ASSIGNMENT 2 SEIS 763 2/10/26

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

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. You do NOT need to split data into training and testing sets.
3. What are the regression coefficients (thetas)?
4. How do you interpret those numbers in thetas?
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?

In [ ]:

```python
import pandas as pd
import statsmodels.api as sm
from scipy.stats import zscore


# 1. Load the patient data from “ML_HW_Data_Patients.csv” file.

data = pd.read_csv("ML_HW_Data_Patients.csv")

# inspect the data to understand its structure and contents.

print(data.head()) # look at gender, last name, and location

print(data.info())

print(data.describe())
```

```
   Age  Diastolic    Gender  Height     LastName                     Location  
\
0   38          93    'Male'      71     'Smith'    'County General Hospital'
1   43          77    'Male'      69   'Johnson'                'VA Hospital'
2   38          83  'Female'      64  'Williams'  'St. Mary's Medical Center'
3   40          75  'Female'      67     'Jones'                'VA Hospital'
4   49          80  'Female'      64     'Brown'    'County General Hospital'

  SelfAssessedHealthStatus  Smoker  Systolic  Weight
0              'Excellent'       1       124     176
1                   'Fair'       0       109     163
2                   'Good'       0       125     131
3                   'Fair'       0       117     133
4                   'Good'       0       122     119
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 10 columns):
 #   Column                    Non-Null Count  Dtype 
---  ------                    --------------  ----- 
 0   Age                       100 non-null    int64 
 1   Diastolic                 100 non-null    int64 
 2   Gender                    100 non-null    object
 3   Height                    100 non-null    int64 
 4   LastName                  100 non-null    object
 5   Location                  100 non-null    object
 6   SelfAssessedHealthStatus  100 non-null    object
 7   Smoker                    100 non-null    int64 
 8   Systolic                  100 non-null    int64 
 9   Weight                    100 non-null    int64 
dtypes: int64(6), object(4)
memory usage: 7.9+ KB
None
              Age   Diastolic      Height      Smoker   Systolic      Weight
count  100.000000  100.000000  100.000000  100.000000  100.00000  100.000000
mean    38.280000   82.960000   67.070000    0.340000  122.78000  154.000000
std      7.215416    6.932459    2.836469    0.476095    6.71284   26.571421
min     25.000000   68.000000   60.000000    0.000000  109.00000  111.000000
25%     32.000000   77.750000   65.000000    0.000000  117.75000  130.750000
50%     39.000000   81.500000   67.000000    0.000000  122.00000  142.500000
75%     44.000000   89.000000   69.250000    1.000000  127.25000  180.250000
max     50.000000   99.000000   72.000000    1.000000  138.00000  202.000000
```

In [ ]:

```
# 2. STANDARDIZATION
# apply z-score only to contin. vars: Age, Height, Weight
continuous_vars = ['Age', 'Height', 'Weight']
data[continuous_vars] = data[continuous_vars].apply(zscore)

# 3. Define predictors
target = 'Systolic'
features = ['Age', 'Gender', 'Height', 'Weight', 'Smoker', 'Location', 'Self

X = data[features].copy()
y = data[target]

# 4. for categorical vars (one hot encoding)
# drop_first=True creates the Reference Groups
```

```
X = pd.get_dummies(X, columns=['Gender', 'Location', 'SelfAssessedHealthStat

# 5. Add Constant
X = sm.add_constant(X)

# 6. Fit Model (Force Float to prevent errors)
model = sm.OLS(y, X.astype(float)).fit()

# 7. Print Results
print(model.summary())
```

```
                             OLS Regression Results
==============================================================================
==
Dep. Variable:                Systolic   R-squared:                        0.5
57
Model:                             OLS   Adj. R-squared:                   0.5
07
Method:                  Least Squares   F-statistic:                      11.
19
Date:                 Mon, 16 Feb 2026   Prob (F-statistic):            3.89e-
12
Time:                         16:13:15   Log-Likelihood:                 -291.
09
No. Observations:                  100   AIC:                              60
4.2
Df Residuals:                       89   BIC:                              63
2.8
Df Model:                           10
Covariance Type:             nonrobust
==============================================================================
============================
                                                coef    std err          t      P
>|t|      [0.025      0.975]
------------------------------------------------------------------------------
----------------------------
const                                       121.1615      1.851     65.449
0.000     117.483     124.840
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
Gender_'Male'                                -1.4794      3.266     -0.453
0.652      -7.968       5.010
Location_'St. Mary's Medical Center'     -0.8565      1.298     -0.660
0.511      -3.436       1.723
Location_'VA Hospital'                       -1.7348      1.133     -1.531
0.129      -3.987       0.517
SelfAssessedHealthStatus_'Fair'          -2.7510      1.511     -1.821
0.072      -5.753       0.251
SelfAssessedHealthStatus_'Good'           0.5864      1.178      0.498
0.620      -1.755       2.928
SelfAssessedHealthStatus_'Poor'           0.4593      1.676      0.274
0.785      -2.871       3.790
==============================================================================
==
Omnibus:                         3.710   Durbin-Watson:                    1.7
47
Prob(Omnibus):                   0.156   Jarque-Bera (JB):                 3.7
23
Skew:                            0.451   Prob(JB):                         0.1
55
Kurtosis:                        2.718   Cond. No.                          1
```

```
2.3
============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
```