# MACHINE LEARNING - ASSIGNMENT 1

**GUNA PRAGNA**
**S987P265**

---

**Q1 Fit a predictive linear regression model to estimate weight of the fish from its length, height and width? (the data source fish.csv can be found here: https://www.kaggle.com/aungpyaeap/fish-market)**
**-Report the coefficients values by using the standard Least Square Estimates**

1) Uploading the dataset Fish and showing the lengths, height and width variables.

In [1]:
```
import matplotlib.pyplot as plt
import numpy as np
from sklearn import datasets, linear_model
import pandas as pd
import statsmodels.api as sm


df=pd.read_excel("/Users/gunapragna/Downloads/Fish.xls")
```

In [12]:
```
X= df [['Length1','Length2','Length3','Height','Width']]

Y= df["Weight"] #we need to predict this weight
```

In [51]:
```
X
```

Out[51]:

| | Length1 | Length2 | Length3 | Height | Width |
|---|---|---|---|---|---|
| **0** | 23.2 | 25.4 | 30.0 | 11.5200 | 4.0200 |
| **1** | 24.0 | 26.3 | 31.2 | 12.4800 | 4.3056 |
| **2** | 23.9 | 26.5 | 31.1 | 12.3778 | 4.6961 |
| **3** | 26.3 | 29.0 | 33.5 | 12.7300 | 4.4555 |
| **4** | 26.5 | 29.0 | 34.0 | 12.4440 | 5.1340 |
| **...** | ... | ... | ... | ... | ... |
| **154** | 11.5 | 12.2 | 13.4 | 2.0904 | 1.3936 |
| **155** | 11.7 | 12.4 | 13.5 | 2.4300 | 1.2690 |
| **156** | 12.1 | 13.0 | 13.8 | 2.2770 | 1.2558 |
| **157** | 13.2 | 14.3 | 15.2 | 2.8728 | 2.0672 |
| **158** | 13.8 | 15.0 | 16.2 | 2.9322 | 1.8792 |

159 rows × 5 columns

2) Splitting the data into texting and training data sets
3) Fitting the Linear Regression and finding the intercept, coefficients

```
In [69]: from sklearn.linear_model import LinearRegression #fitting linear regression
         lin_reg = LinearRegression()
         lin_reg.fit(fish_X_train,fish_Y_train)

Out[69]: LinearRegression()

In [70]: intercept = lin_reg.intercept_
         intercept #the intercept for linear regression

Out[70]: -568.7012643903099

In [71]: coefficients = lin_reg.coef_
         coefficients #the coeffiecients for linear regression

Out[71]: array([  5.23237318,  47.90736985, -40.34550533,  35.92155964,
                 82.39780489])
```

```
fish_X_test = X[-20:]
fish_Y_train = Y[:-20]
fish_Y_test = Y[-20:]
```

## -What is the standard error of the estimated coefficients, R-squared term, and the 95% confidence interval?

- Calculating the standard error with estimated coefficients for length1, length2, length3 , height and width
- R-squared term (uncentered ) for the linear regression is 0.854 for Fish Dataset.

```
In [73]: results.summary()
```

Out[73]:

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Weight | **R-squared (uncentered):** | 0.854 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.849 |
| **Method:** | Least Squares | **F-statistic:** | 179.7 |
| **Date:** | Thu, 30 Sep 2021 | **Prob (F-statistic):** | 2.24e-62 |
| **Time:** | 10:02:42 | **Log-Likelihood:** | -1071.7 |
| **No. Observations:** | 159 | **AIC:** | 2153. |
| **Df Residuals:** | 154 | **BIC:** | 2169. |
| **Df Model:** | 5 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Length1** | 202.0690 | 66.391 | 3.044 | 0.003 | 70.914 | 333.224 |
| **Length2** | -89.5220 | 69.967 | -1.279 | 0.203 | -227.742 | 48.698 |
| **Length3** | -82.6718 | 28.784 | -2.872 | 0.005 | -139.534 | -25.809 |
| **Height** | 55.7740 | 14.470 | 3.854 | 0.000 | 27.188 | 84.360 |
| **Width** | -51.1129 | 33.577 | -1.522 | 0.130 | -117.444 | 15.218 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 59.832 | **Durbin-Watson:** | 0.423 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 143.316 |
| **Skew:** | 1.628 | **Prob(JB):** | 7.57e-32 |
| **Kurtosis:** | 6.322 | **Cond. No.** | 310. |

- 95% confidence interval

| | | |
|---|---|---|
| **Length1** | 70.914463 | 333.223631 |
| **Length2** | -227.741886 | 48.697795 |
| **Length3** | -139.534379 | -25.809164 |
| **Height** | 27.188025 | 84.359934 |
| **Width** | -117.443783 | 15.217952 |

**-Is there any dependence between the length and weight of the fish?**
Yes, the length and weight are interdependent with each other.

**Q2 Using the data source in Q1 fit the linear regression model using Stochastic Gradient Descent (SGD) optimizer.**

Stochastic gradient descent is the machine learning algorithm for finding the optimal parameter configuration and the error of the data network is decreased.
- Fitting SGD in Linear regression model
- Calculating the intercept and coefficients for Fish dataset.

```
In [74]: # stochastic gradient descent

         from sklearn.linear_model import SGDRegressor

In [17]: #sgd_reg = SGDRegressor(max_iter=1000, tol=1e-3, penalty=None, eta0=0.1, random_state=42)

In [86]: sgd_reg.fit(fish_X_train,fish_Y_train) #fitting linear regression into SGD

Out[86]: SGDRegressor(eta0=0.1, penalty=None, random_state=42)

In [87]: intercept = sgd_reg.intercept_ #finding the intercept of sgd
         intercept

Out[87]: array([-4.18482114e+08])

In [77]: coefficients = sgd_reg.coef_ #finding the intercept of sgd
         coefficients

Out[77]: array([-4.51100286e+11, -4.03579289e+11, -4.68025475e+11,  1.19724237e+12,
                -4.73582754e+11])
```

- **Report the difference in the obtained coefficient values due to SGD over Least Square as an optimizer.**

- The coefficient values in the SGD are exponential values but in the least square they are integer values.

- The coefficient estimates for Least Squares depend on the independence of the features. This coefficients are correlate with the variables in the SGD are dependent on the tolerance and the number of iterations performed where the least square does not depend on the.
- The coefficients and linear equations in the Linear Regression was quite simple. Using an iterative process, we will obtain a numerical approximation of this values that is close to the OLS solution, which gave us the exact solution. Whereas in SGD, We iterate step by step to find the best solution. We begin with arbitrary weight values and examine the gradient at the location.

**Q3 Using the data source in Q1 fit the Ridge and Lasso Regression Models.**

- **Report the coefficients for both the models**

```
In [79]: from sklearn.linear_model import Ridge
         ridge_reg = Ridge(alpha=1, solver="sag", random_state=42)
         ridge_reg.fit(fish_X_train, fish_Y_train) #fitting the ridge regression model
         ridge_reg.coef_ #calculating the coefficients

Out[79]: array([ 22.22715472,  31.12956146, -39.53060896,  36.90045452,
                 79.94890184])

In [83]: from sklearn.linear_model import Lasso
         lasso_reg = Lasso(alpha=0.01,max_iter = 10000)
         lasso_reg.fit(fish_X_train, fish_Y_train) #fitting the Lasso regression model
         lasso_reg.coef_ #calculating the coefficients

Out[83]: array([  5.59599589,  47.5204698 , -40.30663794,  35.92760158,
                 82.40712417])
```

- **Report the attribute(s) least impacting the weight of the fish**.

The attributes like width impact the weight of the fish in the least because the mod of the coefficient value of the width is the lowest and the highest is for the length1.