Heart Failure Prediction

Mi Zhang

11/27/2021

Abstract

Heart failure prediction has been a prevalent public concern for a long period. In this report, I explored a heart disease-related database and aimed to find the best-fit regression model for heart failure prediction. A Generalized Linear Mixed-Effects Regression Model had been chosen as the best fit model for a binary outcome variable among other model-fitting methods and allowed for building complex mixed effects with random effects. This report found that male patients than female patients, with the higher indicator of type II diabetes, with symptoms of exercise-induced angina, with a flat ECG reading on exercised ST segment, and with asymptomatic chest pain type are at higher risk of heart disease.

Introduction

According to the Centers for Disease Control and Prevention, heart disease is the leading cause of death in the United States and the population of heart disease patients continuously increased in recent years. Prediction of heart failure depends on many health-related factors and accurate prediction of heart failure is almost impossible. However, it would be beneficial if there is a statistical way of approaching the heart failure prediction and cautioning the patient's cardiac health as early as possible. The heart database used in this report is publicly available on Kaggle's website and the database is collected from five different countries. This report will focus on how different health factors are associated with heart disease and use a generalized linear mixed regression model for heart failure prediction.

EDA

First of all, I want to understand what is the major factor of heart disease in the data. Therefore, I decided to make a correlation plot to find the relationship between the variables and the outcome. However, one trade-off here is that ggcorrplot can only plot numeric variables. To do that, I did some researches on those categorical variables which can lead to heart disease and convert them into ordered numeric variable. For example, in the variable "ChestPaintype" the severity for heart disease is ASY > NAP > ATA > ASY, for "RestingECG" is LVH > ST > Normal, and for exercised peak of "ST_Slope" is Flat > Down > Up.

Figure 1 is a correlation plot in which I found that the variable: "ChestPainType", "ExerciseAngina", "ST_Slope", and "Oldpeak" have a higher correlation with heart disease. However, converting categorical variables into ordered numeric variables can create bias. Therefore, I decided to dig deeper into those binary and categorical variables.

Figure 2 displays several bar plots in which I majorly focus on binary and categorical variables. From the plots, I found that male patients, patients with ASY (asymptomatic) chest pain type, patients with blood sugar greater than 120 mg/dl, patients with exercise induced angina, and patients' exercise peak of ST slope is flat in ECG reading have a higher chance of having heart disease. For all other continuous variables, I plot them using box plots for EDA, but there is no significant founding. More EDA can be found in the Appendix.

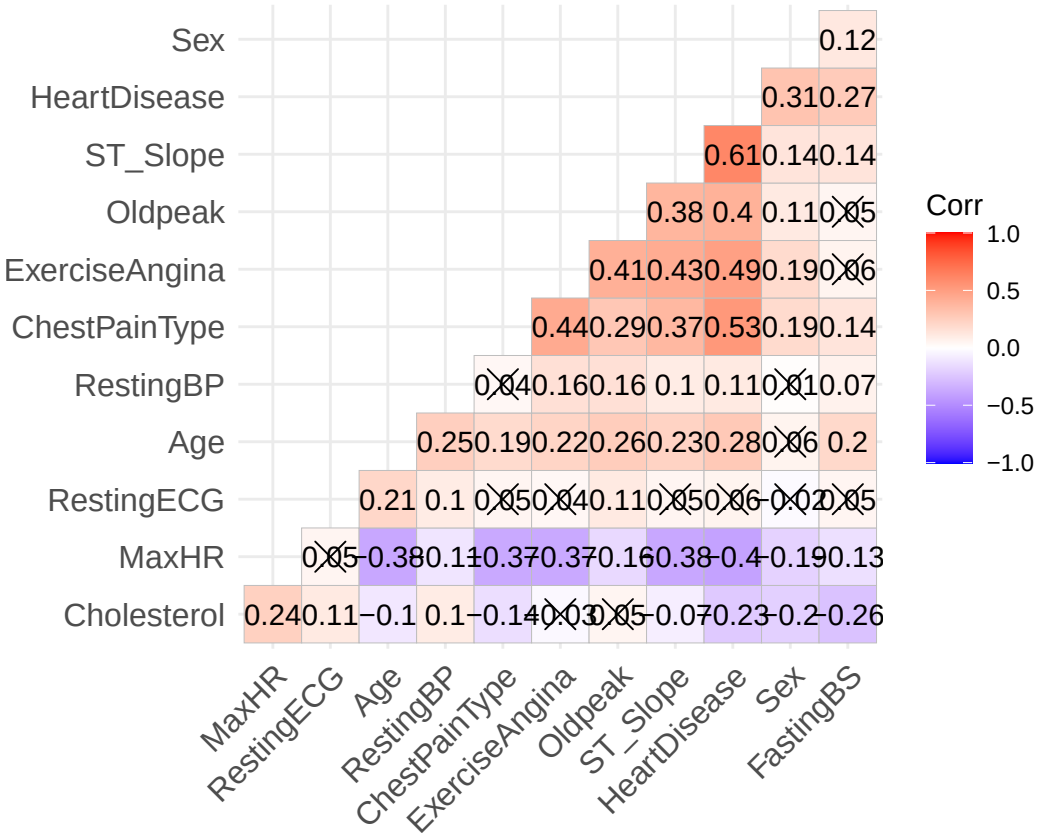


Figure 1: Correlation Plot of Each Variables

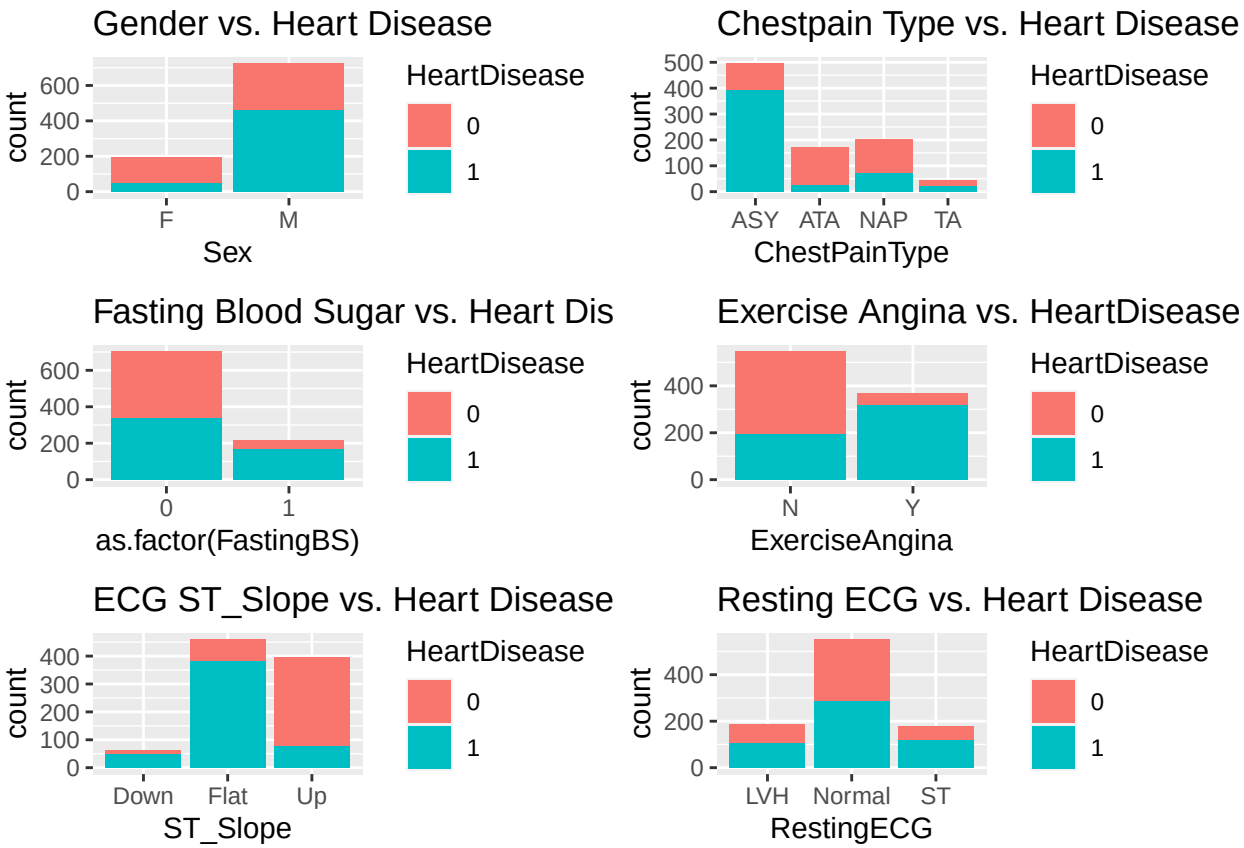


Figure 2: Barplots for Comparison of Different Variables to Heart Disease

Model Fitting and Reult

Based on the EDAs, only a few variables are chosen for the model which includes: "Age", "Sex", "ChestPainType", "ExerciseAngina", "FastingBS", and "ST_Slope". Below I use the "stan_glmer" to estimate a mixed effect of a logistic regression model with variables mentioned above as predictors, add "FastingBS" and "ExerciseAngina" for the random slope and the random intercept is "ChestPainType". The function of the model is shown below:

stan_glmer(data=heart, HeartDisease ~ as.numeric(Age) + + FastingBS +Sex+ ExerciseAngina + ChestPainType + ST_Slope + (1+FastingBS|ChestPainType)+ (0+ExerciseAngina|ChestPainType),family = binomial(link="logit"), refresh=0)

The Generalized linear mixed model is usually hard to interpret, but plots make it easier to understand. Figure 3 shows that male patients than female patients with FastingBS greater than 120 mg/dl (higher indicates type 2 diabetes) and have symptoms of exercise-induced angina have a higher probability of heart disease. A fixed effect table of the model can be found in the Appendix for more information.

The estimates can be interpreted essentially as always. For example, a patient with Fasting blood sugar equal to 1 (greater than 120 mg/dl and therefore diabetic) is associated with 1.2418381 higher expected log odds of having heart disease. Similarly, a patient who has a symptom of exercise-induced angina is expected to have 1.0700240 higher log odds of having heart disease than people who don't have that symptom.

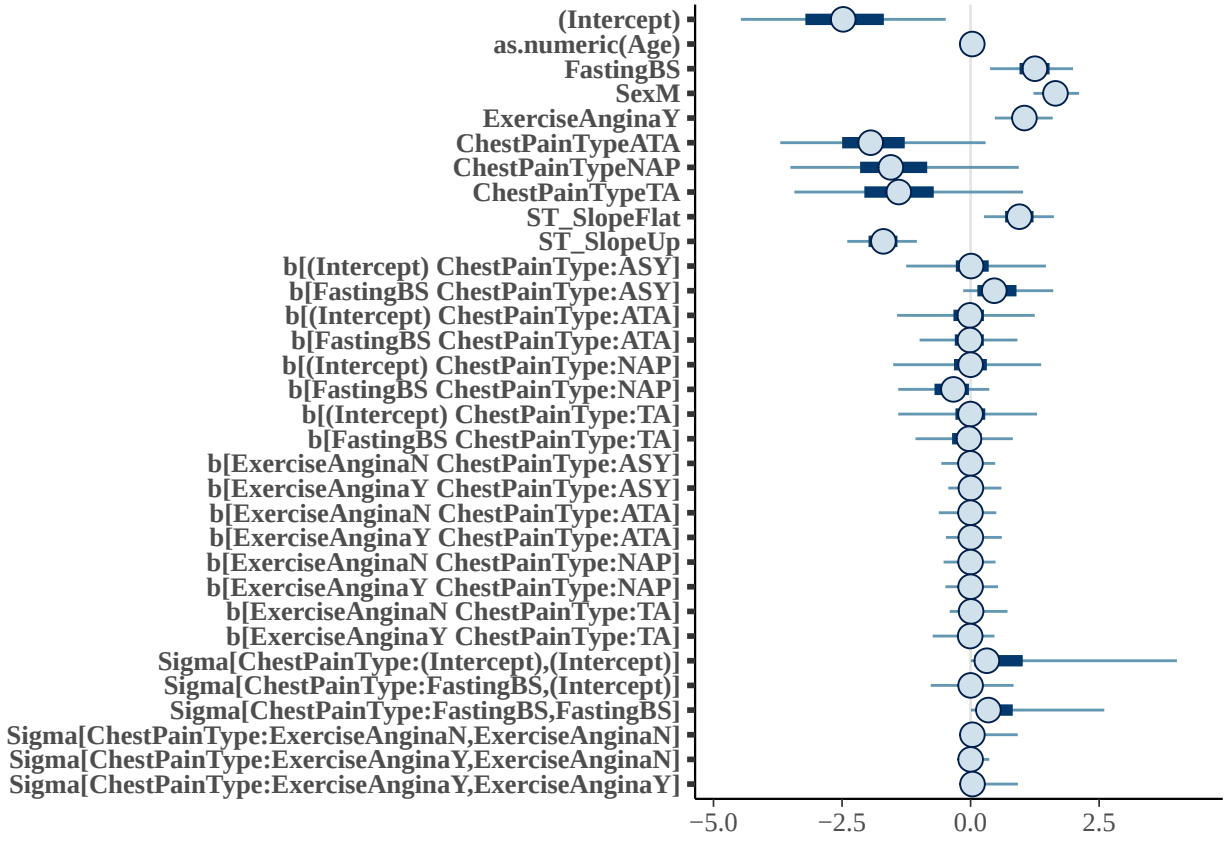


Figure 3: Plot for Generalized Multilevel Model

Model Validation

The validation of the generalized linear mixed model is limited. We can only use the posterior predictive checks plot and binned residual plot for validation. Both posterior predictive checks plot and binned residual

plot can be found in the Appendix. The posterior predictive checks plot displays the comparison distribution of the observed outcome of heart disease patient (y) and simulated outcome of heart disease (y_rep) in which both distributions match well and proves the model fits well.

The binned residual plot shows that a few points lie outside the confidence limits and there is no obvious pattern in the plot which also proves the effectiveness of the model.

Discussion

Based on the result of the Generalized Linear Mixed Regression model, it is no surprise that many predictors have matched the finding of the Exploratory Data Analysis. For example, male patients with diabetes and show symptoms of angina are at higher risk of heart failure. But there are some variables that I previously thought of as important factors for heart disease that fail to prove their importance in the modeling process, such as cholesterol and maximum heart rate. And from the data collecting view in which this database contains three variables collected using ECG screening, we can conclude that ECG is effective for heart abnormality detection. In addition, analysis of medical-related data requires certain knowledge about the field. For example, I assume the elevation or depression on ECG reading of ST-segment as factors for higher risk of heart disease. But GLMER modeling result shows that the flat ST segment has a greater impact on heart disease. Therefore, I went back to do my research and found during exercise, elevated ST-segment is considered normal while flat ST-segment on ECG should bring into caution for heart disease.

Through the entire model fitting process, I was surprised by how mixed effects logistc regression model can build complex mixed effect for the binary outcome, but also found binary outcome variable is limited for analysis in many ways. For example, stan_lmer() can only analyze numeric outcomes while binary outcome variables are factors in general. In addition, binary outcome variables can easily result in zero inflation when using the linear regression model. Matching is useful for the binary outcome variable, but this database contains symptoms rather than treatments which makes it hard to use the matching method. I think the other way to improve this analysis will be using machine learning, I would like to redo this analysis using machine learning in the future when I master the knowledge.

References

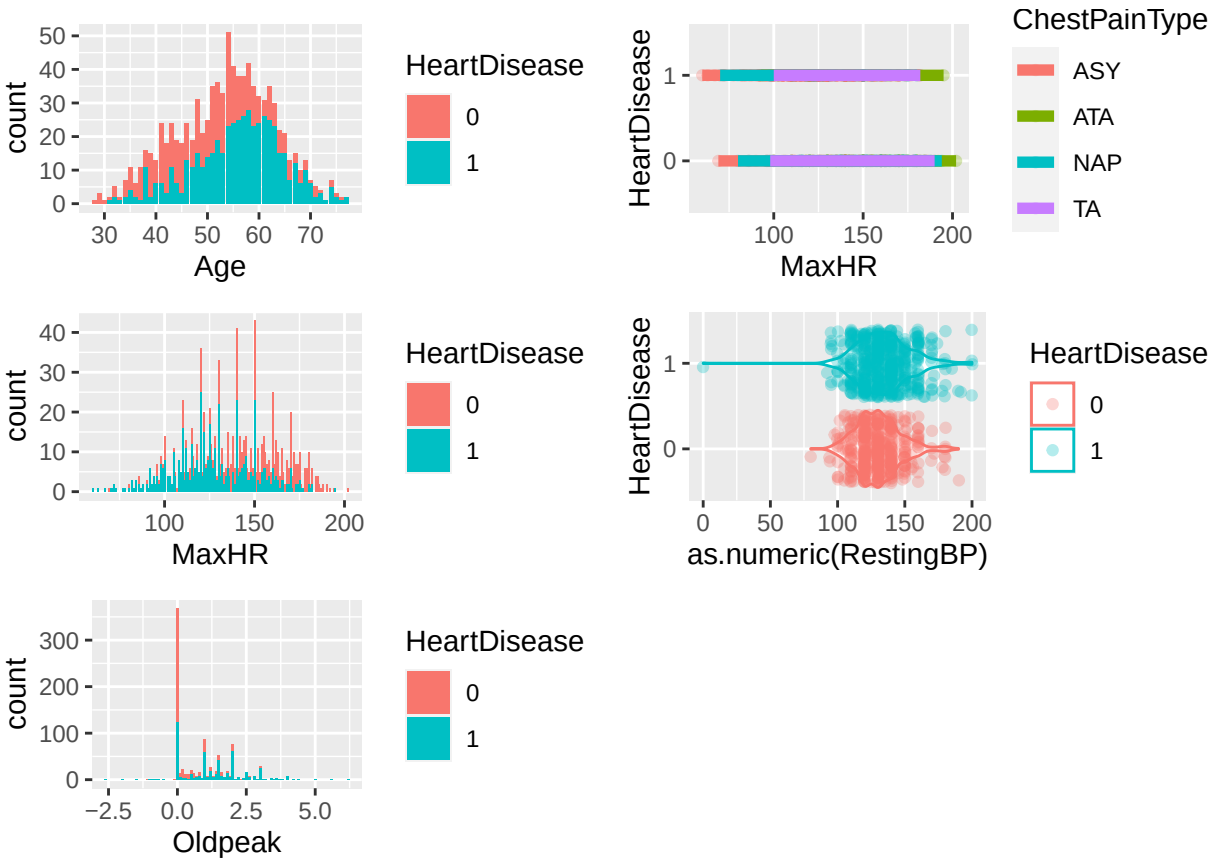
- Noah Greifer, 2021, "MatchIt: Getting Started". https://cran.r-project.org/web/packages/MatchIt/vignettes/MatchIt.html
- Centers for Disease Control and Prevention, "Heart Disease". https://www.cdc.gov/heartdisease/index.htm
- UCLA Statistical Consulting, "MIXED EFFECTS LOGISTIC REGRESSION | R DATA ANALYSIS EXAMPLES". https://stats.idre.ucla.edu/r/dae/mixed-effects-logistic-regression/
- Anthony H. Kashou; Hajira Basit; Ahmad Malik, 2021, "ST Segment". https://www.ncbi.nlm.nih.gov/books/NBK459364/
- Keith McNulty, "Handbook of Regression Modeling in People Analytics: With Examples in R, Python and Julia". https://peopleanalytics-regression-book.org/multinomial-logistic-regression-for-nominal-category-outcomes.html
- onah Gabry and Ben Goodrich, "Estimating Generalized (Non-)Linear Models with Group-Specific Terms with rstanarm". https://mc-stan.org/rstanarm/articles/glmer.html

Appendix

data descripion

Variables	Description
Age	age of the patient[years]
Sex	sex of the patient[M:Male, F:Female]
ChestPainType	chest pain type[TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
RestingBP	resting blood pressure[mm Hg]
Cholesterol	serum cholesterol[mm/dl]
FastingBS	fasting blood sugar[1: if FastingBS > 120 mg/dl, 0: otherwise]
RestingECG	resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes criteria]
MaxHR	maximum heart rate achieved [Numeric value between 60 and 202]
ExciseAngina	exercise-induced angina [Y: Yes, N: No]
Oldpeak	oldpeak = ST [Numeric value measured in depression]
ST_Slope	the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down:downsloping]
HeartDisease	output class [1: heart disease, 0: Normal]

Full EDA



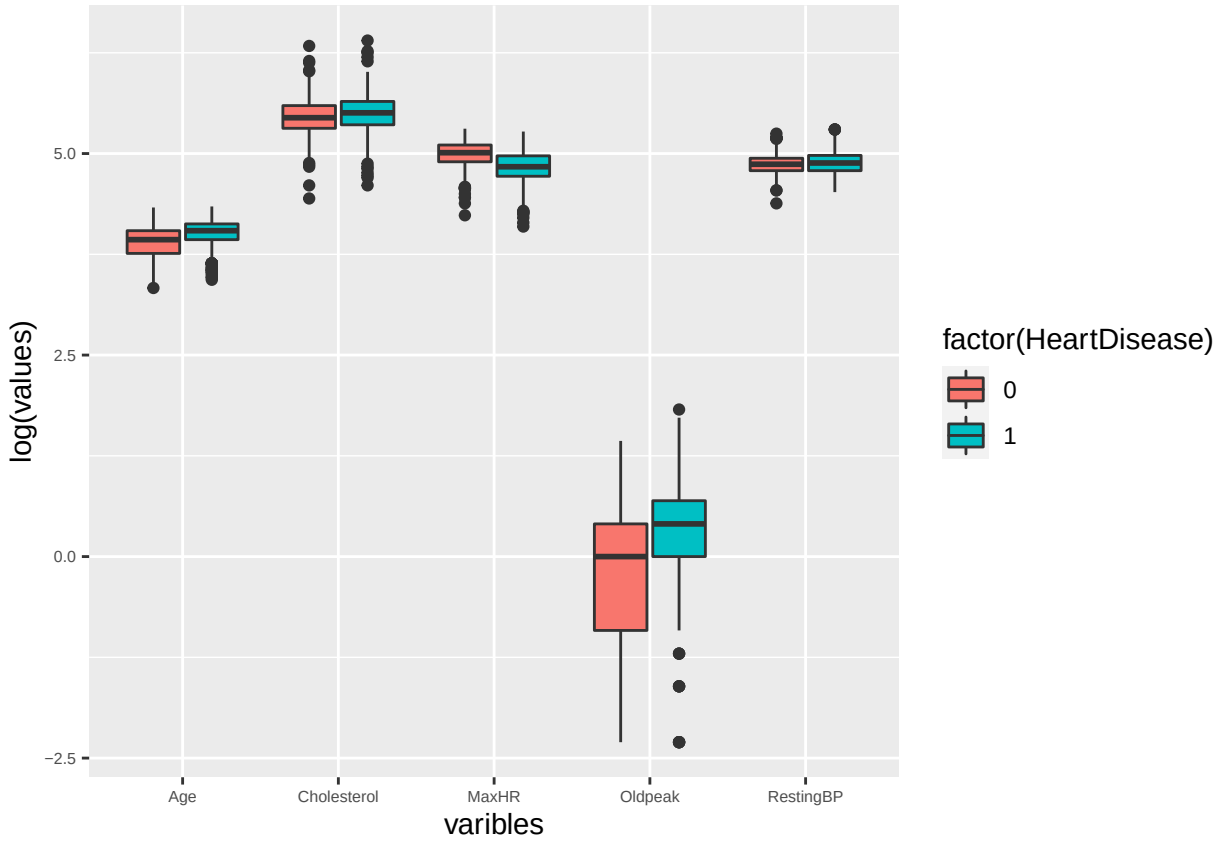


Figure 4: Box Plots for Continous Variables

lmer model

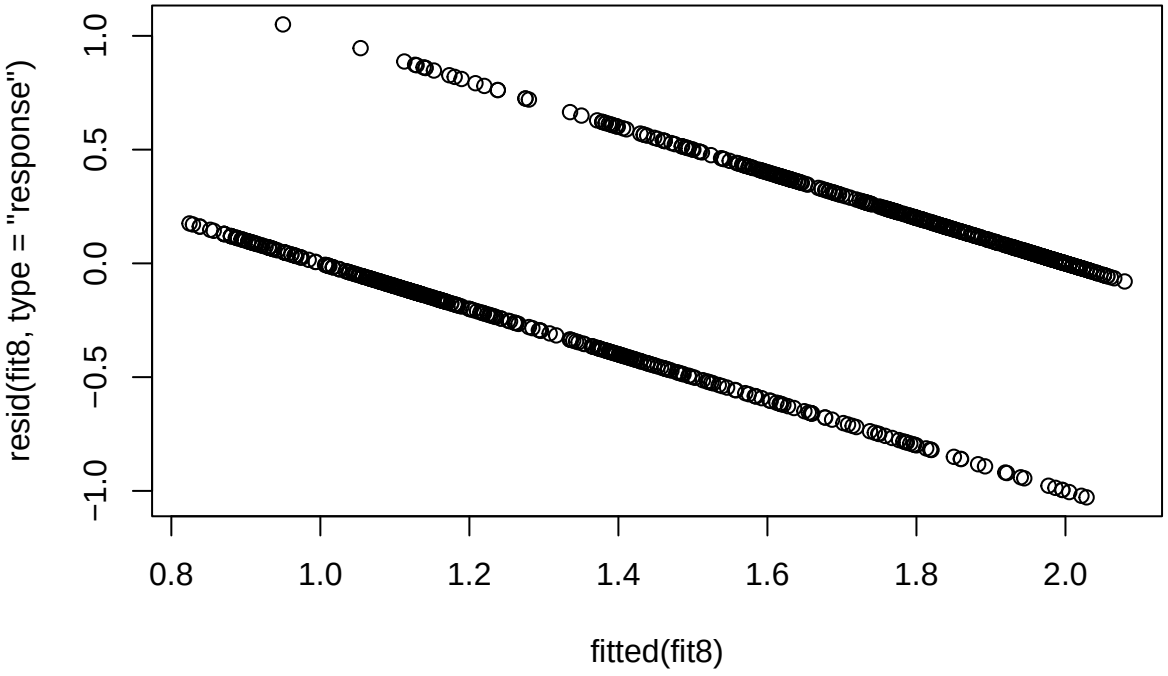


Figure 5: Residual Plot for lmer

more plots for stan_glmer

	x
(Intercept)	-2.4770012
as.numeric(Age)	0.0324299
FastingBS	1.2510264
SexM	1.6505296
ExerciseAnginaY	1.0466536
ChestPainTypeATA	-1.9432997
ChestPainTypeNAP	-1.5518464
ChestPainTypeTA	-1.3967719
ST_SlopeFlat	0.9490179
ST_SlopeUp	-1.6956497

```
## $ChestPainType
```

Matching_it

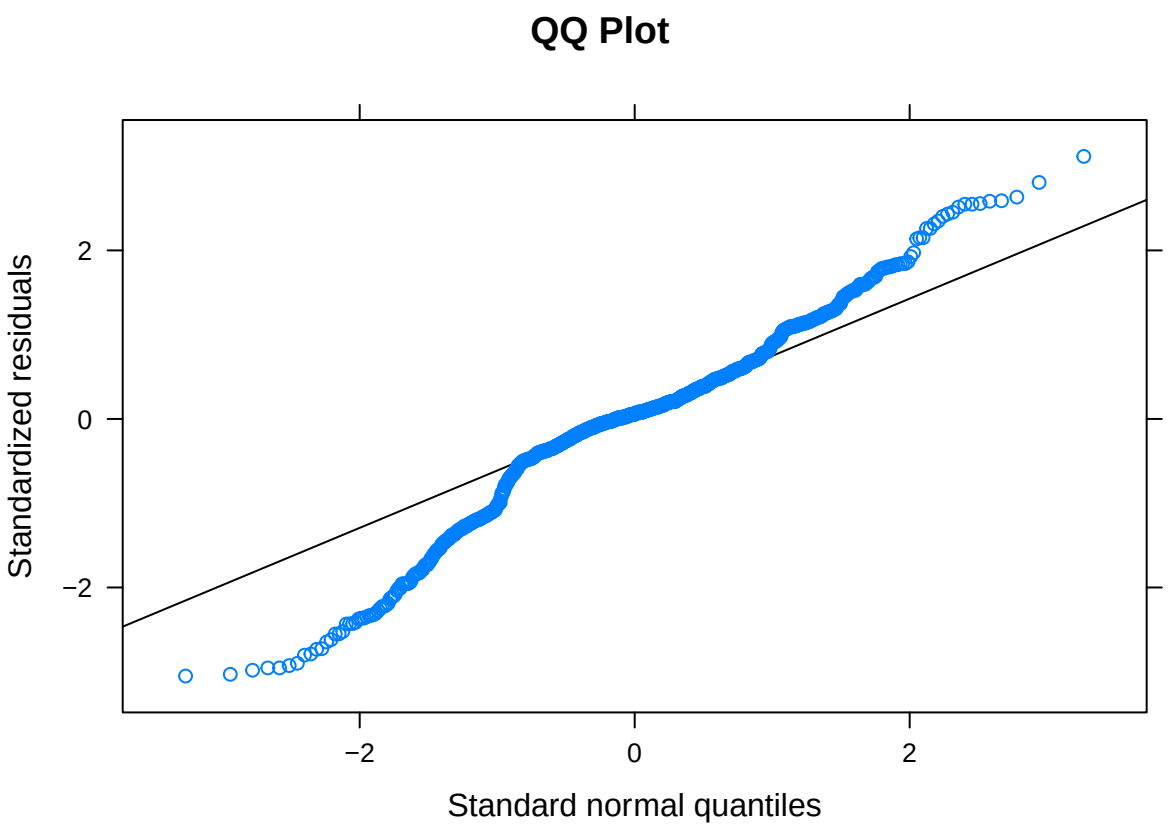


Figure 6: QQ Plot for lmer

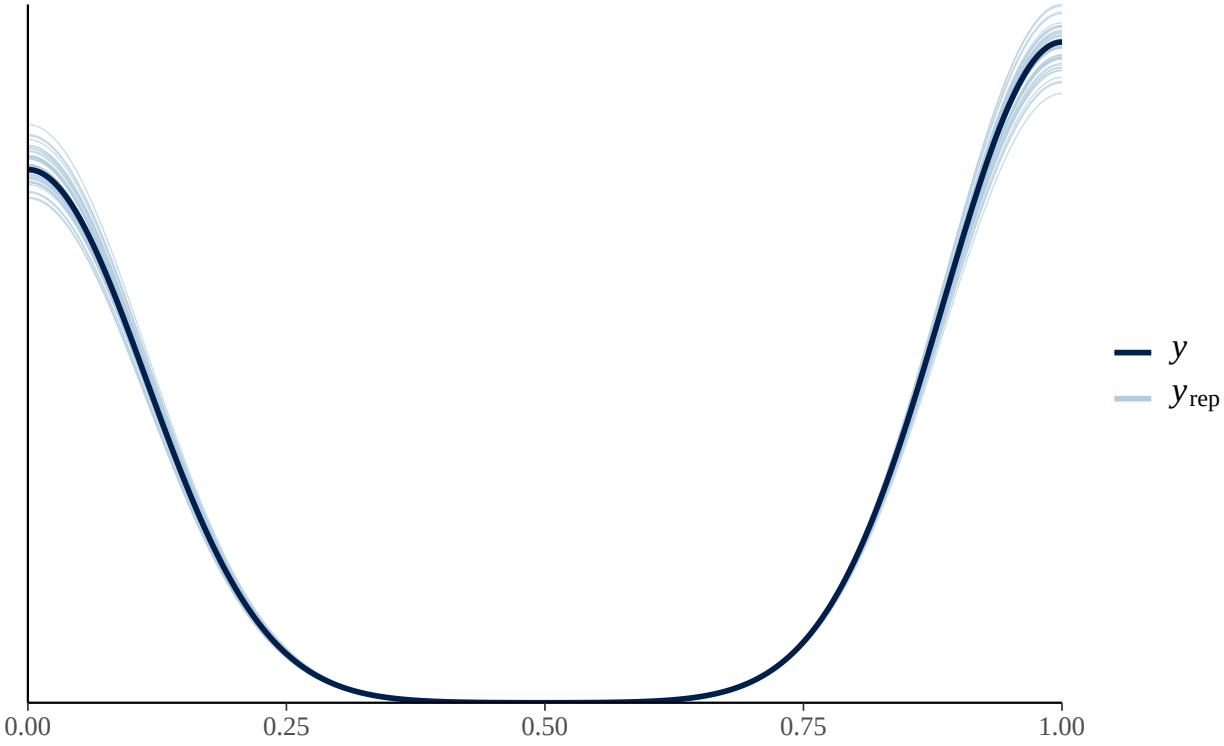


Figure 7: Posterior Predictive Checks Plot

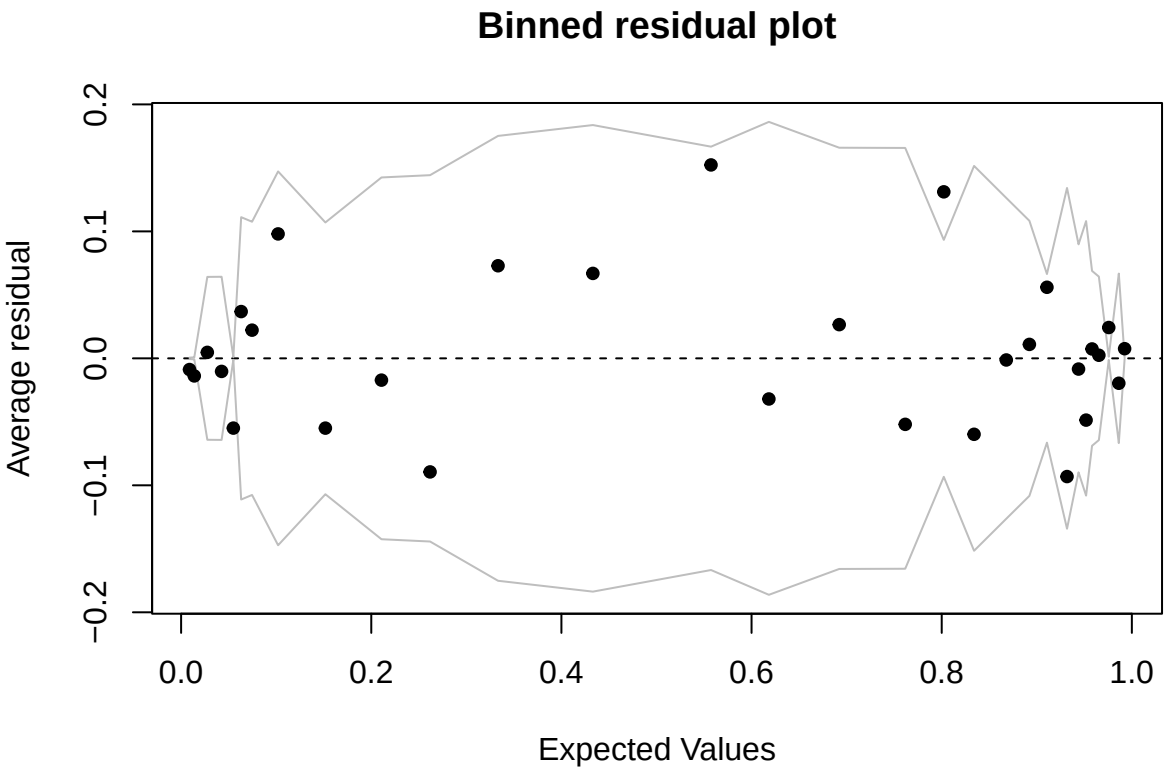


Figure 8: Binned Residual Plot for model

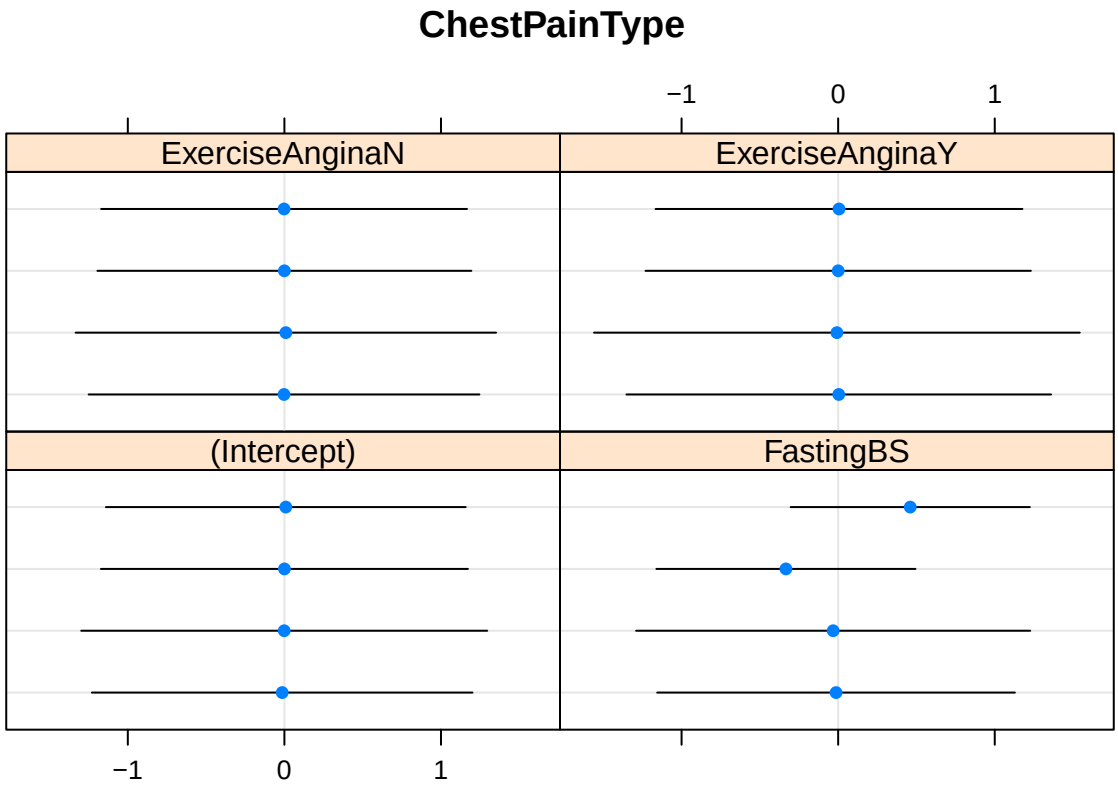


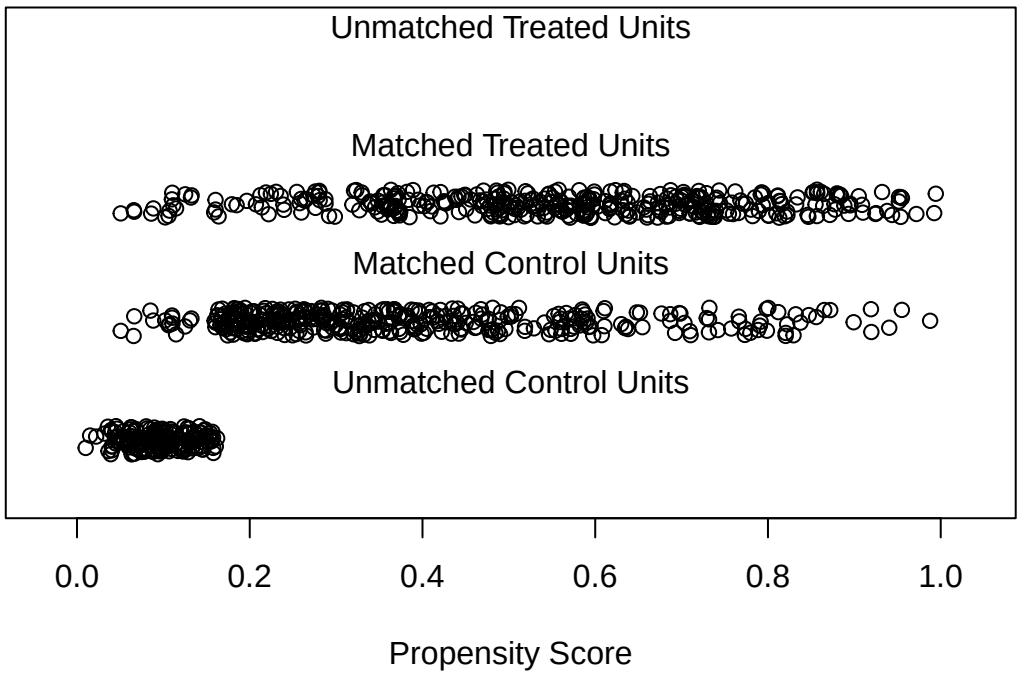
Figure 9: Dot Plot for GLMER

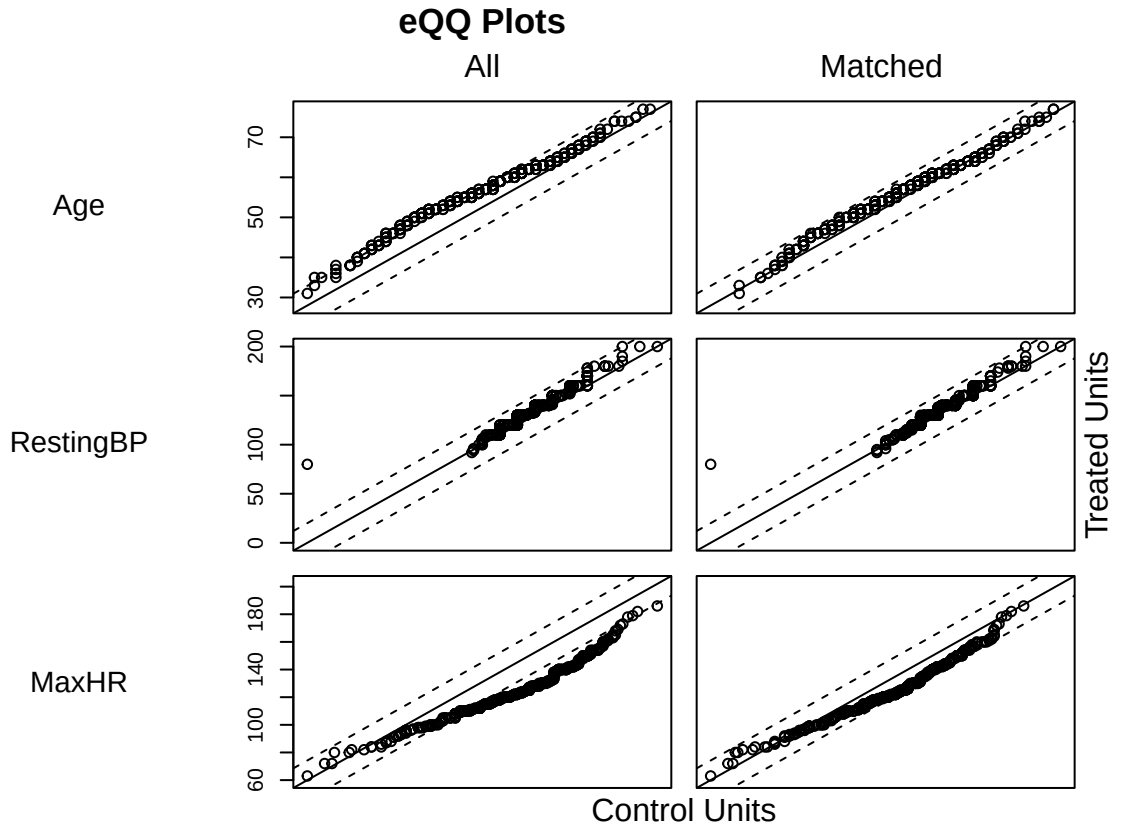
Table 3: Table 2: Sample sizes

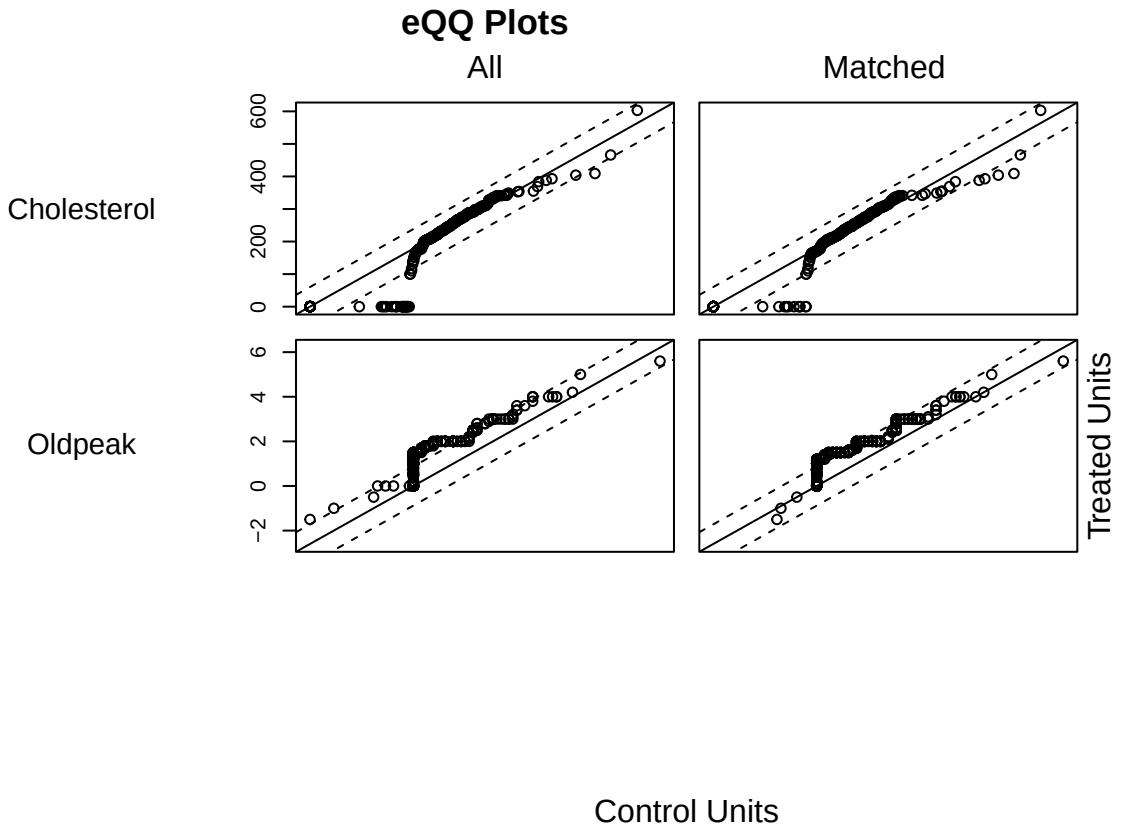
	Control	Treated
All (ESS)	547	371
All	547	371
Matched (ESS)	371	371
Matched	371	371
Unmatched	176	0
Discarded	0	0

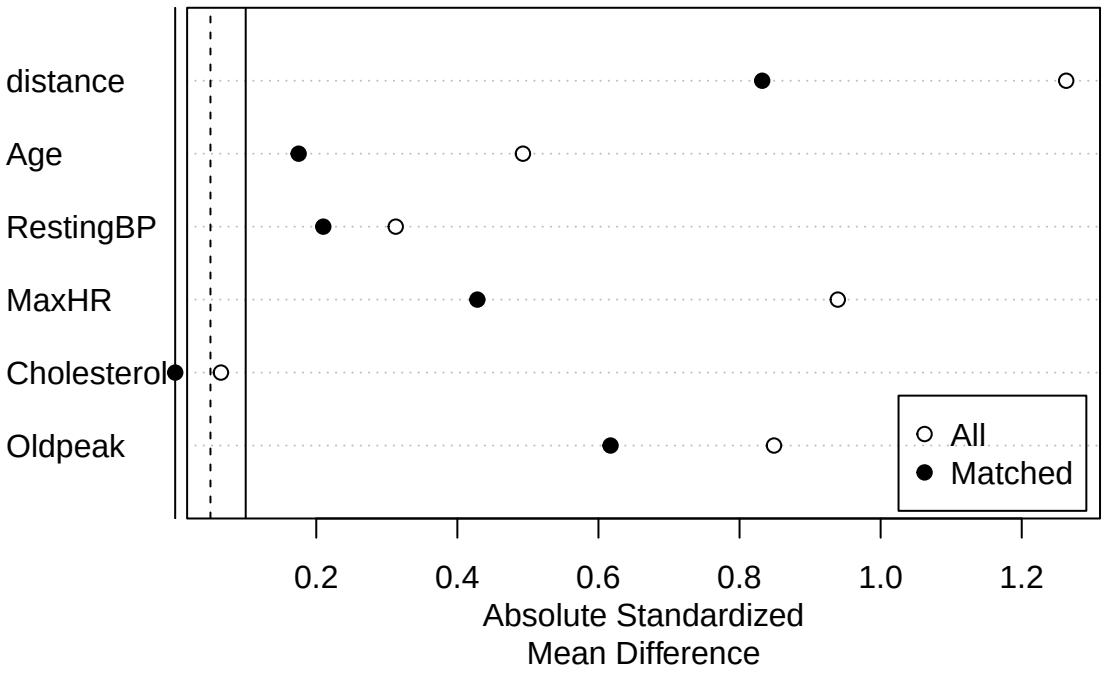
Table 4: Table 3: Summary of balance for matched data

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
distance	0.56	0.38	0.83	1.14	0.19	0.41	0.83
Age	55.98	54.51	0.18	0.87	0.03	0.11	1.16
RestingBP	135.88	131.95	0.21	1.03	0.04	0.13	1.09
MaxHR	125.36	134.12	-0.43	0.77	0.08	0.24	1.01
Cholesterol	194.26	194.26	0.00	1.11	0.04	0.09	1.12
Oldpeak	1.42	0.77	0.62	1.08	0.12	0.37	0.90




eQQ Plots
All
Matched
Age
RestingBP
MaxHR
30
50
70
0
50
100
200
60
100
140
180
Treated Units
Control Units


eQQ Plots
All
Matched
Cholesterol
Oldpeak
0
200
400
600
−2
0
2
4
6
Treated Units
Control Units


distance
Age
RestingBP
MaxHR
Cholesterol
Oldpeak
All
Matched
0.2
0.4
0.6
0.8
1.0
1.2
Absolute Standardized
Mean Difference