In [ ]:

Download (right-click, save target as ...) this page as a jupyterlab notebook Lab29

Laboratory 29: Multiple Regression

Medrano, Giovanni

R11521018

ENGR 1330 Laboratory 29 - In Lab

Example

Explore the data set heart.data.csv and determine the effect that the independent variables biking and smoking have on the dependent variable heart disease using a multiple linear regression model.

In [1]:
```
# Load the necessary packages
import numpy as np
import pandas as pd
import statistics
import math
from matplotlib import pyplot as plt
import statsmodels.formula.api as smf
```

Get the datafile

In [2]:
```
import requests # Module to process http/https requests
remote_url="http://54.243.252.9/engr-1330-webroot/8-Labs/Lab29/heart.data.csv"  # set t
rget = requests.get(remote_url, allow_redirects=True)  # get the remote resource, follo
open('heart.data.csv','wb').write(rget.content); # extract from the remote the contents
```

Read into a dataframe. The database original source \@ https://www.scribbr.com/statistics/multiple-linear-regression/ does not report the units on each variable, a good guess is **biking** is miles per week, **smoking** is packs per week, and **heart.disease** is possibly hospital admissions per 100000 for coronary complications.

After the read we need some shenigagins to get the column names meaningful.

In [7]:
```
heartattack = pd.read_csv('heart.data.csv')
data = heartattack.rename(columns={"biking":"Bike","smoking":"Smoke","heart.disease":"D
data.head(3)
```

Out[7]:

	Unnamed: 0	Bike	Smoke	Disease
0	1	30.801246	10.896608	11.769423
1	2	65.129215	2.219563	2.854081
2	3	1.959665	17.588331	17.177803

In [8]:
```
Now build a linear model
```

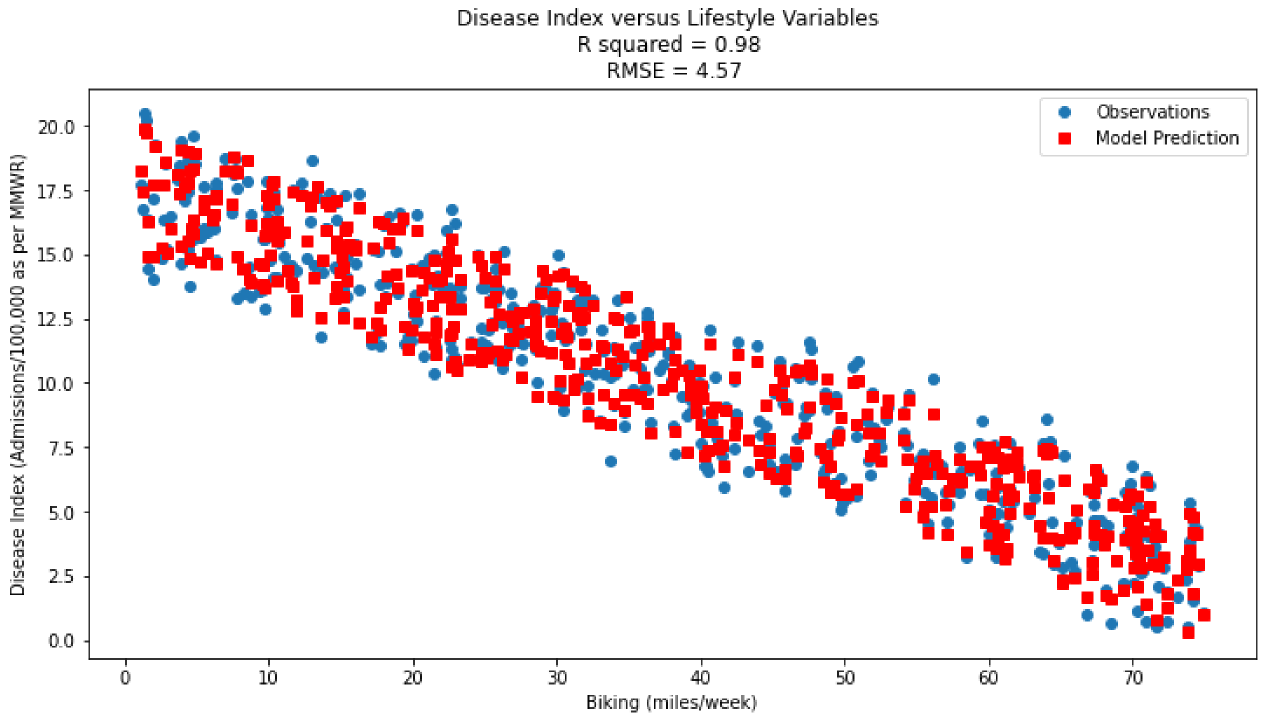
```
  File "<ipython-input-8-5753f2df4b64>", line 1
    Now build a linear model
        ^
SyntaxError: invalid syntax
```

In [13]:
```
# Initialise and fit linear regression model using `statsmodels`
model = smf.ols('Disease ~ Bike + Smoke', data=data)
model = model.fit()
#print(model.summary())
# dir(model) # activate to find attributes
intercept = model.params[0]
slope = model.params[1]
Rsquare = model.rsquared
RMSE = math.sqrt(model.mse_total)
print('y= '+str(slope)+"*x + " + str(intercept))
```

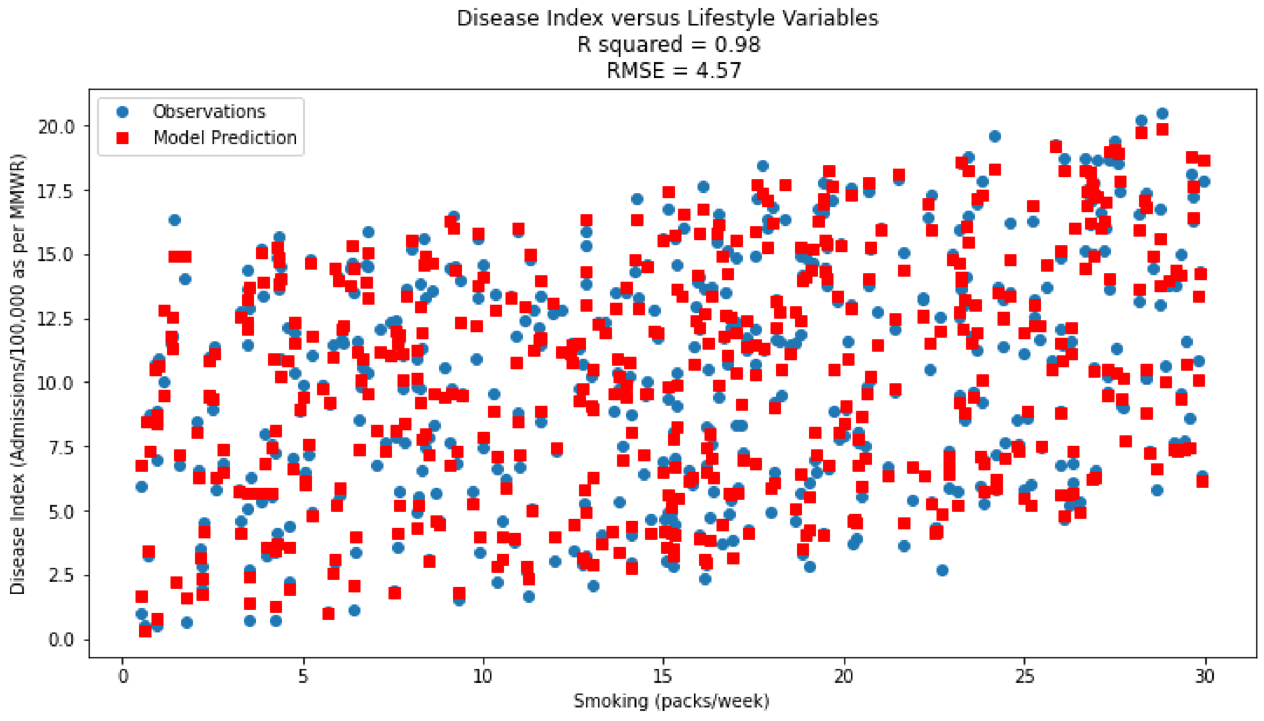
```
y= -0.2001330519586229*x + 14.984657987225836
```

To find the various values a visit to Here is useful! Below we will construct a title line that contains the equation, RMSE, and R-square using type casting and concatenation, then pass it to the plot.

In [14]:
```
# Predict values
heartfail = model.predict()

titleline = 'Disease Index versus Lifestyle Variables \n'  + 'R squared = ' + str(round
# Plot regression against actual data - What do we see?
plt.figure(figsize=(12, 6))
plt.plot(data['Bike'], data['Disease'], 'o')           # scatter plot showing actual da
plt.plot(data['Bike'], heartfail, marker = 's' ,color ='r', linewidth=0)   # regression
plt.xlabel('Biking (miles/week)')
plt.ylabel('Disease Index (Admissions/100,000 as per MMWR)')
plt.legend(['Observations','Model Prediction'])
plt.title(titleline)

plt.show()
```



In [15]:
```
titleline = 'Disease Index versus Lifestyle Variables \n'  + 'R squared = ' + str(round
# Plot regression against actual data - What do we see?
plt.figure(figsize=(12, 6))
plt.plot(data['Smoke'], data['Disease'], 'o')           # scatter plot showing actual d
plt.plot(data['Smoke'], heartfail, marker = 's' ,color ='r', linewidth=0)   # regressio
plt.xlabel('Smoking (packs/week)')
plt.ylabel('Disease Index (Admissions/100,000 as per MMWR)')
plt.legend(['Observations','Model Prediction'])
plt.title(titleline)

plt.show()
```



Now lets learn about the actual model

In [16]:
```
print(model.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                Disease   R-squared:                       0.980
Model:                            OLS   Adj. R-squared:                  0.980
Method:                 Least Squares   F-statistic:                 1.190e+04
Date:                Mon, 01 Aug 2022   Prob (F-statistic):               0.00
Time:                        17:52:31   Log-Likelihood:                -493.68
No. Observations:                 498   AIC:                             993.4
Df Residuals:                     495   BIC:                             1006.
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     14.9847      0.080    186.988      0.000      14.827      15.142
Bike          -0.2001      0.001   -146.525      0.000      -0.203      -0.197
Smoke          0.1783      0.004     50.387      0.000       0.171       0.185
==============================================================================
Omnibus:                        2.794   Durbin-Watson:                   1.917
Prob(Omnibus):                  0.247   Jarque-Bera (JB):                2.582
Skew:                          -0.141   Prob(JB):                        0.275
Kurtosis:                       3.211   Cond. No.                         125.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifi
ed.
```

Exercise 1

Interpret the results.

1. Is there a correlation with smoking and disease (based on our data)?
2. Is there a correlation with biking and disease (based on our data)?
3. Which is the better lifestyle choice, if the goal is reduced disease? How much better?
4. Are the parameters significant (i.e. non-zero and zero is not contained in the estimation interval)?

1.The correlation between smoking and disease is .178. Which means there is a positive correlation even if it isn't really strong.

2. Yes there is a correlation between biking and disease but it is negative. It is -.2001 which is almost the inverse of num1. Meaning there is a weak negative correlation.

3. Based on the data the better lifestytle choice is to bike as it leads to a lower disease rate. It is almost the exact opposite of smoking. Meaning it is better for you.

4. The parameters are both significant as they are non-zero.

Exercise 2

Using the tools from Lab 28, produce labeled plots of prediction intervals for disease index using biking and smoking as predictor varuables (note the work is already done above, you just need to access the object and make and label the plots)

Tips:

The data are not ordered, hence we choose to plot only markers even for the models, to get usual plots using lines, you need to sort the fitted results and plot those - its a bit of a hassle and left as a bonus problem.

Thank you for letting us skip this exercise 2 archer :D

```
In [ ]:
```