P6110 Final Project:
Estimating the Risk of Heart Disease

Jun Zhai (jz3181)
Department of Biostatistics

Columbia University MAILMAN SCHOOL OF PUBLIC HEALTH

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH
BIOSTATISTICS

ABSTRACT

- To help better diagnose heart diseases, we build two logistic regression model with 'diag' as response variable, and select several test results measures as predictors using stepwise methods.
- The ordinal logistic regression model shows that patients with higher ST depression induced by exercise, serum cholesterol and resting blood pressure are more likely to develop heart disease.
- Both models suggest that the likelihood to be diagnosed with heart disease also has association with patients' age, gender, chest pain type, thallium stress test, number of major vessels colored by fluoroscopy, etc.

OBJECTIVE

- The objective of this project is to predict the likelihood of heart disease by developing statistical models based on the various medical tests' results and related heart disease diagnosis outcomes.
- We aim to improve the accuracy of prediction and find variables which have the most significant contribution to the model.
- We hope to figure out patients with what kind of medical test results are more likely to develop heart disease, and implement treatments and interventions in early phases, eventually prevent or slower the process of disease.

METHOD

- Missing Values in the data set will be located and recoded for the later model building.
- Pearson's correlation is used to calculate the correlation between the continuous variables.
- Scatter plot and density plot are used to explore the distribution of variables and potential collinearity between the continuous variables.
- Response variable 'diag' is treated as binary outcome and ordinal outcome, and respectively fit into a simple logistic regression model and a ordinal logistic regression model.
- Stepwise is used a model selection methods during the model building.
- Statistical tests, such as ANOVA, are implemented to study whether there is significant differences of particular test results between different groups of patients.
- Odds ratio and predictive probability of regression models are utilized to discover the important test results that would significantly contribute to heart disease diagnosis.

RESULT

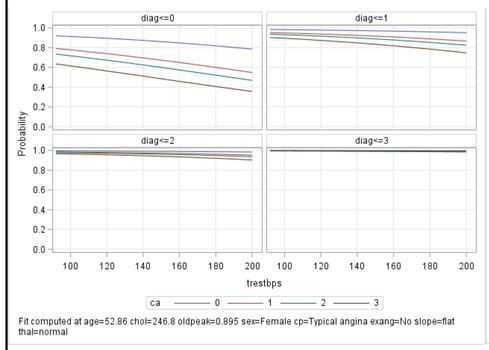


Figure 1. Predicted cumulative probabilities for diag by trestbps and ca

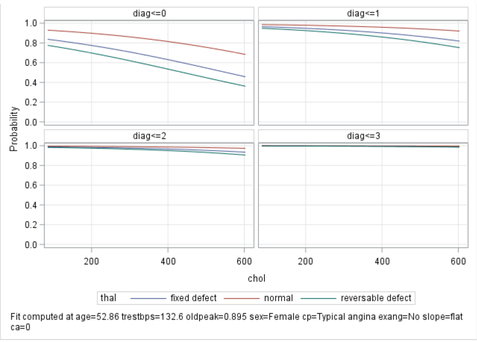


Figure 2. Predicted cumulative probabilities for diag by chol and thal

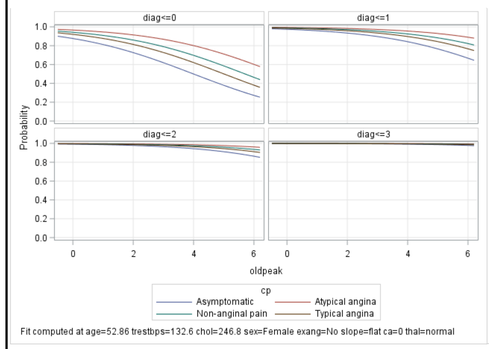


Figure 3. Predicted cumulative probabilities for diag by chol and cp

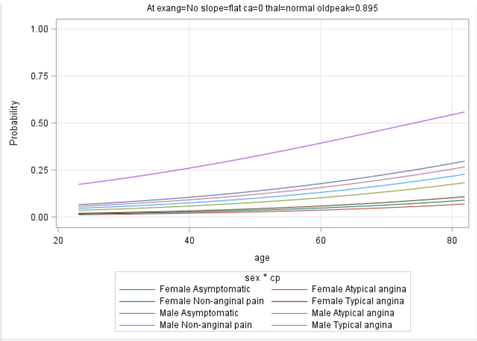
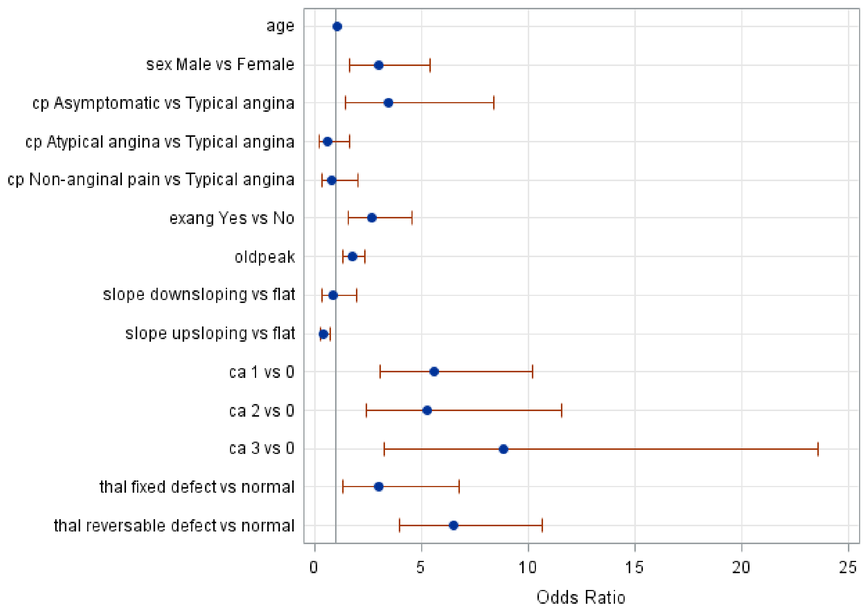


Figure 4. Predicted cumulative probabilities for diag by age and sex*cp

Figure 5. Odds Ratio with 95% CI in logistic regression model



CONCLUSION

- The most important test results in both simple logistic model and ordinal model are thal, ca, cp, oldpeak, which have the biggest significance.
- Only two variables are included in the final ordinal model, while not in the simple logistic model with binary response outcome, they are trestbps and chol, suggesting they may predict ordinal diagnosis results of heart disease.
- When chol, trestbps, oldpeak, age increase, the probability of getting diagnosed as heart disease increase as well. So elder patients and who have higher serum cholesterol or resting blood pressure or ST depression induced by exercise are more likely to develop heart disease.
- Patients who have asymptomatic angina chest pain are most likely to be diagnosed with heart disease, followed by typical angina chest pain, non-anginal pain and atypical pain.
- Male patients have larger likelihood to develop heart disease comparing with female patients.
- Patients with more number of major vessels colored by fluoroscopy have higher risk of having heart disease.
- Having defect in the thallium stress test will increase the risk, and reversable defect has higher odds ratio than fixed defect in the logistic regression model.
- Having experienced exercised induced angina will have higher risk comparing to not having experienced.
- Downsloping and upsloping slope of peak exercise ST segment are not having significantly different odds ratio comparing to flat slope, which may not provide too much sufficient information in the predictive model.
- Other test results in the data set, such as patients' fasting blood sugar value, resting electrocardiographic results, maximum heart rate achieved, are not included in both models, suggesting they may not be too helpful in predicting the risk of having heart disease.

BIBLIOGRAPHY

- Hosmer, D. and Lemeshow, S. (2000). Applied Logistic Regression (Second Edition). New York: John Wiley and Sons, Inc.
- Agresti, A. (2013). Categorical Data Analysis (Third Edition). Hoboken, New Jersey: John Wiley & Sons, Inc.
- Agresti, A. (1996). An Introduction to Categorical Data Analysis. New York: John Wiley & Sons, Inc.
- Marasinghe, M and Koehler, K. (2018). Statistical Data Analysis Using SAS (Second Edition). Heidelberg: Springer, Inc.