# Introduction to Big Data

## Lecture 10. Regression

```
from google.colab import drive
drive.mount('/content/drive')
```

```
import pandas as pd
import seaborn as sns
```

## Regression with Boston Housing Data

CRIM - per capita crime rate by town

ZN - proportion of residential land zoned for lots over 25,000 sq.ft.

INDUS - proportion of non-retail business acres per town.

CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)

NOX - nitric oxides concentration (parts per 10 million)

RM - average number of rooms per dwelling

AGE - proportion of owner-occupied units built prior to 1940

DIS - weighted distances to five Boston employment centres

RAD - index of accessibility to radial highways

TAX - full-value property-tax rate per $10,000

PTRATIO - pupil-teacher ratio by town

B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

LSTAT - % lower status of the population MEDV - Median value of owner-occupied homes in $1000's

```
from statsmodels.formula.api import ols
```

```
housing_df = pd.read_csv('[DIRECTORY]/HousingData.csv')
housing_df
```

```
sns.pairplot(housing_df[['MEDV','CRIM','LSTAT']])
```

## Simple Regression

```
statsOLSModel = ols('MEDV ~ CRIM', data=housing_df)
statsOLSModel
```

```
statsOLSModel_res = statsOLSModel.fit()
statsOLSModel_res
```

```
print(statsOLSModel_res.summary())
```

```
statsOLSModel_res.params
```

## Visualization

```
sns.lmplot(x='CRIM',y='MEDV', data=housing_df,
           scatter_kws = {'color':'red'}, line_kws={'color':'black'})
```

## Multiple Regression

```
statsOLSModel_all = ols('MEDV ~ CRIM+ZN+INDUS+CHAS+NOX+RM+AGE+DIS+RAD+TAX+PTRATIO+B+LSTAT',
                        data=housing_df)
statsOLSModel_all
```

```
statsOLSModel_all_res = statsOLSModel_all.fit()
statsOLSModel_all_res
```

```
print(statsOLSModel_all_res.summary())
```

```
housing_df.corr()
```

✓ 0초 오전 9:20에 완료됨