Applied Statistics MATH 661 Assignment #3

September 25, 2019

Andres Castellano

1 Question 2 Excercise 2.8 Protein and Fat

A data set containing the dietary infomation of meals from a sample or the population of first year college students can be used to study the relationship between protein and fat intake of those students. For example, the protein and fat amounts of every meal in the data set can be plotted on different axes of a scatter diagram. In this way, it may be determined whether there is a relationship between this variables.

2 Question 3 Excercise 2.24, 2.25, and 2.26 Bone Strenght

2.1 IPC 2.24

[1]:
```
import pandas as pd
import numpy as np

data = {'Nondominant':[15.7,25.2,17.9,19.1,12.0,20.0,12.3,14.4,
                      15.9,13.7,17.7,15.5,14.4,14.1,12.3],
       'Dominant':[16.3,26.9,18.7,22.0,14.8,19.8,13.1,17.5,20.1,
                   18.7,18.7,15.2,16.2,15.0,12.9]}
frata = pd.DataFrame(data,
                           index=[np.arange(1,16)],
                           columns=pd.Index(['Nondominant','Dominant'],
                                               name='ID'))
%matplotlib inline
frata
```

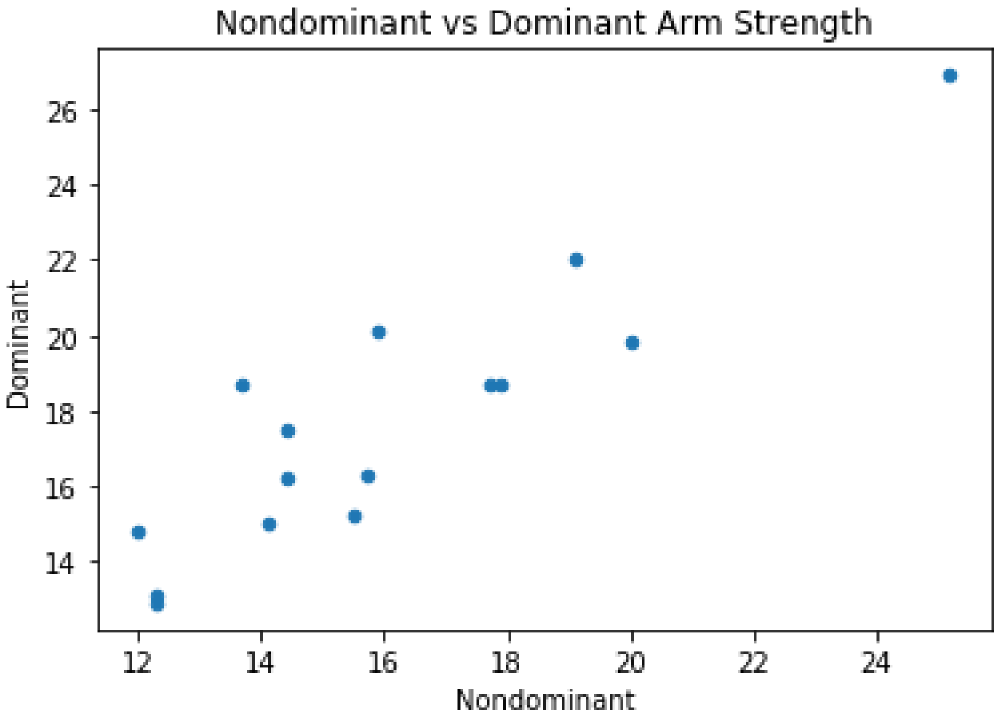
[1]:
```
ID  Nondominant  Dominant
1          15.7      16.3
2          25.2      26.9
3          17.9      18.7
4          19.1      22.0
```

```
5            12.0      14.8
6            20.0      19.8
7            12.3      13.1
8            14.4      17.5
9            15.9      20.1
10           13.7      18.7
11           17.7      18.7
12           15.5      15.2
13           14.4      16.2
14           14.1      15.0
15           12.3      12.9
```

2.1.1 a) Create a Scatterplot of The Data with the Nondominant and Dominant arm strengths on the x and y axis respectively.

[2]:
```
%matplotlib inline
frata.plot.scatter('Nondominant','Dominant',
                   title='Nondominant vs Dominant Arm Strength')
```

[2]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x23442a6e5f8>
```



2.1.2 b,c,d, and e) Describe the overall pattern in the scatterplot and any striking deviations from the pattern

The Data appears to follow a linear behavior densely distributed between 12,000 and 20,000 Newton-Meters. One notable instance occurs approximately at 26,000 Newton Meters. Although the instance seems to behave according with the overall trend of the data, it is outside of the densely populated area mentioned above. However, it is not likely that the instance is an outlier since it appears to follow the overall linear trend of the rest of the data. It is possible that the sample did not include enought data between 22000 and 24000 Newton Meters.

2.2 IPC 2.25

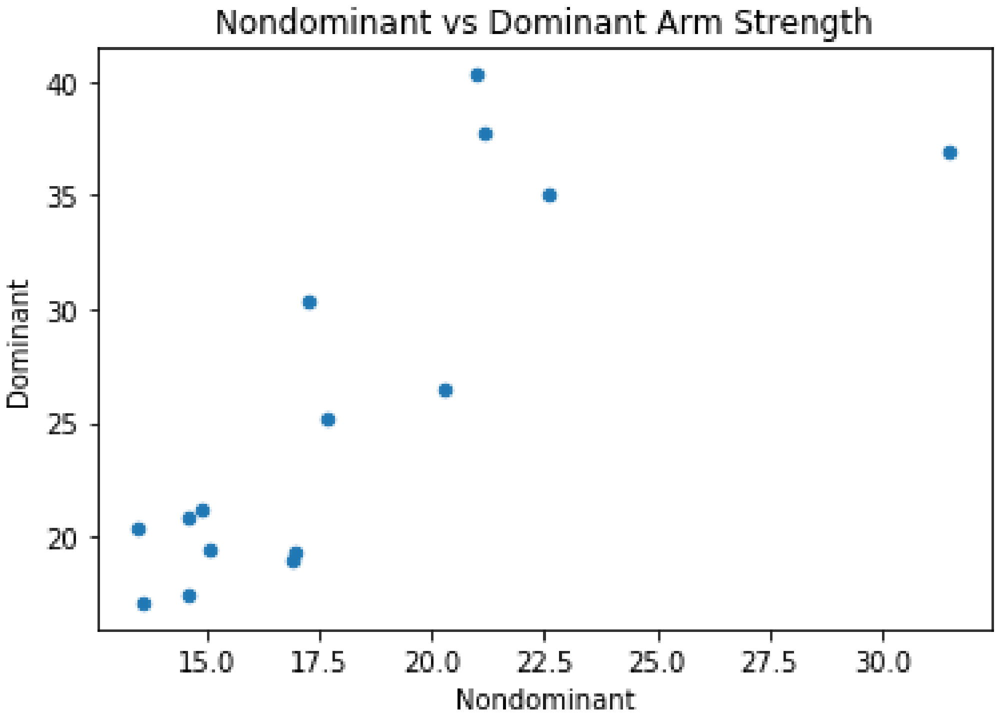
[3]:
```
data2 = {'Nondominant':[17.0,16.9,17.7,21.2,21.0,14.6,31.5,14.9,
                        15.1,13.5,13.6,20.3,17.3,14.6,22.6],
        'Dominant':[19.3,19.0,25.2,37.7,40.3,20.8,36.9,21.2,
                     19.4,20.4,17.1,26.5,30.3,17.4,35.0]}
frata2 = pd.DataFrame(data2,
                  index = [np.arange(16,31)],
                  columns = pd.Index(['Nondominant','Dominant'],name='ID'))
frata2
```

[3]:
```
ID  Nondominant  Dominant
16         17.0      19.3
17         16.9      19.0
18         17.7      25.2
19         21.2      37.7
20         21.0      40.3
21         14.6      20.8
22         31.5      36.9
23         14.9      21.2
24         15.1      19.4
25         13.5      20.4
26         13.6      17.1
27         20.3      26.5
28         17.3      30.3
29         14.6      17.4
30         22.6      35.0
```

2.2.1 a) Make a scatter plot of the data.

[4]:
```
%matplotlib inline
frata2.plot.scatter('Nondominant','Dominant',
                   title='Nondominant vs Dominant Arm Strength')
```

[4]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x23442a4fdd8>
```



2.2.2 b,c,d,e) Describe the overall pattern in the scatterplot and any striking deviations from the pattern.

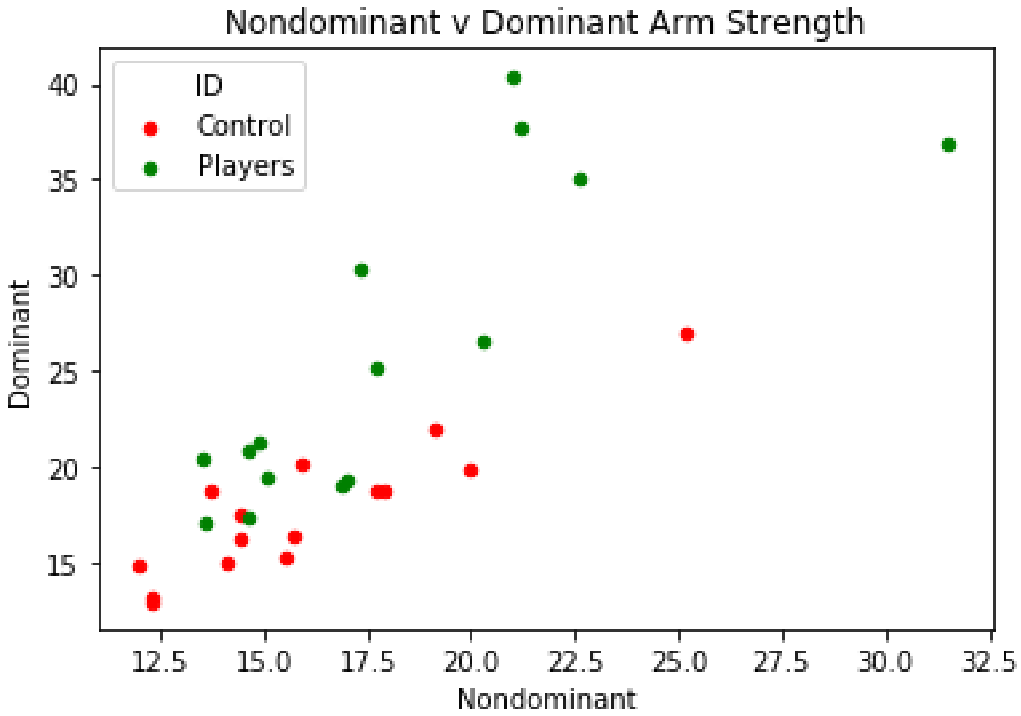
The data is clustered between 15,000 and 22,500 Newton-Meters. However, the data does not seem to be linearly related. In fact, at first glance, the data appears to follow a positive exponential distribution except for one possible outlier found at approximately 34,000 Newton-Meters.

2.3 IPC 2.26

2.3.1 a) Create a Scatter plot of both sets of data with different markers for the control group.

[5]:

```
ax1 = frata.plot(kind='scatter',x='Nondominant',y='Dominant',
                color='r',label='Control')
ax2 = frata2.plot(kind='scatter',x='Nondominant',y='Dominant',
                color='g',ax=ax1,
                 title='Nondominant v Dominant Arm Strength',label='Players')
```



2.3.2 b) Discussion

From the scatter plot of both samples, it is clear the samples do not belong to the same population. There is a clear difference between arm strenghts of dominant vs nondominant hands of Baseball Players compared to the general population of young men.

---

3 Question 5 Excercise 2.47 and 2.48 Bone Strength Continued.

3.1 IPC 2.47

3.1.1 a) Find the correlation between bone strenght of dominant and non dominant arms of the controls.

[6]:
```
frata.corr(method='pearson')#Defines Correlation between Columns of
                                        #DataFrame
```

[6]:
```
ID           Nondominant  Dominant
ID
Nondominant     1.000000  0.904894
Dominant        0.904894  1.000000
```

3.1.2 b) Look at the scatterplot of this data and determine whether the correlation found in 2.47a is a good numerical summary of the graphical display in the scatterplot.

As expected, the .904 value found in 2.47a is a strong indicator of the linear correlation observed in 2.24.

3.2 IPC 2.48

3.2.1 a) Find the correlation between bonde strenght of the dominant and non dominant arms of baseball players.

```
[7]: frata2.corr(method='pearson')#Computes Column wise correlation of
                                  #DataFrame
```

```
[7]: ID            Nondominant  Dominant
     ID
     Nondominant      1.000000  0.793589
     Dominant         0.793589  1.000000
```

3.2.2 b) Is the correlation coefficient found in 2.48a a good summary of the graphical data observed in 2.25a?

Yes, a correlation coefficient of .79 indicates a weak correlatino between the observed values. As observed on the scatterplot, the data does not resemble a linear distribution and the numerical and graphical analysis corroborate this result.

---

4 Question 6 IPC Excercises 2.68, 2.69, 2.70, 2.71, 2.72 Bone Strength (Continued)

4.1 IPC 2.68

4.1.1 a) Plot the data using nondominant strength as an explanatory variable for the dominant arm strength.

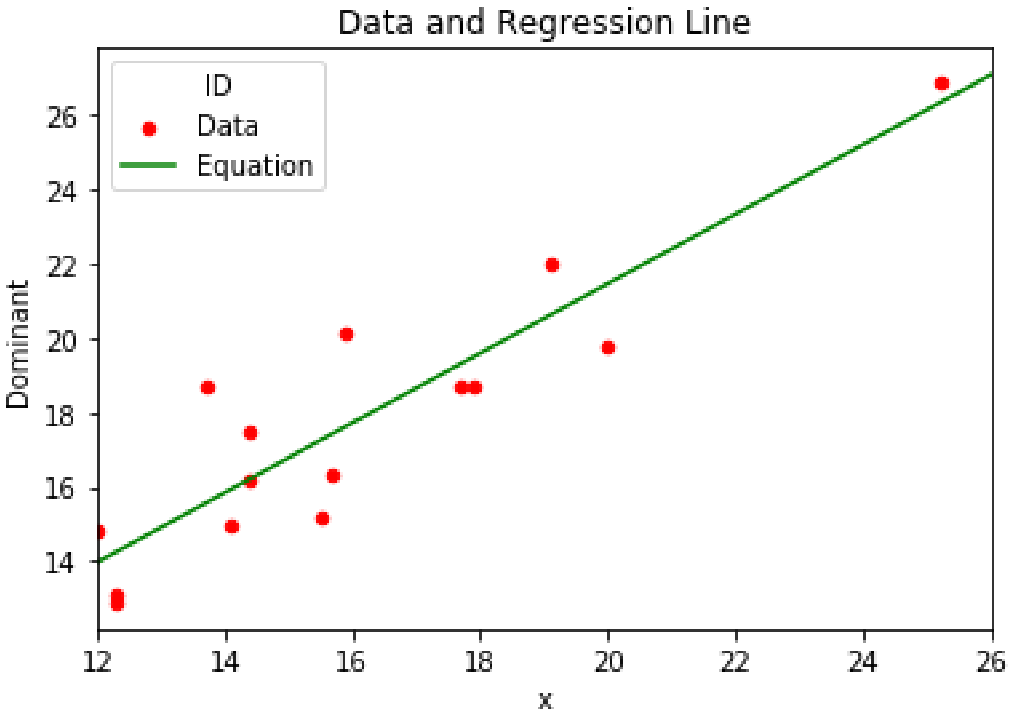
See Chart in 2.68b

4.1.2 b) Add a line with the equation y = 2.74+.936*x to the previous chart.

```
[8]: ax3 = frata.plot(kind='scatter',x='Nondominant',y='Dominant',
                     color='r',label='Data')
     def func(args):
         y=args*.936+2.74
         return y
     x = np.arange(12,27)
     Eframe = pd.DataFrame([x,func(x)],index=['x','y']).T

     ax4 = Eframe.plot(kind='line',x='x',y='y',
                      color='g',ax=ax3,
                       label='Equation',
                       title='Data and Regression Line')
     print(ax3==ax4)
```

```
True
```



4.1.3 c) Discussion

As indicated by the chart above and previous discussion, this data set appears to be linearly distributed. The data points and trendline show a clear linear relationship.

4.2 IPC 2.69

4.2.1 a) Plot the data using nondominant strength as an explanatory variable for the dominant arm strength.

See Chart in 2.69b

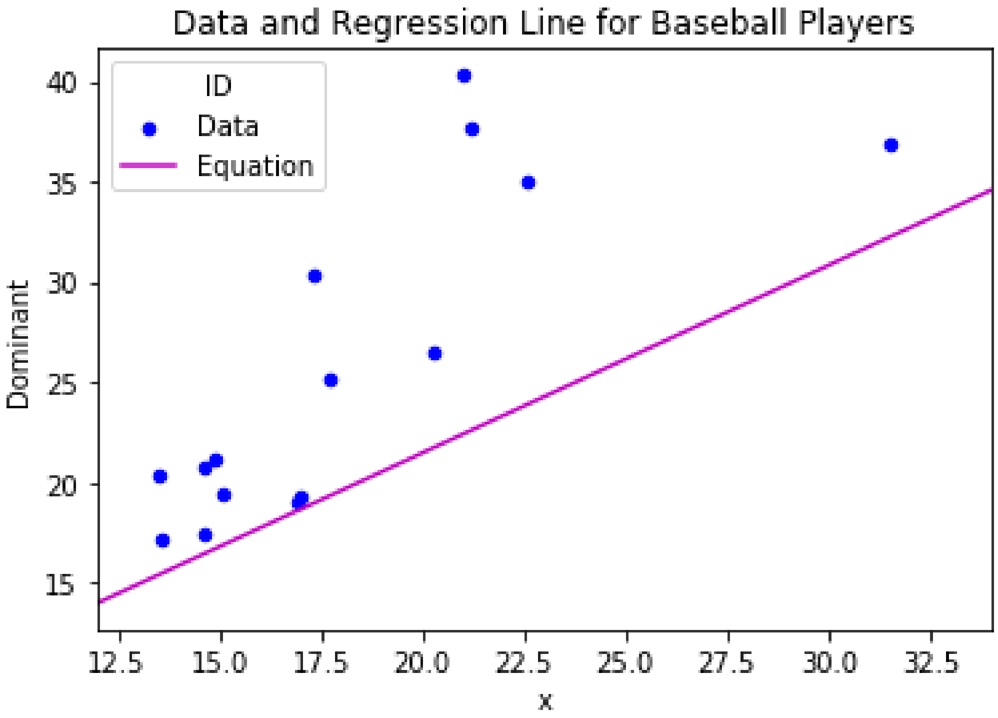
4.2.2 b) Add a line with the equation 'y = .886+1.373*x' to the previous chart.

[9]:

```
ax5 = frata2.plot(kind='scatter',x='Nondominant',y='Dominant',
                  color='b',label='Data')
def func2(args):
    y=args*1.373+.886
    return y
x = np.arange(12,35)
Eframe2 = pd.DataFrame([x,func(x)],index=['x','y']).T

ax6 = Eframe2.plot(kind='line',x='x',y='y',
                   color='m',ax=ax5,
                    label='Equation',
                    title='Data and Regression Line for Baseball Players')
print(ax3==ax4)
```

```
True
```



4.2.3 c) Discussion

From the data points it is evident that the data does not follow a linear distribution well. By plotting the best fit line on the same graph, it can be shown that the data points do not lie anywhere near the line that best approximates the line and so it can de concluded that nondominant arm strenght is not a good explanatory variable for dominant arm strenght of baseball players

4.3 IPC 2.70

4.3.1 Assuming that 1 cmˆ4 is equal to 1 Nm, we can use the equation from 2.68b to estimate the bone strength of the nondominant arm of a young man who does not play baseball.

We can do this because we have established a linear relationship between nondominant and dominant arm bone strength for young man who do not play baseball. The equation f(16) = 2.74+.936(x) yields a dominant arm bone strength of 17.716 Newton meters/1000 or 17,716Nm.

4.4 IPC 2.71

4.4.1 Predict the dominant arm bone strength of a baseball player using his nondominant arm bone strength.

We do not have an appropriate way to predict dominant arm strength of baseball players. The results in this assignment are not appropriate since the given equation for baseball players cannot be used to estimate dominant arm bone strength.

5 Question 8 IPC Excercises 2.127 and 2.128

5.1 IPC 2.127

[10]:

```
phata = pd.DataFrame({'Hospital A':[63,2037,2100],
       'Hospital B':[16,784,800]},
                     index=['Died','survived','Total'])
```

[11]:

```
phata
pctdieda = (100)*phata.loc['Died','Hospital A']/phata.loc['Total','Hospital A']
pctdiedb = (100)*phata.loc['Died','Hospital B']/phata.loc['Total','Hospital B']
#print(pctdieda,'Percent of Patients died in Hospital A and',
     # pctdiedb,'Percent of patients died in Hospital B')

print(('%4.2f%% of Patients in Hospital A died and '+
      '%4.2f%% of patients in Hospital B died.') %
      (pctdieda,pctdiedb))
```

```
3.00% of Patients in Hospital A died and 2.00% of patients in Hospital B died.
```

5.2 IPC 2.128

[12]:

```
food_cond = pd.DataFrame({'Hospital A':[6,594,600],
                         'Hospital B':[8,592,600]},
                        index=['Died','Survived','Total'])
print('\033[1m'+
      '\033[4m'+
      'Patients in Good Condition \n\n'+
      '\033[0m',
      food_cond)
```

```
Patients in Good Condition            Hospital A  Hospital B
Died                 6           8
Survived           594         592
Total              600         600
```

[13]:

```
print('\033[1m'+
      '\033[4m'+
      'Patients in Poor Condition \n\n'+
      '\033[0m',
      foor_cond)
```

```
      ␣
↪---------------------------------------------------------------------------

      NameError                                 Traceback (most recent call␣
↪last)

      <ipython-input-13-61a7cdf9de27> in <module>
        3       'Patients in Poor Condition \n\n'+
        4       '\033[0m',
  ----> 5       foor_cond)


      NameError: name 'foor_cond' is not defined
```

5.2.1 a) In which hospitals do patients marked in poor condition before surgery fare better?

[ ]:

```
pctdieda = (100)*foor_cond.loc['Died','Hospital A']/foor_cond.
 ↪loc['Total','Hospital A']
pctdiedb = (100)*foor_cond.loc['Died','Hospital B']/foor_cond.
 ↪loc['Total','Hospital B']
#print(pctdieda,'Percent of Patients died in Hospital A and',
     # pctdiedb,'Percent of patients died in Hospital B')

print(('%4.2f%% of Patients marked in poor condition at Hospital A '+
       'died and %4.2f%% of patients marked in poor condition at '+
       'Hospital B died.') %
      (pctdieda,pctdiedb))
```

5.2.2 b) In which hospitals do patients marked in good condition before surgery fare better?

[ ]:

```
pctdieda = (100)*food_cond.loc['Died','Hospital A']/food_cond.
 ↪loc['Total','Hospital A']
```

```
pctdiedb = (100)*food_cond.loc['Died','Hospital B']/food_cond.
 ↪loc['Total','Hospital B']
#print(pctdieda,'Percent of Patients died in Hospital A and',
     # pctdiedb,'Percent of patients died in Hospital B')

print(('%4.2f%% of Patients marked in good condition at Hospital A '+
       'died and %4.2f%% of patients marked in good condition at '+
       'Hospital B died.') %
      (pctdieda,pctdiedb))
```

5.2.3 c) Recommendation for someone facing surgery.

Someon facing surgery and choosing between these two hospitals should consider hospital A since it features a lower percentage of fatalities for patients marked in good as well as poor condition before surgery.

5.2.4 d) Discussion

There are a lot more data points for Hospital A than for Hospital B. It is possible that the difference in data sample size has a direct impact in the comparison of our statistics.

[ ]: