**Graduate Program in Software SEIS 763: Machine Learning Assignment #2 (100 points) Due Date: September 27st**

Write a program (Python or Matlab) to find results / answers to the following tasks:

Name : Byunghun Lee

1. Load the patient data from “ML\_HW\_Data\_Patients.csv” file.
2. Use variables **Age, Gender, Height, Weight, Smoker, Location, SelfAssessedHealthStatus** to build a linear regression model to predict the systolic blood pressure.

3.What are the regression coefficients (thetas)?

inputs = Patients\_df[['Age', 'Weight', 'Smoker','Gender', 'County General Hospital','VA Hospital',"St. Mary's Medical Center", 'Excellent','Good', 'Fair', 'Poor' ]]

model.coef\_

array([ 0.09132731, -0.01515041, 9.87541468, 0.5488521 , 0.57694361,

0.17733968, -0.75428329, 0.55602622, 0.77978667, -2.11802701,

0.78221412])

* A coefficient of zero represents no linear relationship. As one variable increases, there is no tendency in the other variable to either increase or decrease.

1. How do you interpret those numbers in thetas?

Clearly, Weight theta is almost zero and **SelfAssessedHealthStatus (Fair)** have a much lower weightage, and i can see why. It has a tiny contribution, and even that is probably accidental. This is an important thing to keep in mind. I wasn’t able to find a relationship that doesn't exist, no matter what machine learning technique or optimization algorithm I apply.

1. If you need to identify one outlier record, which record is a potential outlier? How do you reach this conclusion?

Weight between 200-209 is only one person

1. If you need to identify one or few useless features (independent variables or predictors), which one(s) will you choose? Why do you reach this conclusion?

**SelfAssessedHealthStatus (Fair) is – 2.324657 W so it is**

**SelfAssessedHealthStatus (Fair) .**

**As we can see coefficient theta in this column, it is very subjective feeling to Systolic blood pressure.**