PRAGYA VERMA – (101276204)

SEIS 763: 01 Machine Learning

Assignment #2

---

1 - Load the patient data from "ML_HW_Data_Patients.csv" file.

```
#import libraries

import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
import statsmodels.formula.api as smf
import statsmodels.api
import statsmodels.api as sm
import scipy.stats as stats
import pylab
import matplotlib.pyplot as plot



#Load dataset
Pt_data=pd.read_csv("ML_HW_Data_Patients.csv")

#print top 5 rows of dataset
print(Pt_data.head())
```

2 - Use variables Age, Gender, Height, Weight, Smoker, Location, SelfAssessedHealthStatus to build a linear regression model to predict the systolic blood pressure.

```
#Extracting variables to be used in linear regression

Extvar_Model= Pt_data[["Age", "Gender",
"Height","Weight","Smoker","Location","SelfAssessedHealthStatus","Systolic"
]]
print(Extvar_Model.head())

#scale the numerical columns uniformly
#Using StandardScaler/Zscaler to uniformly scale the numerical columns.
BP_scaler = StandardScaler()
Extvar_Model[['Age','Height', 'Weight']] =
BP_scaler.fit_transform(Extvar_Model[['Age','Height', 'Weight']])
print(Extvar_Model.head())

#create dummy variables for Categorical columns and drop first category
Extvar_Model = pd.get_dummies(Extvar_Model,drop_first=True)
print(Extvar_Model.head())

#Rename column names which are long
Extvar_Model.rename(columns = {"Gender_'Male'":"Male","Location_'St. Mary's
Medical Center'": "Loc_StMaryMedCtr", "Location_'VA Hospital'":
"Loc_VAHosp", "SelfAssessedHealthStatus_'Fair'":
"HealthStat_Fair","SelfAssessedHealthStatus_'Good'":
```

```
"HealthStat_Good","SelfAssessedHealthStatus_'Poor'": "HealthStat_Poor" },
inplace = True)
print(Extvar_Model.head())

# creating formula for regression model .
Formula = 'Systolic ~ Age + Height + Weight + Smoker + Male +
Loc_StMaryMedCtr + Loc_VAHosp + HealthStat_Fair + HealthStat_Good +
HealthStat_Poor'

#Fitting Linear Regression model and displaying the coefficients.
model = smf.ols(formula= Formula, data = Extvar_Model).fit()
model.params.sort_values()
print(model.params.sort_values())


print(model.resid.sort_values())
print(model.summary())
```

3 – What are the regression coefficients (thetas).

Variables	θ (Theta)
HealthStat_Fair	-2.750968
Loc_VAHosp	-1.734841
Male	-1.479391
Loc_StMaryMedCtr	-0.856501
Weight	-0.354757
HealthStat_Poor	0.459343
Age	0.576204
HealthStat_Good	0.586379
Height	1.325387
Smoker	9.673087
Intercept	121.161481

4 - How do you interpret those numbers in thetas?

For Numeric Variable (Age, Weight, height) – Presence of numerica variables means how much blood pressure is going to change when we add or subtract 1 unit in from these variables. Since we have normalized the numeric columns, we use the standard deviation as a unit.

Example: If age goes by 1 unit, or one standard deviation, their systolic blood pressure will go up by **+0. 576204**.

For Categorical Variable (Health Status, Location, Gender, Smoker) – Categorical variable represents whether something is true or not. When categorical variables are presents, theta factor will affect the blood pressure.

Example: if a patient smokes, his/her blood pressure will go up by **+9. 673087** if we compare it to a person who does not smoke**.**

5 - 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?

Ans: To check which variable or feature is useless, we can use theta and p value of variables.

```
print(model.summary())
```

High P-value (> 0.05): The feature is not significant and we can remove.

Low P-value (≤ 0.05): The feature is significant.

- We can see that Weight **coefficient** has the lowest value **-0.354757**
- Weight has the highest **p-value 0.819.** It has minimal effect and can be removed from systolic blood pressure prediction.

```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept                  121.1615      1.851     65.449      0.000     117.483     124.840
Male[T.True]                -1.4794      3.266     -0.453      0.652      -7.968       5.010
Loc_StMaryMedCtr[T.True]    -0.8565      1.298     -0.660      0.511      -3.436       1.723
Loc_VAHosp[T.True]          -1.7348      1.133     -1.531      0.129      -3.987       0.517
HealthStat_Fair[T.True]     -2.7510      1.511     -1.821      0.072      -5.753       0.251
HealthStat_Good[T.True]      0.5864      1.178      0.498      0.620      -1.755       2.928
HealthStat_Poor[T.True]      0.4593      1.676      0.274      0.785      -2.871       3.790
Age                          0.5762      0.481      1.198      0.234      -0.380       1.532
Height                       1.3254      0.717      1.850      0.068      -0.098       2.749
Weight                      -0.3548      1.543     -0.230      0.819      -3.421       2.712
Smoker                       9.6731      1.046      9.249      0.000       7.595      11.751
==============================================================================
```