Module 5 - Exercise 2.1 Filipp Krasovsky 11-28-2020

In [1]:
```
import pymysql.cursors as sql
import pymysql
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import statsmodels.api as sm
```

In [48]:
```
# 2.1 Use Python to explore the relationship of different variables to models per gallon
# (mpg). Find out which of the variables have high correlation with mpg. Report those
# values. Build a regression model using one of those variables to predict mpg. Do the
# same using two of those variables. Report your models along with the regression line
# equations. (10 points)
```

In [2]:
```
#connect to DB
connection = pymysql.connect(
    host='localhost',
    user='root',
    password='admin',
    db='500b',
    charset='utf8mb4',
    cursorclass=sql.DictCursor)

#Create a function to simplify the query process.

try:
    with connection.cursor() as cursor:
        query = "SELECT * FROM mpg"
        cursor.execute(query)
        df = pd.DataFrame(cursor.fetchall())
finally:
    connection.close()
```

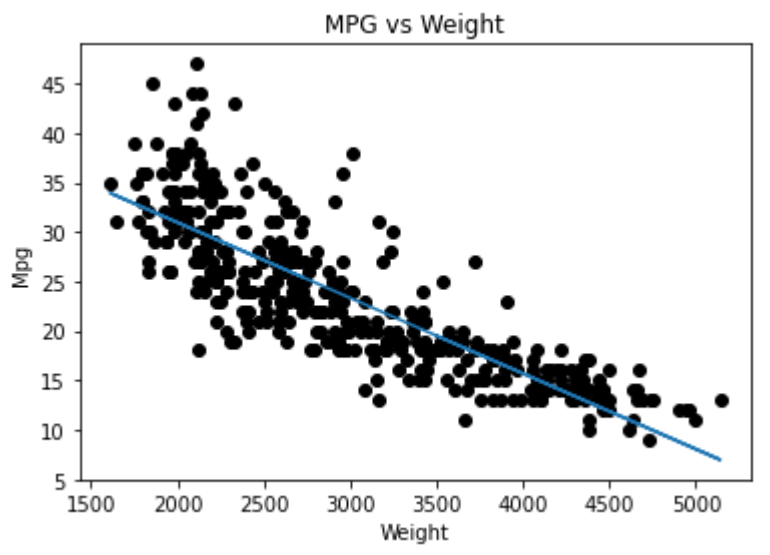
In [3]:
```
#conclusion: displacement and weight seem to have the strongest association with mpg
(df.corr()["mpg"])
```

Out[3]:
```
mpg             1.000000
cylinders      -0.776796
displacement   -0.804304
horsepower     -0.777683
weight         -0.831535
model year      0.582750
origin          0.563667
Name: mpg, dtype: float64
```

In [4]:
```
#model 1: we regress mpg against weight.
Y = df["mpg"]
X = df["weight"]
X = sm.add_constant(X)

lr_model_1 = sm.OLS(Y,X).fit()
print(lr_model_1.summary())

plt.scatter(df["weight"],Y,color='black')
plt.title("MPG vs Weight")
plt.xlabel("Weight")
plt.ylabel("Mpg")
m, b = np. polyfit(df["weight"], Y, 1)
plt. plot(df["weight"], m*df["weight"] + b)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.691
Model:                            OLS   Adj. R-squared:                  0.691
Method:                 Least Squares   F-statistic:                     874.0
Date:                Sat, 28 Nov 2020   Prob (F-statistic):          1.27e-101
Time:                        14:16:40   Log-Likelihood:                -1130.5
No. Observations:                 392   AIC:                             2265.
Df Residuals:                     390   BIC:                             2273.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         46.2287      0.800     57.809      0.000      44.657      47.801
weight        -0.0076      0.000    -29.563      0.000      -0.008      -0.007
==============================================================================
Omnibus:                       40.521   Durbin-Watson:                   0.821
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               58.131
Skew:                           0.712   Prob(JB):                     2.38e-13
Kurtosis:                       4.239   Cond. No.                     1.13e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.13e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Out[4]:
```
[<matplotlib.lines.Line2D at 0x9e27c70>]
```



In [82]:
```
#model 2: we regress mpg against weight and displacement.
Y = df["mpg"]
X = df[["weight","displacement"]]
X = sm.add_constant(X)

lr_model_2 = sm.OLS(Y,X).fit()
print(lr_model_2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.698
Model:                            OLS   Adj. R-squared:                  0.696
Method:                 Least Squares   F-statistic:                     448.9
Date:                Sat, 28 Nov 2020   Prob (F-statistic):          8.72e-102
Time:                        12:54:27   Log-Likelihood:                -1126.4
No. Observations:                 392   AIC:                             2259.
Df Residuals:                     389   BIC:                             2271.
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           43.8052      1.165     37.610      0.000      41.515      46.095
weight          -0.0058      0.001     -8.086      0.000      -0.007      -0.004
displacement    -0.0164      0.006     -2.840      0.005      -0.028      -0.005
==============================================================================
Omnibus:                       44.856   Durbin-Watson:                   0.845
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               67.316
Skew:                           0.753   Prob(JB):                     2.41e-15
Kurtosis:                       4.362   Cond. No.                     1.66e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.66e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The equation for the two lines are as follows:

model 1: mpg = 46.2287 - 0.0076 *weight*
model 2: mpg = 43.8052 - 0.0058 weight - 0.0164 * displacement