small.

After checking the reduced model, we will use ANOVA to check which model will be better. Since the p-value is 0.9263, we can conclude that the reduced model is better than the full model. We also compare the AIC BIC and adjected R square. Since both the AIC and BIC of the full model are greater than the reduced model and adjusted R square for the full model is smaller than the reduced model, it will lead to the same conclusion as before, which the reduced model is better.

Final Model: ejection_fraction_hat ~ 0.00001142 platelets -4.335 Sex -2.782 diabetes

3.3 Goodness of Final Model:

To validate the model, we need to use the same predictors but test data to check. Then we get the test data. Which is: ejection_fraction_hat ~ -000001.126 platelets -0.5832 Sex + 6.854 Diabetes and we will put the relevent graph and table on the appendix.

For the check model, we need to check condition 1 and condition 2. And based on the graph we get; we conclude a linear combination of the predictors. We also check the residual plot. They all have no pattern, which means this model satisfies the linearity of relationship, uncorrelated error, and constant variance assumption. And for qq-plot, we conclude that this model satisfies the normality. As usual, we checked the outlier, leverage points, and VIF. They all have less effect on the model. Then we will compare the reduced model with the check model. We find that the difference between them is slightly small enough. Which means the reduce model is valid and reduce model will be the final model for the data. Thus, the final model will be:

ejection_fraction_hat ~ 0.00001142 platelets -4.335 Sex -2.782 Diabetes

4.Discussion Section:

4.1 Final Model Interpretation and Importance

This paper we try to figure out the relevent predictors that may lead the rejection fraction that to prevent the heart ajection. Based on the previous exploring on the data. We conclude that our final model will be:ejection_fraction_hat ~ 0.00001142 platelets -4.335 Sex -2.782 diabetes. Which means with the other predictors constant, with the one unit increasing on platelets, the risk of leading ejection fraction will increase 0.00001142 units. In addition, with the other predictors constant, on average, the difference ejection fraction between the male and female is 4.335. And with the other predictors constant, on average, the difference ejection fraction between the patients who are diabetes and the patients who are not diabetes is 2.782.We find that the sex, platelets and diabetes have the significant relation with the rejection fraction.

However, there are several leverage points and outliers i3n our model. Outliers affect the estimated regression line. The leverage points may affect the estimated regression line.

4.2 Limitation:

The adjust R square for our final model is quite small which means the linear model may not sa perfect as the non-linear model. In the future study, we can learn to use the non-linear model to express this data. In addition, the sample size for our model is a bit small, we can collect more data for exploring.

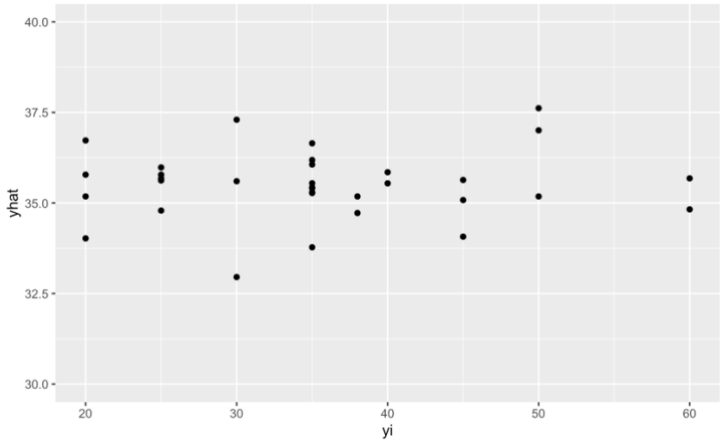
Reference:

Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5

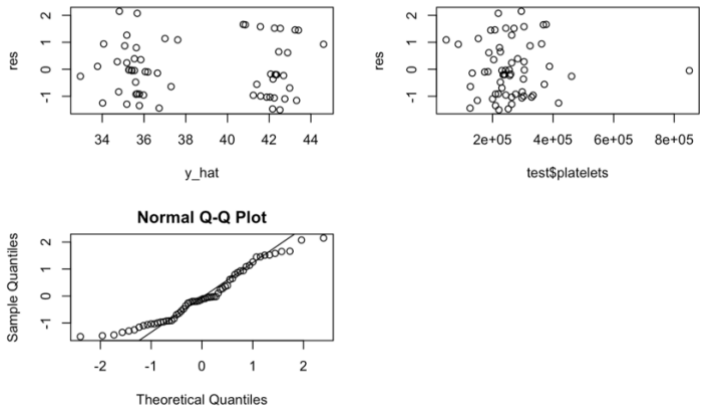
Pandey, A. , Khan, H. , Newman, A. , Lakatta, E. , Forman, D. , Butler, J. & Berry, J. (2017). Arterial Stiffness and Risk of Overall Heart Failure, Heart Failure With Preserved Ejection Fraction, and Heart Failure With Reduced Ejection Fraction. Hypertension, 69 (2), 267-274. doi: 10.1161/HYPERTENSIONAHA.116.08327.

Shah, K. S., Xu, H., Matsouaka, R. A., Bhatt, D. L., Heidenreich, P. A., Hernandez, A. F., Devore, A. D., Yancy, C. W., & Fonarow, G. C. (2017). Heart Failure With Preserved, Borderline, and Reduced Ejection Fraction: 5-Year Outcomes. Journal of the American College of Cardiology, 70(20), 2476–2486. https://doi.org/10.1016/j.jacc.2017.08.074

Appendix:



Graph 3: Condition 1 for Test Model



Graph 4: Residual Plot & QQ-Plot for Check Model

Table4: Test Model Explore

Reduce Model: ejection_fraction_hat ~ platelets + Sex + Diabetes (data = test)

Outliers	Levarage Points	Influential point
2	60	None

Variable:	platelets	Sex	Diabetes
VIF	1.021293	1.014793	1.035164
Estimate	-1.126e-05	-5.832e-01	6.854e+00