Multiple Linear Regression

Using scikit-learn to implement Multiple Linear Regression. Check how Engine Size is related to Co2Emissions. Create a model using train set,test using test set .

import matplotlib.pyplot as plt import pandas as pd import pylab as pl import numpy as np %matplotlib inline

In [4]:
```
df = pd.read_csv("FuelConsumptionCo2.csv")
#df.head()
#df.describe()
```

Creating train and test dataset:

Train/Test Split dataset to mutually exclusive. We can use 80% of the entire data for training, and the 20% for testing. We create a mask using np.random.rand().

In [5]:
```
cdf = df[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_CITY','FUELCONSUMPTION_
cdf.head(9)
```

Out[5]:

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_CITY	FUELCONSUMPTION_HWY	FUELCONSUMP
0	2.0	4	9.9	6.7	
1	2.4	4	11.2	7.7	
2	1.5	4	6.0	5.8	
3	3.5	6	12.7	9.1	
4	3.5	6	12.1	8.7	
5	3.5	6	11.9	7.7	
6	3.5	6	11.8	8.1	
7	3.7	6	12.8	9.0	
8	3.7	6	13.4	9.5	

In [7]:
```
msk = np.random.rand(len(df)) < 0.8
train = cdf[msk]
test = cdf[~msk]
```

Multiple Linear Regression

Multiple Linear Regression is an extension of linear regresssion and multiple variables are used to predict Co2Emissions. Co2Emissions are predicted using FuelConsumption-Comp ,Engine size and cylinders of car Creating model using Training Set

In [8]:
```
from sklearn import linear_model
regr = linear_model.LinearRegression()
x = np.asanyarray(train[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_COMB']])
y = np.asanyarray(train[['CO2EMISSIONS']])
regr.fit (x, y)
# The coefficients
print ('Coefficients: ', regr.coef_)
```

```
Coefficients:  [[11.80178358  7.36961366  9.24897056]]
```

In [9]:
```
y_hat= regr.predict(test[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_COMB']]
x = np.asanyarray(test[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_COMB']])
y = np.asanyarray(test[['CO2EMISSIONS']])
print("Residual sum of squares: %.2f"
      % np.mean((y_hat - y) ** 2))

# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(x, y))
```

```
Residual sum of squares: 580.69
Variance score: 0.85
```

Creating another multiple linear regression model with the same dataset using FUEL CONSUMPTION in CITY and FUEL CONSUMPTION in HWY instead of FUELCONSUMPTION_COMB.

In [11]:
```
regr = linear_model.LinearRegression()
x = np.asanyarray(train[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_CITY','F
y = np.asanyarray(train[['CO2EMISSIONS']])
regr.fit (x, y)
print ('Coefficients: ', regr.coef_)
y_= regr.predict(test[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_CITY','FUE
x = np.asanyarray(test[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_CITY','FU
y = np.asanyarray(test[['CO2EMISSIONS']])
print("Residual sum of squares: %.2f"% np.mean((y_ - y) ** 2))
print('Variance score: %.2f' % regr.score(x, y))
```

```
Coefficients:  [[11.82627841  7.2316827   5.41337158  3.73318229]]
Residual sum of squares: 580.31
Variance score: 0.85
```

In [ ]: