Nate Evans

Mike Mooney

BMI665: Scripting

11/24/18

HW#6 Write-up

# Overview

The analysis of this data relies is built of the heart.dat file data (<https://sakai.ohsu.edu/access/content/group/BMI-565-665-DL-F18/Week%208/heart.dat>) and concludes in a multiple linear model to predict the ‘old\_peak’ variable, which is a measure of ST depression and a sign of decreased oxygen to the heart muscle (myocardial ischemia), based on three correlated variables: age, max\_hr, resting\_bp.

The analysis of this dataset can be run by navigating to the script directory in terminal and running the command:

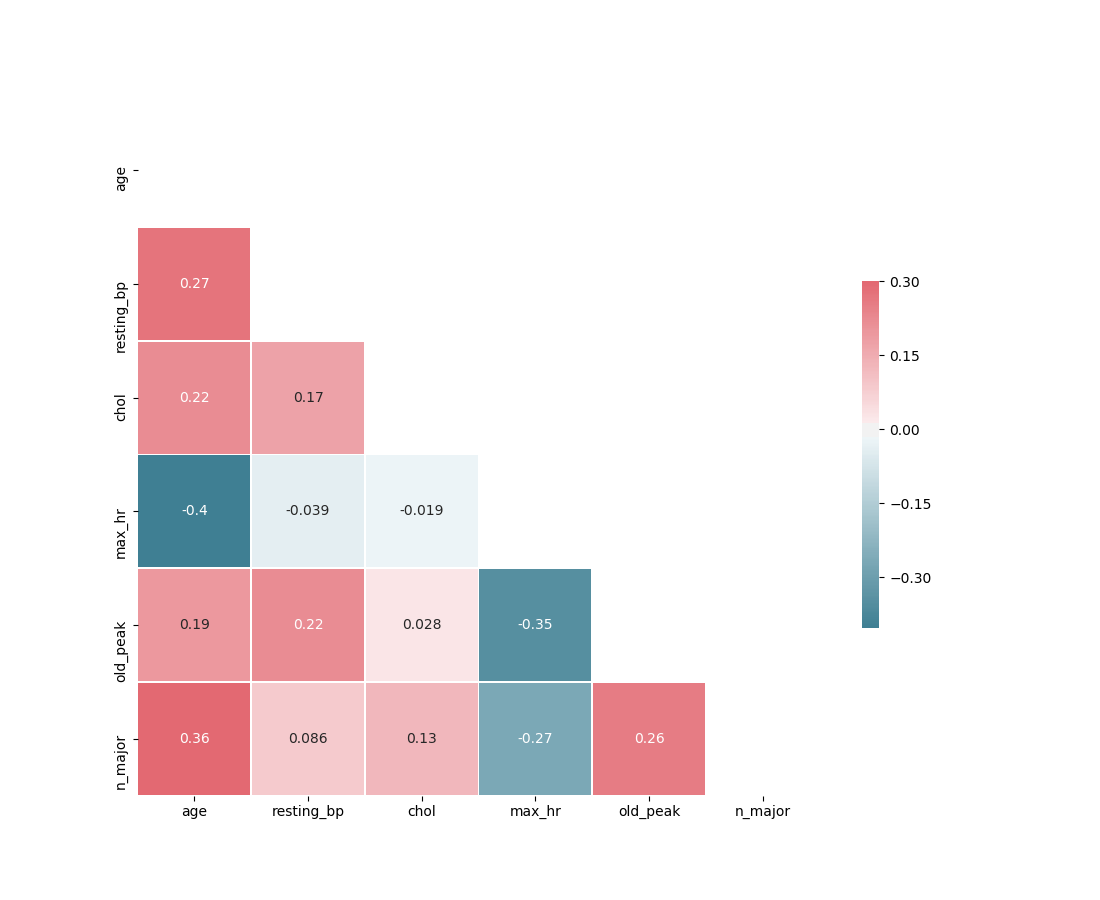
$ python3 evans\_hw6.py

The resulting will be saved in the script directory

./outputs/

# EDA

To begin, we explored the dataset and extracted variables that showed a high correlation (> 0.2) with the dependent variable. Note that, since we ran correlations between 6 variables, the alpha for statistical significance should be adjusted to avoid p-hacking. Alpha2 ~ 0.05 / 6 = 0.0083.

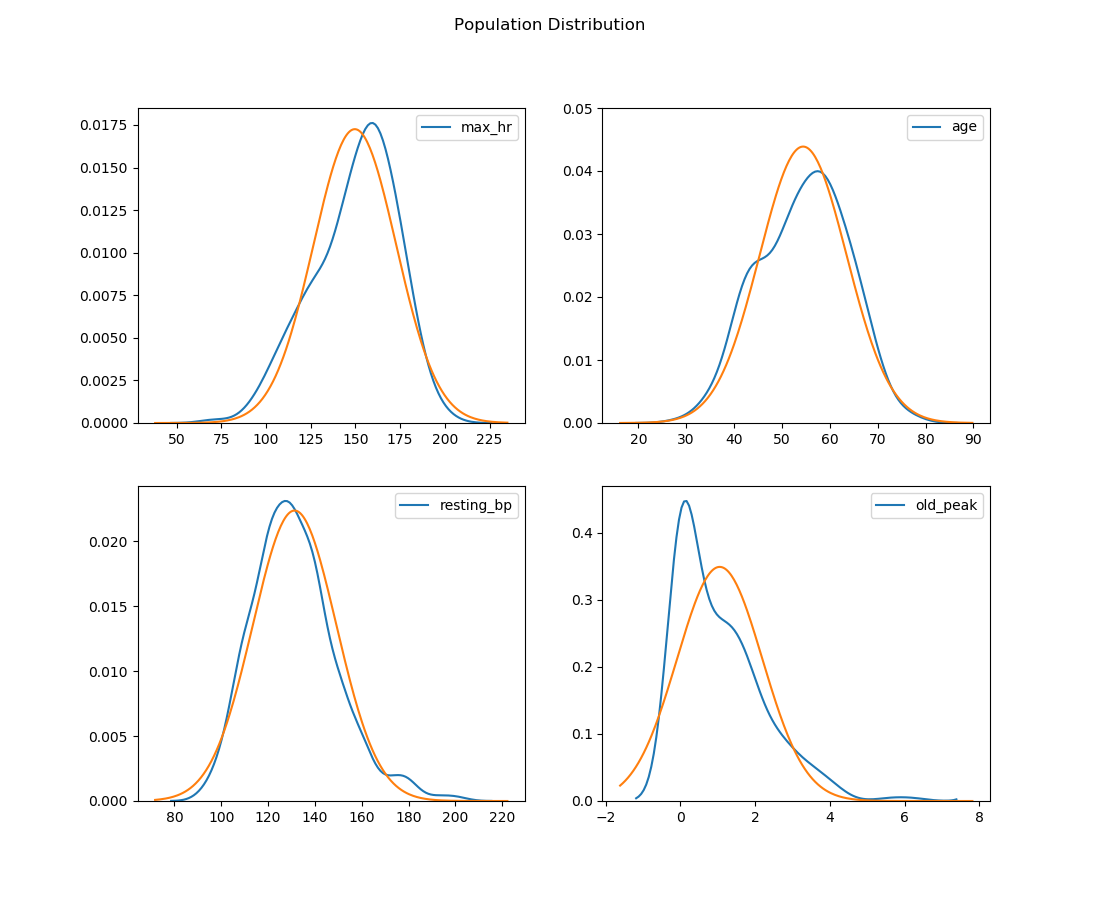


From the above heat map plot, we identify the variables:

1. ‘resting\_bp’
2. ‘max\_hr’
3. ‘age’

It should be noted that these variables show significant correlation and therefore should be used together in a model with reservations to avoid collinearity.

Next, we focus on these three variables and plot the population distribution to evaluate if they follow a normal distribution.



The probability distributions plots above (heart data: blue, scaled normal distribution : orange) make it easy to say that all independent variables follow normal distributions. The dependent variable (old\_peak) shows some right skew and kurtosis which may bias model accuracy towards lower values.

# Linear Regression

Next, we used each variable separately to predict old\_peak using a simple linear model by Ordinary Least Squares regression. The models are summarized in the following table.

Table 1: SLR summary

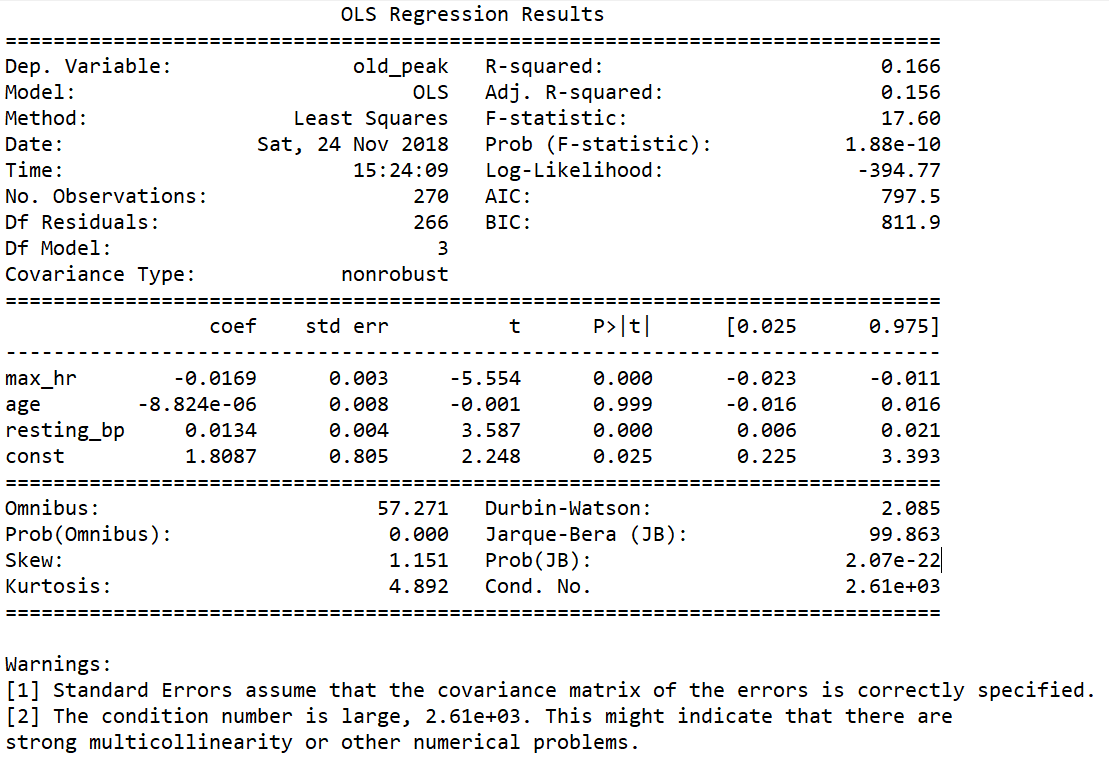
|  |  |  |  |
| --- | --- | --- | --- |
| model | ‘old\_peak ~ max\_hr’ | ‘old\_peak ~ resting\_bp’ | ‘old\_peak ~ age’ |
| (coef, int) | (-0.0173, 3.63) | (-0.0173, 3.63) | (0.0244, -0.279) |
| R2 | .122 | 0.05 | 0.038 |
|  |  |  |  |

Our models have p-values and F-statistics that suggest statistical significance, however, they have fairly low predictive power, with only max\_hr accounting for more than 10% of variability in old\_peak.

These variable predictive power makes sense biologically as patients with high maximum heart rates may have greater capacity to cycle oxygen into their lungs and absorb it, thus it makes sense that patients with higher max\_hr have lower old\_peak. resting\_bp is a common disease predictor and is linked to hardening arterial walls which can decrease tissues capacity to absorb oxygen from the blood, thus it makes sense that patients with higher resting blood pressures have higher old\_peak. Lastly, we see a positive correlation with age and old\_peak but I would propose that this is collinearity based on the biological trend for max heart rate to decrease with age. I would suggest not including age in the final model.

# Multiple Linear Regression

Lastly, we attempt to predict old\_peak using all three variables.



From the above model summary, we see that this model has slightly more predictive power than any of the individual models, however, some of the predictor critical values show less predictive validity. Specifically, the confidence interval for age coefficient includes 0, meaning that it’s directionality can’t be trusted. Furthermore, because of the collinearity between these variables I would be leery of using this model.