MULTIVARIATE LINEAR REGRESSION OF HEART DISEASE ATTRIBUTES TO BLOOD PRESSURE

Anjali Ramesh, Saroop Samra, Harmeena Sandhu, Ashna Sood, Urmi Suresh

University of California, San Diego, Department of Cognitive Science

Background

Healthcare is a vital point of research in order to best help patients with certain conditions. Blood pressure specifically often has no symptoms, and yet if untreated high blood pressure can be a large contributor to more severe health conditions such as a stroke or heart attack **[2]**.

According to the CDC, 1 in 3 U.S. adults are not even aware of their high blood pressure which means their blood pressure is going untreated **[3]**.

Due to blood pressure presenting itself almost invisibly, we wanted to further understand what can best predict the blood pressure on the patient, which is useful information that could help diagnose someone with high blood pressure and get them the care that they need.

Research Question

Using the provided dataset, which attribute(s) can be used to best predict a patient's blood pressure?

Dataset

To answer this question, we are looking at the UCI Statlog (Heart) data set **[1]**.

This dataset consists of 14 attributes with 270 samples. The attributes are the following:

- age
- sex
- chest pain type
- blood pressure
- cholesterol
- blood sugar
- resting ECG
- heart rate
- angina,
- ST depression
- peak exercise ST segment,
- blood vessels
- thal
- **Label:** heart disease (1 or 2)

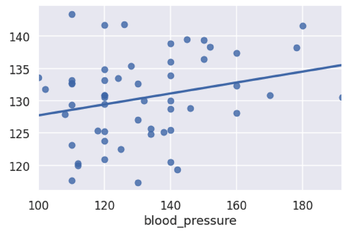
While performing our analysis, we used one hot encoding to turn the nominal variables into usable values.

Methods

We created **four different multivariate linear regression models** with different combinations of variables , which attempt to narrow down the best predictors of blood pressure

Each graph show the relationship betwen the predicted Y value and the test Y values

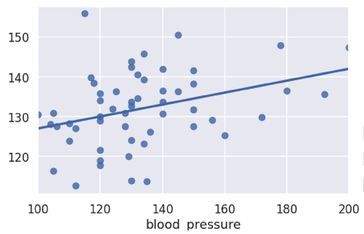
Model 1: all continuous and binary variables, excluding the nominal variables



Regression Equation: Y = (0.5452)age + (-3.7447)gender + (0.0103) cholesterol + (9.0994)blood_sugar + (0.0941)heart_rate + (2.2694)angina + (3.289)ST_depression + (-1.3174)ST_peak_exercise_slope + (-1.531)major_vessels + (2.8935) heart_disease + 80.6436

MAE: 15.680
MSE: 388.671
RMSE: 19.714

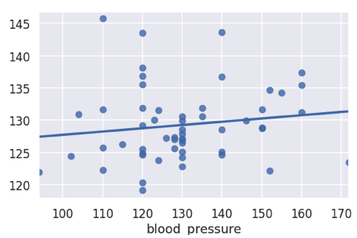
Model 2: all variables, using one-hot encoding



Regression Equation: Y = (0.4467)age + (-2.9617)gender + (0.0479)cholesterol + (5.0634)blood_sugar + (0.1377)heart_rate + (1.0485)angina + (2.7976)ST_depression + (-1.0946)ST_peak_exercise_slope + (-1.693)major_vessels + (3.8217)heart_disease + (6.4162)Chest_Pain_1 + (-3.6774)Chest_Pain_2 + (-1.9342)Chest_Pain_3 + (-0.8046)Chest_Pain_4 + (-9.1334)Electrocardiographic_0 + (15.965)Electrocardiographic_1 + (-6.8316)Electrocardiographic_2 + (-2.3412)Thal_type_3 + (2.2544)Thal_type_6 + (0.0867)Thal_type_7 + 79.541

MAE: 15.505
MSE: 414.345
RMSE: 20.355

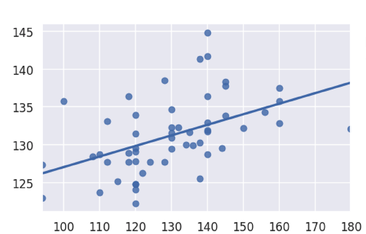
Model 3: 3 most positively and 3 most negatively correlated variables



Regression Equation: Y = (0.414)age + (-2.2348)gender + (0.047)cholesterol + (0.0817)heart rate + (3.3651)ST depression + (9.0969)Chest Pain 1 + (-2.1377)Chest Pain 2 + (-4.8307)Chest Pain 3 + (-2.1285)Chest Pain 4 + 84.7231

MAE: 12.267
MSE: 265.139
RMSE: 16.283

Model 4: 2 most correlated variables



Regression Equation: Y = (0.3638)age + (2.8664)ST depression + 108.7843

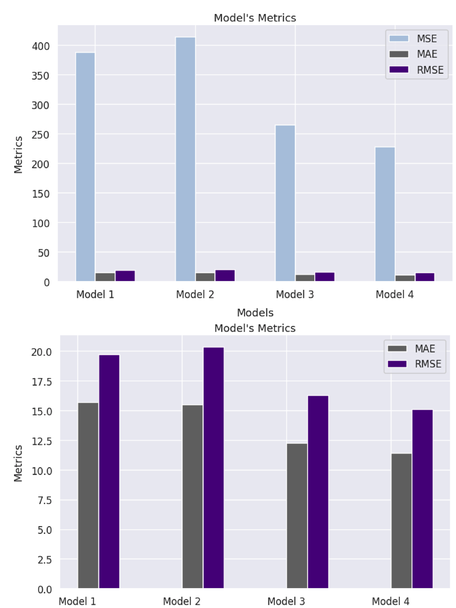
MAE: 11.412
MSE: 228.274
RMSE: 15.108

Results

Based on each model's error metrics, **Model 4 performed the best, followed by Model 3, Model 1, and Model 2.** From the errors we can see that even after adding more attributes to the multivariate Model 2, this model is still outperformed by Model 1. By adding more attributes, the test error did not improve and therefore Model 1 is a better fit for the data than Model 2.

Model 3 is where the attributes began to become more narrowed down, and this model has lower test error than both Model 1 and Model 2.

Out of all the models, Model 4 had the lowest test error, which is what we expected. This model uses only the two most correlated variables, age and ST depression, which is mostly likely the reason why Model 4 has performed the best. Model 4 is the best model for predicting blood pressure with less, but more important attributes selected. **This shows that adding more attributes does not in this case help predict blood pressure better.**



Discussion

Answering our research question, age and ST depression are the best attributes to use to predict a patient's blood pressure.

Our exploration of the raw data allowed us to visualize the apparent relationships between each variable against blood pressure. Ultimately, the correlation matrix allowed us to generate our most accurate model, model 4, which used only two variables that were the most correlated according to the matrix. Exploring the raw data was a necessary step in understanding what we were working with, but the correlation matrix itself was what allowed us to confidently choose certain variables for our models.

Future Plans:

- implement a decision stump to further validate our models
- compared the results we got to other similar heart disease datasets
- find a larger dataset to allow us to be more sure of our results

References

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.
[2]"The Facts About High Blood Pressure." Www.heart.org, 30 Nov. 2017,
[3] "5 Surprising Facts About High Blood Pressure." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 9 Nov. 2020