Matthew B. Ledesma Machine Learning Prelims March 7, 2025

1. Why did you choose this dataset?

I chose the Heart Disease Risk dataset because I had a hunch that it might show me the difference in predicting health related issues using Linear Regression and Multi-Linear Regression. I think having more independent variables or symptoms would further narrow down the accuracy whether a person is at risk of a certain disease or not. It also intrigued me since most of the data was categorical instead of numerical.

2. What is your conclusion upon conducting linear regression on this dataset?

For the first Linear Regression, I chose to compare the Age as the independent variable and the Heart Risk as the dependent variable. This made sense to me since as we age, we usually have a higher risk of diseases and complications. This however only gave us an R^2 score of 0.37, meaning the it only captures 37% of the heart risk differences. However, the Mean Squared Error (MSE) was quite lower (0.16) compared to the Heart Risk Variance (0.25), which may describe that the model is doing a decent job of predicting close to the true values.

With the Multi-Linear Regression, we achieved a higher R^2 score of 0.87 which means the model captures 87% of the differences. The MSE (0.03) was also significantly lower than the Heart Risk Variance (0.25).

In response to my curiosity mentioned in the first question, the data I have gathered and analyzed showed that the accuracy was much better when using Multi Linear Regression. This is most probably due to the presence of more symptoms that strengthen the relationship between them and the resulting heart risk. This is also the reason why professional medical practitioners cannot diagnose diseases with one symptom alone; they need an array of symptoms that will help them determine the specifics of the problem.

3. How relevant is linear regression today?

I think linear regression is a great tool that helps us analyze and predict certain conditions. Although it has its limits in producing results especially when encountering datasets that are filled with categorical data, there are still useful insights we can gain from it. From analyzing the data and training our own models, we will be able to help produce more hypotheses and conclusions that may benefit research.