HW 1

Idit Belachsen 032583940 Dana Drahler 305702003

Q1

A

נתונים:

$$Y_1 = w_1 + w_2 X_1 + \epsilon_1 \qquad\qquad Y_2 = w_1 + w_2 X_2 + \epsilon_2$$

נחשב את $\widehat{w_1}, \widehat{w_2}$ ע"י שיטת הריבועים הפחותים:

$$\sum_{i=1}^{2} (\epsilon_i)^2 = \sum_{i=2}^{2} \left(Y_i - \widehat{Y_i}\right)^2 = \sum_{i=2}^{2} (Y_i - \widehat{w_1} - \widehat{w_2} X_i)^2$$

<u>נגזור לפי $\widehat{w_1}$ ונשווה ל-0</u>

$$\frac{d \sum_{i=1}^{2} (\epsilon_i)^2}{d\widehat{w_1}} = -2 \sum_{i=2}^{2} Y_i - \widehat{w_1} - \widehat{w_2} X_i = -2 \left(\sum_{i=2}^{2} Y_i - \sum_{i=2}^{2} \widehat{w_1} - \sum_{i=2}^{2} \widehat{w_2} X_i \right)$$
$$= -2 \sum_{i=2}^{2} Y_i + 2 \sum_{i=2}^{2} \widehat{w_1} + 2 \sum_{i=2}^{2} \widehat{w_2} X_i = 0$$

נצמצם ונקבל:

$$\sum_{i=2}^{2} Y_i - \sum_{i=2}^{2} \widehat{w_1} - \sum_{i=2}^{2} \widehat{w_2} X_i = 0$$

נזכור ש-

$$\bar{Y} = \frac{\sum_{i=2}^{2} Y_i}{2} \quad \rightarrow \quad \sum_{i=2}^{2} Y_i = 2\,\bar{Y}$$

$$\bar{X} = \frac{\sum_{i=2}^{2} Y_i}{2} \quad \rightarrow \quad \sum_{i=2}^{2} X_i = 2\,\bar{X}$$

נציב ונקבל:

$$2\,\bar{Y} - 2\widehat{w_1} - 2\widehat{w_2}\bar{X} = 0$$

ע"י העברת אגפים נקבל את משוואה 1:

$$\widehat{w_1} = \bar{Y} - \widehat{w_2}\bar{X}$$

<u>באופן דומה, נגזור לפי $\widehat{w_2}$ ונשווה ל-0:</u>

$$\frac{d \sum_{i=1}^{2} (\epsilon_i)^2}{d\widehat{w_2}} = -2 \sum_{i=2}^{2} (Y_i - \widehat{w_1} - \widehat{w_2} X_i) * X_i = 0$$

נציב את משוואה 1 ונקבל:

$$= -2\sum_{i=2}^{2}(Y_i - \bar{Y} + \widehat{w_2}\bar{X} - \widehat{w_2}X_i) * X_i = -2\sum_{i=2}^{2}[(Y_i - \bar{Y}) - \widehat{w_2}(X_i - \bar{X})] * X_i = 0$$

$$\sum_{i=2}^{2}[(Y_i - \bar{Y}) - \widehat{w_2}(X_i - \bar{X})] * X_i = \sum_{i=2}^{2}(Y_i - \bar{Y}) * X_i - \widehat{w_2}\sum_{i=2}^{2}(X_i - \bar{X}) * X_i = 0$$

$$\widehat{w_2} = \frac{\sum_{i=2}^{2}(Y_i - \bar{Y}) * X_i}{\sum_{i=2}^{2}(X_i - \bar{X}) * X_i}$$

נפתח את המכנה:

$$\sum_{i=2}^{2}(X_i - \bar{X}) * X_i = \sum_{i=2}^{2}\left(X_i^2 - \bar{X}X_i\right) = \sum_{i=2}^{2}X_i^2 - \sum_{i=2}^{2}\bar{X}X_i = \sum_{i=2}^{2}X_i^2 - \bar{X}\sum_{i=2}^{2}X_i$$

נציב $\sum_{i=2}^{2}X_i = 2\bar{X}$ ונקבל:

$$\sum_{i=2}^{2}X_i^2 - \bar{X}\sum_{i=2}^{2}X_i = \sum_{i=2}^{2}X_i^2 - 2\bar{X}^2 = \sum_{i=2}^{2}X_i^2 - \sum_{i=2}^{2}\bar{X}^2 = \sum_{i=2}^{2}\left(X_i^2 - \bar{X}^2\right)$$

נשים לב ש-

$$\begin{aligned}\sum_{i=2}^{2}(X_i - \bar{X})^2 &= \sum_{i=2}^{2}\left(X_i^2 - 2X_i\bar{X} + \bar{X}^2\right) = \sum_{i=2}^{2}X_i^2 - 2\bar{X}\sum_{i=2}^{2}X_i + \sum_{i=2}^{2}\bar{X}^2 \\ &= \sum_{i=2}^{2}X_i^2 - 4\bar{X}^2 + 2\bar{X}^2 = \sum_{i=2}^{2}X_i^2 - 2\bar{X}^2 = \sum_{i=2}^{2}X_i^2 - \sum_{i=2}^{2}\bar{X}^2 \\ &= \sum_{i=2}^{2}(X_i^2 - \bar{X}^2)\end{aligned}$$

ולכן:

$$\sum_{i=2}^{2}\left(X_i^2 - \bar{X}^2\right) = \sum_{i=2}^{2}(X_i - \bar{X})^2$$

נפתח את המונה:

$$\sum_{i=2}^{2}(Y_i - \bar{Y}) * X_i = \sum_{i=2}^{2}Y_iX_i - \sum_{i=2}^{2}\bar{Y}X_i = \sum_{i=2}^{2}Y_iX_i - \bar{Y}\sum_{i=2}^{2}X_i$$

נציב $\sum_{i=2}^{2}X_i = 2\bar{X}$ ונקבל:

$$\sum_{i=2}^{2}Y_iX_i - \bar{Y}\sum_{i=2}^{2}X_i = \sum_{i=2}^{2}Y_iX_i - 2\bar{Y}\bar{X} = \sum_{i=2}^{2}Y_iX_i - 2\bar{Y}\bar{X} - 2\bar{Y}\bar{X} + 2\bar{Y}\bar{X} =$$

$$\sum_{i=2}^{2} Y_iX_i - \sum_{i=2}^{2} \bar{Y}X_i - \sum_{i=2}^{2} Y_i\bar{X} - \sum_{i=2}^{2} \bar{Y}\bar{X} = \sum_{i=2}^{2} (Y_i - \bar{Y})(X_i - \bar{X})$$

ולכן:

$$\widehat{w_2} = \frac{\sum_{i=2}^{2}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=2}^{2} X_i^2 - \bar{X}^2}$$

לסיכום:

$$\widehat{w_1} = \ \bar{Y} - \widehat{w_2}\bar{X}$$

$$\widehat{w_2} = \frac{\sum_{i=2}^{2}(Y_i - \bar{Y}) * (X_i - \bar{X})}{\sum_{i=2}^{2}(X_i - \bar{X})^2}$$

B

האומדים אינם מוטים אם מתקיימים:

$E(\widehat{w_1}) = w_1 \qquad\qquad E(\widehat{w_2}) = w_2$

<u>נבצע חישובים מקדימים:</u>

$$E(\bar{Y}) = E\left(\frac{\sum_{i=2}^{2} Y_i}{2}\right) = \frac{1}{2}E\left(\sum_{i=2}^{2} Y_i\right) = \frac{1}{2}E\left(\sum_{i=2}^{2} w_1 + w_2X_i + \epsilon_i\right)$$

תחת ההנחה ש $E(\epsilon_i) = 0$ נקבל:

$$\frac{1}{2}E\left(\sum_{i=2}^{2} w_1\right) + \frac{1}{2}E\left(\sum_{i=2}^{2} w_2X_i\right) + \frac{1}{2}E\left(\sum_{i=2}^{2} \epsilon_i\right) = \frac{2w_1}{2} + \frac{2w_2\bar{X}}{2} + 0 = w_1 + w_2\bar{X}$$

$$E(\bar{Y}) = w_1 + w_2\bar{X}$$

$$\widehat{w_2} = \frac{\sum_{i=2}^{2}(Y_i - \bar{Y}) * (X_i - \bar{X})}{\sum_{i=2}^{2}(X_i - \bar{X})^2} = \frac{\sum_{i=2}^{2}(Y_i(X_i - \bar{X}) - \bar{Y}(X_i - \bar{X}))}{\sum_{i=2}^{2}(X_i - \bar{X})^2}$$

$$= \frac{\sum_{i=2}^{2}(Y_i(X_i - \bar{X}) - \bar{Y}(X_i - \bar{X}))}{\sum_{i=2}^{2}(X_i - \bar{X})^2} = \frac{\sum_{i=2}^{2} Y_i(X_i - \bar{X}) - \sum_{i=2}^{2} \bar{Y}(X_i - \bar{X})}{\sum_{i=2}^{2}(X_i - \bar{X})^2}$$

$$= \frac{\sum_{i=2}^{2} Y_i(X_i - \bar{X}) - \bar{Y}(\sum_{i=2}^{2} X_i - \sum_{i=2}^{2} \bar{X})}{\sum_{i=2}^{2}(X_i - \bar{X})^2} = \frac{\sum_{i=2}^{2} Y_i(X_i - \bar{X}) - \bar{Y}(2\bar{X} - 2\bar{X})}{\sum_{i=2}^{2}(X_i - \bar{X})^2}$$

$$\widehat{w_2} = \frac{\sum_{i=2}^{2} Y_i(X_i - \bar{X})}{\sum_{i=2}^{2}(X_i - \bar{X})^2}$$

נגדיר:

$$c_i = \frac{X_i - \bar{X}}{\sum_{i=2}^{2}(X_i - \bar{X})^2}$$

$$\sum_{i=2}^{2} c_i = \sum_{i=2}^{2} \frac{X_i - \bar{X}}{\sum_{i=2}^{2}(X_i - \bar{X})^2} = \frac{\sum_{i=2}^{2} X_i - \sum_{i=2}^{2} \bar{X}}{\sum_{i=2}^{2}(X_i - \bar{X})^2} = \frac{2\bar{X} - 2\bar{X}}{\sum_{i=2}^{2}(X_i - \bar{X})^2} = 0$$

$$\sum_{i=2}^{2} c_i^2 = \frac{\sum_{i=2}^{2}(X_i - \bar{X})^2}{\left(\sum_{i=2}^{2}(X_i - \bar{X})^2\right)^2} = \frac{1}{\sum_{i=2}^{2}(X_i - \bar{X})^2}$$

$$\sum_{i=2}^{2} c_i X_i = \frac{\sum_{i=2}^{2}(X_i - \bar{X})X_i}{\sum_{i=2}^{2}(X_i - \bar{X})^2} = \frac{\sum_{i=2}^{2} X_i^2 - \sum_{i=2}^{2} \bar{X}X_i}{\sum_{i=2}^{2}(X_i^2 - 2X_i\bar{X} + \bar{X}^2)} = \frac{\sum_{i=2}^{2} X_i^2 - 2\bar{X}^2}{\sum_{i=2}^{2} X_i^2 - 4\bar{X}^2 + 2\bar{X}^2} = 1$$

<u>נבדוק עבור $\widehat{w_2}$:</u>

נציב ב-

$$\widehat{w_2} = \frac{\sum_{i=2}^{2} Y_i(X_i - \bar{X})}{\sum_{i=2}^{2}(X_i - \bar{X})^2}$$

ונקבל:

$$\widehat{w_2} = \sum_{i=2}^{2} Y_i c_i = \sum_{i=2}^{2} c_i\,(w_1 + w_2 X_i + \epsilon_i) = w_1 \sum_{i=2}^{2} c_i + w_2 \sum_{i=2}^{2} c_i\, X_i + \sum_{i=2}^{2} c_i \epsilon_i$$

$$= w_1 * 0 + w_2 * 1 + \sum_{i=2}^{2} c_i \epsilon_i = w_2 + \sum_{i=2}^{2} c_i \epsilon_i$$

תחת ההנחה ש-c_i אינו סטוכסטי נקבל שהתוחלת של $(\widehat{w_2})$ היא פונקציה לינארית של השגיאות:

$$E(\widehat{w_2}) = E\left(w_2 + \sum_{i=2}^{2} c_i \epsilon_i\right) = E(w_2) + E\left(\sum_{i=2}^{2} c_i \epsilon_i\right) = w_2 + \sum_{i=2}^{2} c_i E(\epsilon_i)$$

תחת ההנחה ש $E(\epsilon_i) = 0$ נקבל:

$$E(\widehat{w_2}) = w_2$$

<u>נבדוק עבור $\widehat{w_1}$:</u>

$$E(\widehat{w_1}) = E(\bar{Y} - \widehat{w_2}\bar{X}) = E(\bar{Y}) - E(\widehat{w_2}\bar{X})$$

תחת ההנחה ש $\bar{X}$ אינו סטוכסטי

$$= w_1 + w_2\bar{X} - \bar{X}E(\widehat{w_2}) = w_1 + w_2\bar{X} - w_2\bar{X} = w_1$$

HW1

November 22, 2018

1 Q2 - Linear Regression

In [347]:
```
import pandas as pd

from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
%config InlineBackend.figure_format = 'retina'
```

1.1 a

In [348]:
```
df = pd.read_csv('parkinsons_updrs_data.csv')
```

In [349]:
```
df.head()
df.columns
```

Out[349]:
```
   subject.ID  age  sex  test_time  motor_UPDRS  total_UPDRS  Jitter.Per  \
0           1   72    0     5.6431       28.199       34.398     0.00662
1           1   72    0    12.6660       28.447       34.894     0.00300
2           1   72    0    19.6810       28.695       35.389     0.00481
3           1   72    0    25.6470       28.905       35.810     0.00528
4           1   72    0    33.6420       29.187       36.375     0.00335

   Jitter.Abs  Jitter.RAP  Jitter.PPQ5    ...     Shimmer.dB  Shimmer.APQ3  \
0    0.000034     0.00401      0.00317    ...          0.230       0.01438
1    0.000017     0.00132      0.00150    ...          0.179       0.00994
2    0.000025     0.00205      0.00208    ...          0.181       0.00734
3    0.000027     0.00191      0.00264    ...          0.327       0.01106
4    0.000020     0.00093      0.00130    ...          0.176       0.00679

   Shimmer.APQ5  Shimmer.APQ11  Shimmer.DDA       NHR     HNR     RPDE  \
0       0.01309        0.01662      0.04314  0.014290  21.640  0.41888
1       0.01072        0.01689      0.02982  0.011112  27.183  0.43493
2       0.00844        0.01458      0.02202  0.020220  23.047  0.46222
3       0.01265        0.01963      0.03317  0.027837  24.445  0.48730
4       0.00929        0.01819      0.02036  0.011625  26.126  0.47188

       DFA      PPE
0  0.54842  0.16006
```

```
1  0.56477  0.10810
2  0.54405  0.21014
3  0.57794  0.33277
4  0.56122  0.19361

[5 rows x 22 columns]
```

Out[349]:
```
Index(['subject.ID', 'age', 'sex', 'test_time', 'motor_UPDRS', 'total_UPDRS',
       'Jitter.Per', 'Jitter.Abs', 'Jitter.RAP', 'Jitter.PPQ5', 'Jitter.DDP',
       'Shimmer', 'Shimmer.dB', 'Shimmer.APQ3', 'Shimmer.APQ5',
       'Shimmer.APQ11', 'Shimmer.DDA', 'NHR', 'HNR', 'RPDE', 'DFA', 'PPE'],
      dtype='object')
```

1.2 b

NOTES TO MYSELF

The data contains data for 42 patients (subjects). For these 42 patients we have 5,875 voice recordings.

general data: subject.ID - Integer that uniquely identifies each subject age - Subject age sex - Subject gender '0' - male, '1' - female test_time - Time since recruitment into the trial. The integer part is the number of days since recruitment.

what we want to predict: motor_UPDRS - Clinician's motor UPDRS score, linearly interpolated total_UPDRS - Clinician's total UPDRS score, linearly interpolated

measuremets (16 biomedical voice measures): Jitter.Per,Jitter.Abs,Jitter.RAP,Jitter.PPQ5,Jitter.DDP - Several measures of variation in fundamental frequency Shimmer,Shimmer.dB,Shimmer.APQ3,Shimmer.APQ5,Shimmer.APQ11,Shimmer.DDA - Several measures of variation in amplitude NHR,HNR - Two measures of ratio of noise to tonal components in the voice RPDE - A nonlinear dynamical complexity measure DFA - Signal fractal scaling exponent PPE - A nonlinear measure of fundamental frequency variation

In [350]:
```
df.describe()
```

Out[350]:
```
         subject.ID          age          sex    test_time  motor_UPDRS  \
count  5875.000000  5875.000000  5875.000000  5875.000000  5875.000000
mean     21.494128    64.804936     0.317787    92.863722    21.296229
std      12.372279     8.821524     0.465656    53.445602     8.129282
min       1.000000    36.000000     0.000000    -4.262500     5.037700
25%      10.000000    58.000000     0.000000    46.847500    15.000000
50%      22.000000    65.000000     0.000000    91.523000    20.871000
75%      33.000000    72.000000     1.000000   138.445000    27.596500
max      42.000000    85.000000     1.000000   215.490000    39.511000

        total_UPDRS   Jitter.Per   Jitter.Abs   Jitter.RAP  Jitter.PPQ5  \
count  5875.000000  5875.000000  5875.000000  5875.000000  5875.000000
mean     29.018942     0.006154     0.000044     0.002987     0.003277
std      10.700283     0.005624     0.000036     0.003124     0.003732
min       7.000000     0.000830     0.000002     0.000330     0.000430
25%      21.371000     0.003580     0.000022     0.001580     0.001820
50%      27.576000     0.004900     0.000035     0.002250     0.002490
```

```
75%      36.399000     0.006800     0.000053     0.003290     0.003460
max      54.992000     0.099990     0.000446     0.057540     0.069560

           ...      Shimmer.dB  Shimmer.APQ3  Shimmer.APQ5  Shimmer.APQ11  \
count      ...     5875.000000   5875.000000   5875.000000    5875.000000
mean       ...        0.310960      0.017156      0.020144       0.027481
std        ...        0.230254      0.013237      0.016664       0.019986
min        ...        0.026000      0.001610      0.001940       0.002490
25%        ...        0.175000      0.009280      0.010790       0.015665
50%        ...        0.253000      0.013700      0.015940       0.022710
75%        ...        0.365000      0.020575      0.023755       0.032715
max        ...        2.107000      0.162670      0.167020       0.275460

       Shimmer.DDA          NHR          HNR         RPDE          DFA  \
count  5875.000000  5875.000000  5875.000000  5875.000000  5875.000000
mean      0.051467     0.032120    21.679495     0.541473     0.653240
std       0.039711     0.059692     4.291096     0.100986     0.070902
min       0.004840     0.000286     1.659000     0.151020     0.514040
25%       0.027830     0.010955    19.406000     0.469785     0.596180
50%       0.041110     0.018448    21.920000     0.542250     0.643600
75%       0.061735     0.031463    24.444000     0.614045     0.711335
max       0.488020     0.748260    37.875000     0.966080     0.865600

               PPE
count  5875.000000
mean      0.219589
std       0.091498
min       0.021983
25%       0.156340
50%       0.205500
75%       0.264490
max       0.731730

[8 rows x 22 columns]
```

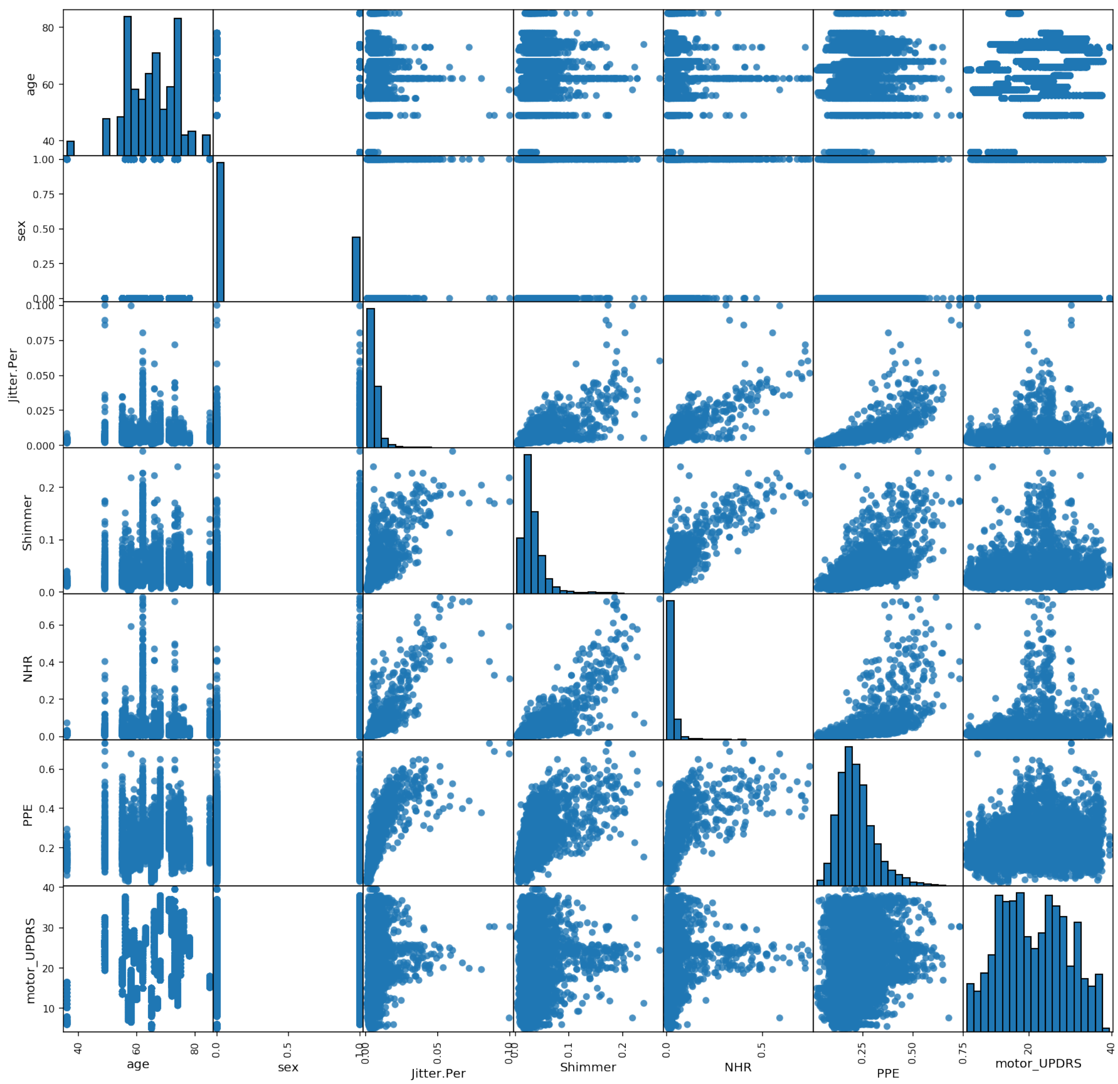
age of subjects - the mean age of the subjects is 64.8 years old, with a standard deviation of 8.8 years. The youngest subject is 36 years old while the oldest is 85.

gender of subjects - since 0 represents a male and 1 - a female, we can see by the “mean sex”, that about 32% of the subjects are females, meaning the majuraty (68%) are males.

1.3 c

```
In [351]: df_6 = df[['age','sex','Jitter.Per','Shimmer','NHR','PPE','motor_UPDRS']]
```

```
In [352]: scatter_plot = pd.plotting.scatter_matrix(df_6, alpha=0.8,s=30,figsize=(15,15),
                                                    ax=None, grid=True, diagonal='hist',
                                                    marker='o', density_kwds=None,
                                                    hist_kwds={'bins':20,'edgecolor':'black'},
                                                    range_padding=0.05)
```



1.4 d

```
In [353]: def leastSquares(X,y): #X and y are numpy arrays

              estimators = np.matmul(np.matmul(np.linalg.inv(np.matmul(np.transpose(X),X)),
                                                 np.transpose(X)),y)

              return estimators
```

1.5 e

```
In [354]: import numpy as np
          import statsmodels.api as sm
```

```
X = df_6[['age','sex','Jitter.Per','Shimmer','NHR','PPE']].values
y = df_6['motor_UPDRS'].values
X2 = sm.add_constant(X)
# np.shape(X)
# np.shape(y)

w = leastSquares(X2,y)
```

1.6 f

```
In [355]: # python's linear regression (sklearn)
          # from sklearn import linear_model
          # import statsmodels.api as sm
          # X2 = sm.add_constant(X)
          # reg = linear_model.LinearRegression()
          # reg2 = reg.fit(X2, y)
          # print(reg2.coef_)

          # python's linear regression (statsmodels)
          import statsmodels.api as sm

          X2 = sm.add_constant(X)
          reg = sm.OLS(y, X2)
          reg2 = reg.fit()
```

Yes, I got the same values. The following table summarizes the results.

```
In [356]: data = {'My_model': w,
                  'Python_model': reg2.params}
          table = pd.DataFrame(data)

          # The following table compares both models' estimators
          print(table)
```

```
     My_model  Python_model
0    3.445982      3.445982
1    0.236592      0.236592
2   -0.160868     -0.160868
3 -101.125995   -101.125995
4   -4.648934     -4.648934
5    7.366135      7.366135
6   14.176249     14.176249
```

1.7 g

We can use a t-test, provided by the the linear regression model from statsmodels: