Slide 1

Hello everyone. Today I am going to do a heart disease data analysis. My goal is to find a good fit to predict if a person has heart disease or not.

Next Slide

Here is a brief introduction about my project. I got the heart disease dataset from Kaggle. It is a website that contain lots of dataset. To analyze this dataset, I want to use logistic regression for model development. After that, I want to use a bootstrapping concept for model selection and validation. Also, I want to mention that I will introduce some simple function in R for statistical modeling.

Next Slide

Here is an outline for my presentation. First, I will introduce logistic regression theory. Then I will talk about how I clean up the data and will proceed to data exploratory analysis. After that, I will discuss the process of modeling bootstrap method for features selection. Then I will end the presentation with some limitations and future research for my project.

Next Slide

In terms of logistic regression, it has been used in the field of biostatistics. We are interested in studying the association between a response variable and single or multiple explanatory variables. Logistic regression is widely applied on binary, nominal, ordinal and polytomous data. Similar to linear regression, logistic regression also comes from Generalized Linear Models (GLM), which involved linear regression models using the variances to be a function of its predicted value through a link function. The purpose is to look for the best fitting model which can be used to interpret the relationship between a dichotomous response variable and one or more independent variables.

Let’s look at a general form of simple logistic regression model. This is a general form. We have the function of x equal to the exponential function of beta zero and beta 1 times x divided 1 plus the numerator. In this case, the function of x represents the average probability of the occurrence of an event. It can only be range from 0 to 1.

Next Slide

To achieve the properties of a linear regression model, a transformation can be looking like this equation right here. g(x) is a continuous linear function. The outcome variable follows a binomial distribution and it describes the distribution of with 0 mean and pi of x variance. Binary outcome variables are most commonly denoted as 0 or 1, and it means existence or non-existence of a specific trait in our interest.

Next Slide

Estimation of maximum likelihood is necessary when fitting the logistic regression model. Maximum likelihood method can be used to get the parameters that maximize the probability of obtaining the observed data. The logistic likelihood function can be expressed as this function right here. Since the likelihood equations are nonlinear in logistic regression, iterative weighted least squares algorithms are required for their calculations and it can be done in computer.

Next Slide

After obtaining the values, we would like to test how significant these regression coefficients are. Likelihood Ratio test is used to test the goodness-of-fit between our fitted model and the saturated model, which is the model with the same numbers of parameters as data points. We can then observe if our fitted model fits the dataset better. Deviance (D statistic) acts as the sum of squares residual, which accounts for the discrepancy between the data and our estimation. By comparing the D statistic with and without the predictor variable in interest, there is a G statistic acting as the partial F-test in linear regression model. G can be expressed as this function right here. Under the null hypothesis, the statistic G follows a chi-square distribution

distribution with 1 degree of freedom.

Next slide

Confidence interval estimation not only tells us the possible range around the estimate, but it also reveals the stability of the estimate. The endpoints of a CI for logit of the fitted value can be calculated by using Wald-based confidence interval. When the MLE estimators do not meet normal assumptions, It is said that likelihood ratio test could be more suitable for testing whether the regression coefficients and the fitted model is significant or not. To compute the corresponding confident intervals for the coefficient in interest, we can first plot a graph of profile log-likelihood function vs coefficient values by using Venzon and Moolgavkar method that maximizing the log-likelihood. Then, we can calculate specific profile log-likelihood for the fitted value. Eventually, we can draw a horizontal line from that calculated profile log-likelihood value in the plot. The two end points, where the horizontal line intercepting with the curve on x-axis, will be the lower and upper confidence interval.

Next slide

Moving on to the data cleaning process. The data I got from Kaggle is already a good dataset. However, some values are not in the right class as I wanted. So, here is a diagram of the process that I did to clean up the data. First, I added description to the factor variables. This will help me in the long run where I want to see the description of each factor in a variable, I don’t have to look online or on Kaggle. Then, I added reference level for all the factor variables. This will help me understand which factor the reference level is to compare too. After that, I added bins for all the numerous variables. This will help me visualize the variable better. The data ended up having data.frame as its class with 303 rows and 14 columns.

Next slide

Moving on to the data exploratory analysis of this project. I want to explore each variable in the dataset by plotting the variables against the target variable. Also fitting a simple logistic regression would also help in determent the significant of the variable.

Now, I want to start by exploring the variable gender. As you can in the plot on the right, this plot shows gender factor on the x axis, count of gender on the y axis, and average proportion of having heart disease on the second y axis. It is clearly showing that on average, male has a lower change of having heart disease than women. This plot also shows that the proportion of male patient in the data is about double the proportion of female in the data. When looking at the summary statistics on the bottom left, you can see that the variable gender is a significant variable with a really low p value. And the distribution table is just the number in the plot.

Next slide

Let’s look at age. As you can see on the plot, the average proportion of having heart disease tend to decrease as age is increasing until 60 years old. Then it increases as age after 60 years old increases. Rest blood pressure also seem to be very significant with a negative trend.

Next slide

As you can see here, the type of chest pain is very different if a patient has a 0 compare to other type of chest paint. If we assign a new binary factor to chest pain type to have 0 if patient has a 0 in chest pain type and 1 if patient has a 1,2, or 3 in chest pain type, then this would be more significant.

On the right plot, we can also see that if a person has fasting blood sugar, then they would be less likely to have heart disease.

Next slide

To make this presentation short, I combined the rest of the variables in 2 slides. I want to point out that exercise induced angina, ST depression induced by exercise relative to rest, and maximum heart rate seem very significant to contribute to the model.

Next slide

Last but not least, I want to point out that vessels have a really steep negative trend except for when vessels is 4. But this indicate that on average, as the number of vessels increases, the average proportion of having heart disease decreases.

Next slide

Now that we finished exploring the data, I want to introduce the bootstrapping method for our model development process. First, I want to talk about how I will fit the model. I use a function called glm in R with family = binomial to model our logistic regression. Y represent response variable and x represent one or more explanatory variables. The data we use here is from the training set. Now let’s discuss the bootstrap process. I use the for loop to loop over all of my condition. First, I want to split the train and test set into random proportion of 60, 70, 80, 90, and 99 percent. By doing this, I want to see if the model is consistent if there is more or less training data. Then I want to run each of the model 10 times and take the average for more accurate results. Since I am doing random partition, each time I run the model, I would get a different result, so doing it 10 times and take the average would give me more stable results. After that, I specify the explanatory variable to be run. Here I want to loop over all of my variable for a simple regression, then pair the variable, and so on. Last but not least, I want to judge the model by looking at two validation metrics including AIC and area under the curve.

Next slide

Here is a glimpse of the model iteration log. The columns I have here is proportion of random partition, row count for train set and test set, how many times we ran the model, target variable, number of features in the model, independent variables, AIC, and AUC. There are more than 2000 iteration for this log, but I am only showing a few. The row inside the red rectangle is the model that I am selecting for my final model. It has a really high AUC which would be very accurate when predicting. And it has almost the least AIC. In addition, I also check for the random proportion, AUC for this model is about the same when the train set has different number of rows with random proportion is 90, 60 and 70.

Next slide

Here is my final model. It includes gender, maximum heart rate, exercise induced angina, and number of major vessels. On the right corner, there is a summary statistics table showing all variables have very low p values and don’t have 0 in their confident interval. This indicate that all variable in the model are significant.

Next slide

Despite a good fit that we have, there are some limitation and future research ideas I want to highlight. Since the AUC is too high near 90%, I want to be suspicious and think that it’s too good to be true. We might need more data for this dataset to have more accurate results. In the future research, I want to discover more variables that could be an important factor in predicting the proportion of having heart disease. One variable might be what other disease the patient might have. Some disease could lead to have heart disease. Other variable can be the type of work that patient do. Some work required more energy from people which lead the heart to be more active. Last but not least, I want to suggest looking into different types of model methods including xgboost, ridge regression, and random forest to compare with our current model.

Next slide

For more information about this project, here are some useful resources that your can visit.

Next slide

I know I went over the process very quickly. If you have any question regarding the data analysis and R function, please email me. Thank you for listening. And have a great summer professor.