Introduction to Big Data

Lecture 10. Regression

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

```
Mounted at /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
```

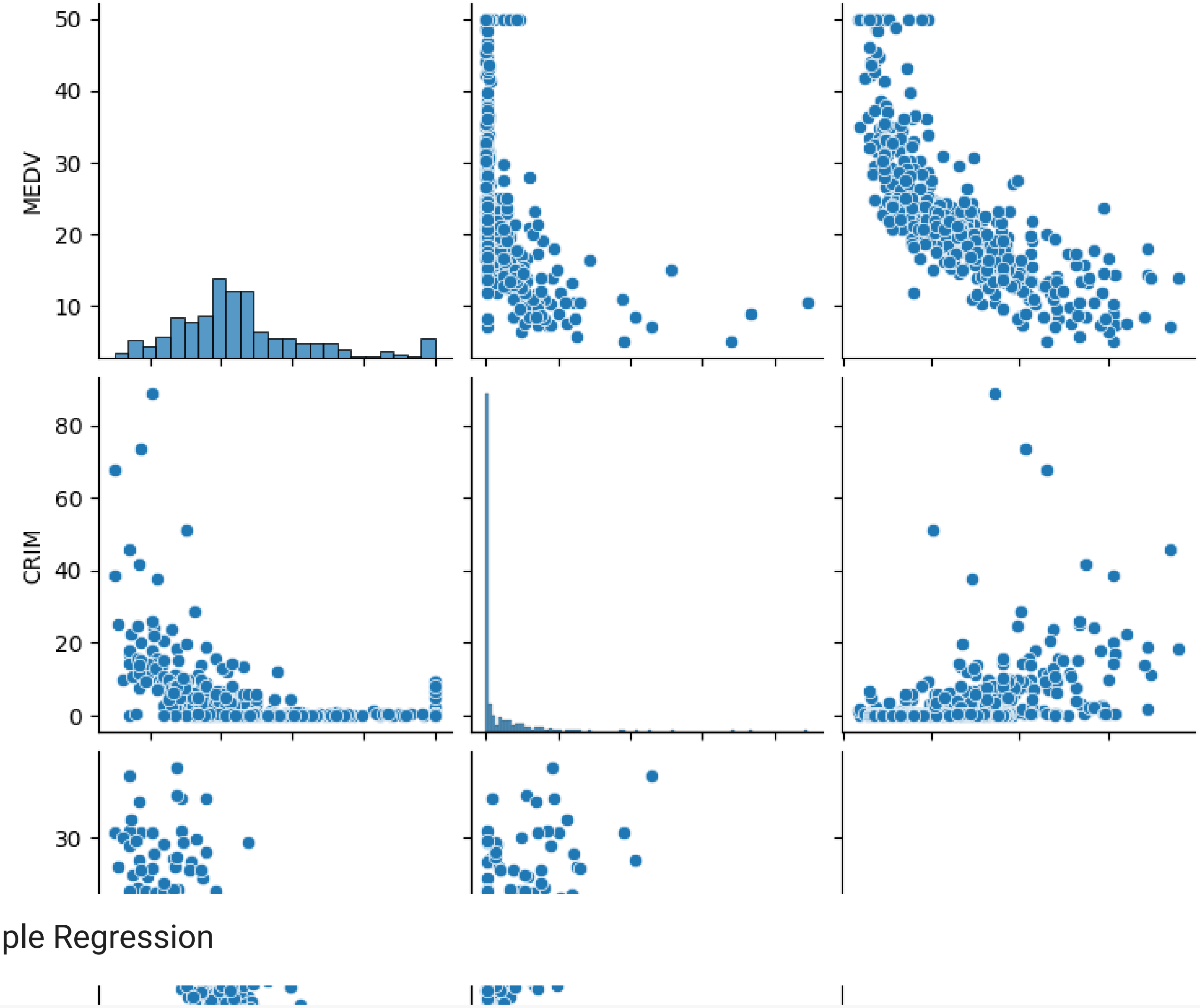
```
housing_df = pd.read_csv('/content/drive/MyDrive/[Lecture]/IntBigData/BigData_Python/10_Regression/HousingData
housing_df
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	NaN	36.2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
501	0.06263	0.0	11.93	0.0	0.573	6.593	69.1	2.4786	1	273	21.0	391.99	NaN	22.4
502	0.04527	0.0	11.93	0.0	0.573	6.120	76.7	2.2875	1	273	21.0	396.90	9.08	20.6
503	0.06076	0.0	11.93	0.0	0.573	6.976	91.0	2.1675	1	273	21.0	396.90	5.64	23.9
504	0.10959	0.0	11.93	0.0	0.573	6.794	89.3	2.3889	1	273	21.0	393.45	6.48	22.0
505	0.04741	0.0	11.93	0.0	0.573	6.030	NaN	2.5050	1	273	21.0	396.90	7.88	11.9

506 rows × 14 columns

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

<seaborn.axisgrid.PairGrid at 0x7a0951a37490>



Simple Regression

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

<statsmodels.regression.linear_model.OLS at 0x7a090de214b0>

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

```
<statsmodels.regression.linear_model.RegressionResultsWrapper at 0x7a090b7bedd0>
```

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

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   MEDV   R-squared:                       0.153
Model:                            OLS   Adj. R-squared:                  0.151
Method:                 Least Squares   F-statistic:                     87.54
Date:                Thu, 23 Nov 2023   Prob (F-statistic):           3.08e-19
Time:                        13:11:01   Log-Likelihood:                -1722.2
No. Observations:                 486   AIC:                             3448.
Df Residuals:                     484   BIC:                             3457.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     23.8792      0.412     57.978      0.000      23.070      24.689
CRIM          -0.4086      0.044     -9.356      0.000      -0.494      -0.323
==============================================================================
Omnibus:                      137.385   Durbin-Watson:                   0.764
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              296.868
Skew:                           1.505   Prob(JB):                     3.44e-65
Kurtosis:                       5.367   Cond. No.                         10.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

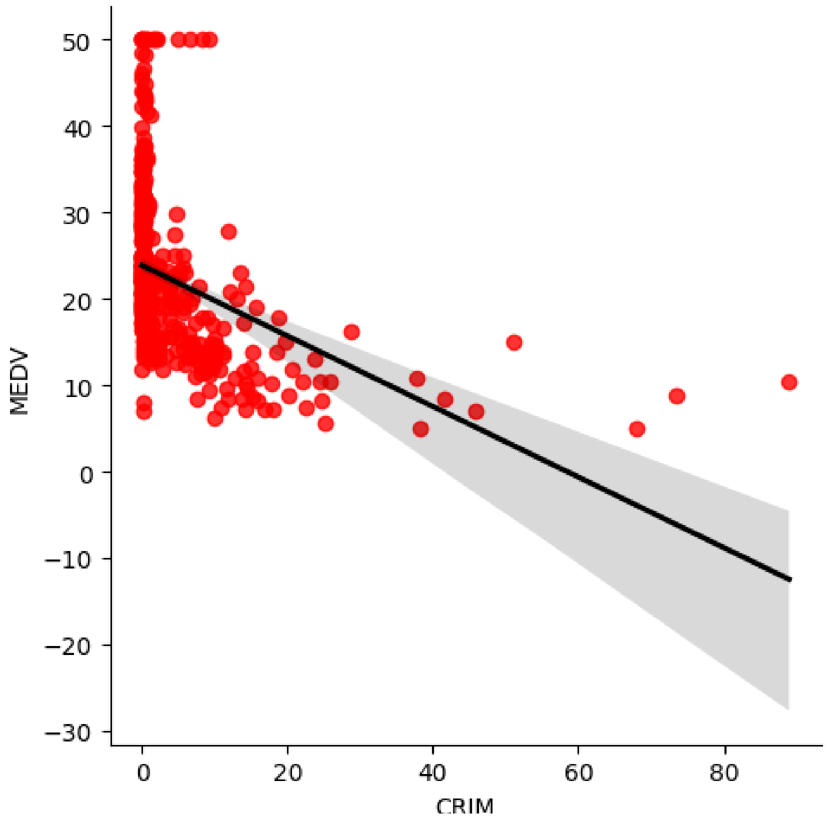
```
statsOLSModel_res.params
```

```
Intercept    23.879229
CRIM         -0.408635
dtype: float64
```

Visualization

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

<seaborn.axisgrid.FacetGrid at 0x7a090b600e50>



Multiple Regression

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

```
<statsmodels.regression.linear_model.OLS at 0x7a090b3480d0>
```

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

```
<statsmodels.regression.linear_model.RegressionResultsWrapper at 0x7a090b3a2500>
```

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

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   MEDV   R-squared:                       0.767
Model:                            OLS   Adj. R-squared:                  0.759
Method:                 Least Squares   F-statistic:                     96.29
Date:                Thu, 23 Nov 2023   Prob (F-statistic):          1.75e-111
Time:                        13:11:14   Log-Likelihood:                -1143.4
No. Observations:                 394   AIC:                             2315.
Df Residuals:                     380   BIC:                             2370.
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     32.6801      5.681      5.752      0.000      21.509      43.851
CRIM          -0.0976      0.032     -3.007      0.003      -0.161      -0.034
ZN             0.0489      0.014      3.397      0.001       0.021       0.077
INDUS          0.0304      0.066      0.461      0.645      -0.099       0.160
CHAS           2.7694      0.925      2.993      0.003       0.950       4.588
NOX          -17.9690      4.243     -4.235      0.000     -26.311      -9.627
RM             4.2833      0.471      9.100      0.000       3.358       5.209
AGE           -0.0130      0.014     -0.898      0.370      -0.041       0.015
DIS           -1.4585      0.211     -6.912      0.000      -1.873      -1.044
RAD            0.2859      0.069      4.125      0.000       0.150       0.422
TAX           -0.0131      0.004     -3.324      0.001      -0.021      -0.005
PTRATIO       -0.9146      0.141     -6.506      0.000      -1.191      -0.638
B              0.0097      0.003      3.251      0.001       0.004       0.015
LSTAT         -0.4237      0.055     -7.700      0.000      -0.532      -0.315
==============================================================================
Omnibus:                      161.243   Durbin-Watson:                   1.247
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              904.814
Skew:                           1.657   Prob(JB):                    3.33e-197
Kurtosis:                       9.643   Cond. No.                     1.57e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.57e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
housing_df.corr()
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
CRIM	1.000000	-0.191178	0.401863	-0.054355	0.417130	-0.219150	0.354342	-0.374166	0.624765	0.580595	0.281110	-0.381411	0.444943
ZN	-0.191178	1.000000	-0.531871	-0.037229	-0.513704	0.320800	-0.563801	0.656739	-0.310919	-0.312371	-0.414046	0.171303	-0.414193
INDUS	0.401863	-0.531871	1.000000	0.059859	0.764866	-0.390234	0.638431	-0.711709	0.604533	0.731055	0.390954	-0.360532	0.590690
CHAS	-0.054355	-0.037229	0.059859	1.000000	0.075097	0.104885	0.078831	-0.093971	0.001468	-0.032304	-0.111304	0.051264	-0.047424
NOX	0.417130	-0.513704	0.764866	0.075097	1.000000	-0.302188	0.731548	-0.769230	0.611441	0.668023	0.188933	-0.380051	0.582641
RM	-0.219150	0.320800	-0.390234	0.104885	-0.302188	1.000000	-0.247337	0.205246	-0.209847	-0.292048	-0.355501	0.128069	-0.614339
AGE	0.354342	-0.563801	0.638431	0.078831	0.731548	-0.247337	1.000000	-0.744844	0.458349	0.509114	0.269226	-0.275303	0.602891
DIS	-0.374166	0.656739	-0.711709	-0.093971	-0.769230	0.205246	-0.744844	1.000000	-0.494588	-0.534432	-0.232471	0.291512	-0.493328
RAD	0.624765	-0.310919	0.604533	0.001468	0.611441	-0.209847	0.458349	-0.494588	1.000000	0.910228	0.464741	-0.444413	0.479541
TAX	0.580595	-0.312371	0.731055	-0.032304	0.668023	-0.292048	0.509114	-0.534432	0.910228	1.000000	0.460853	-0.441808	0.536110
PTRATIO	0.281110	-0.414046	0.390954	-0.111304	0.188933	-0.355501	0.269226	-0.232471	0.464741	0.460853	1.000000	-0.177383	0.375966
B	-0.381411	0.171303	-0.360532	0.051264	-0.380051	0.128069	-0.275303	0.291512	-0.444413	-0.441808	-0.177383	1.000000	-0.369889
LSTAT	0.444943	-0.414193	0.590690	-0.047424	0.582641	-0.614339	0.602891	-0.493328	0.479541	0.536110	0.375966	-0.369889	1.000000
MEDV	-0.391363	0.373136	-0.481772	0.181391	-0.427321	0.695360	-0.394656	0.249929	-0.381626	-0.468536	-0.507787	0.333461	-0.735822