**Programming Assignment01**

|  |
| --- |
| **Submission guide**  1. Write answer following questions in this file  2. Write your code using provided Jupyter notebook file   * Do not import other packages that are not imported in the given file * After completing your code, run script and submit with the printed results for answering questions in this word file. |

1. Apply a multiple linear regression on the given dataset

The following code loads a dataset.

|  |
| --- |
| data=pd.read\_csv('https://drive.google.com/uc?export=download&id=1ssBNxmds4zmmJbAHzJUB0\_UyyfyMtoHT') |

The given dataset aims to predict crime rate(y) using several explanatory variables related with the unit regions.

[INPUT]

- M: percentage of males aged 14-24

- So: whether it is in a Southern state. 1 for Yes, 0 for No.

- Ed: mean years of schooling

- Po1: police expenditure in 1960

- Po2: police expenditure in 1959

- LF: labour force participation rate

- M.F: number of males per 1000 females

- Pop: state population

- NW: number of non-whites resident per 1000 people

- U1: unemployment rate of urban males aged 14-24

- U2: unemployment rate of urban males aged 35-39

- GDP: gross domestic product per head

- Ineq: income inequality

- Prob: probability of imprisonment

- Time: average time served in prisons

[OUTPUT]

- y: crime rate in an unspecified unit region

(1) Find the top 5 input variables that show the high linear correlation with the target based on the correlation coefficient. (5pts)

['Po1','Po2','GDP','Prob','Pop']

y

Po1 0.687604

Po2 0.666714

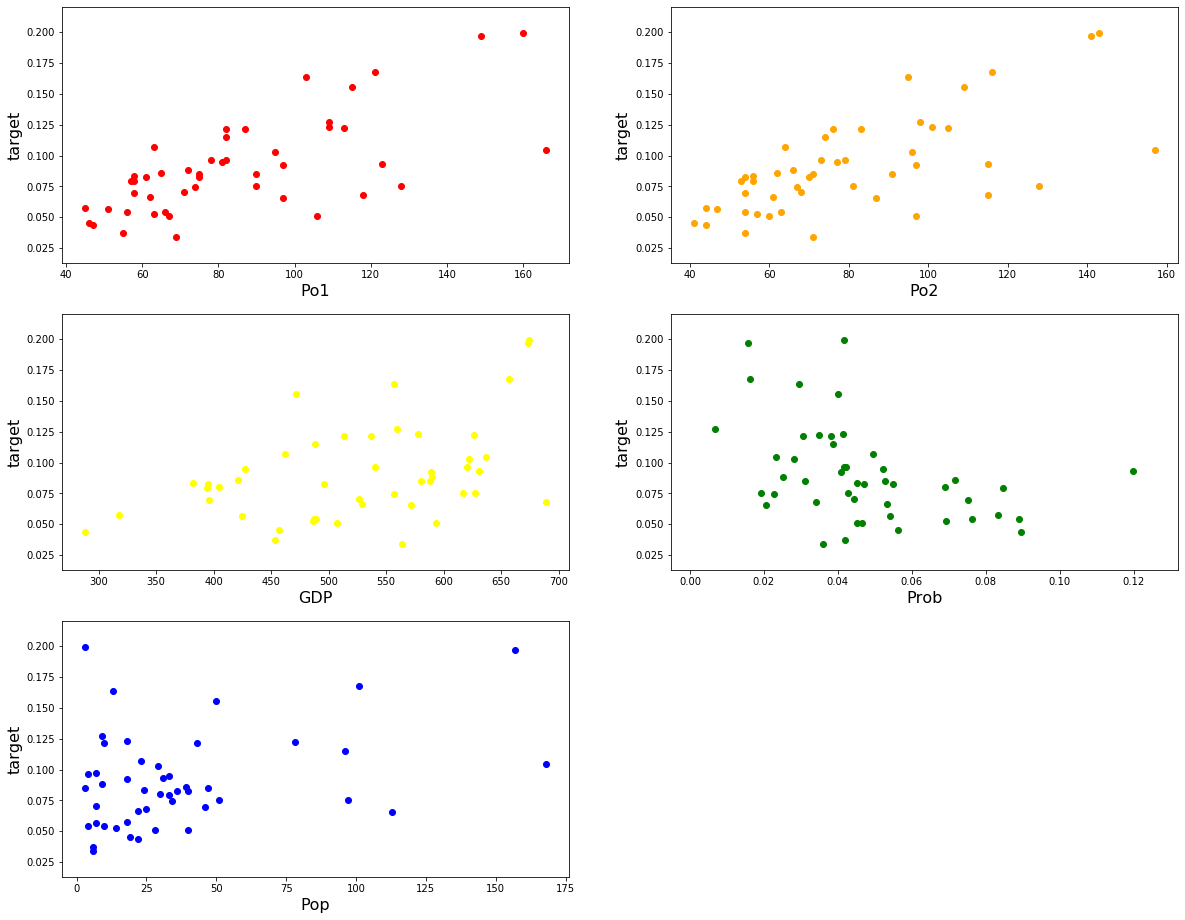
Pop 0.337474

GDP 0.441320

Prob -0.427422

y 1.000000

(2) Draw pairwise scatter plots – one scatter plot illustrates the relationship between the input variable selected in Question (1) and output target (Paste figures here) (5pts)



(3) Train a linear regression model (**M1**) using the selected variables in Question (1) and fill the

following table. (10pts)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Coefficient () |  |  | p-value |
| Intercept | 0.09228757 | 3.53310848e-02 | 2.61207852 | 0.01261335 |
| Po1 | 2.62474439e-03 | 1.22760388e-03 | 2.13810369 | 0.03866971 |
| Po2 | -1.49015707e-03 | 1.30985232e-03 | -1.13765273 | 0.26203435 |
| GDP | -1.53743853e-04 | 7.62150761e-05 | -2.01723675 | 0.05041619 |
| Prob | -4.13347529e-01 | 2.21041940e-01 | -1.86999593 | 0.06881562 |
| Pop | -1.38774123e-04 | 1.31692449e-04 | -1.05377434 | 0.29831023 |

(4) Calculate VIF for the variables of M1. Given that multicollinearity is severe when there is a variable with a VIF value of greater than 10, find the most reasonable way to get a better model based on the calculated VIF values. (10pts)

VIF1

Po1 80.348322

Po2 80.975316

GDP 3.264732

Prob 1.524780

Pop 1.517425

Po1 & po2 are bigger than 10. So I have to delete one of them. I delete Po2, because I think Po1 is more important Po2, Po1’s correlation coefficient to y is bigger than Po2.

So I remake VIF2 ['Po1','GDP','Prob', 'Pop']

VIF\_2

Po1 3.438328

GDP 3.187330

Prob 1.524200

Pop 1.512607

All value is less than 10. So this model is satisfied VIF.

(5) Based on the way you provide in **Question (4),** train a new regression model (**M2**) and create the same table for M2 as the table in Question (3).

Based on VIF values calculated in Question (4), build a new regression model.

(5pts)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Coefficient () |  |  | p-value |
| Intercept | 0.09580457 | 3.53216465e-02 | 2.71234723 | 9.71961612e-03 |
| Po1 | 1.25836611e-03 | 2.54856883e-04 | 4.93754022 | 1.37093413e-05 |
| GDP | -1.67094553e-04 | 7.55759282e-05 | -2.2109494 | 3.26751946e-02 |
| Prob | -4.18254557e-01 | 2.21791484e-01 | -1.8858008 | 6.64201624e-02 |
| Pop | -1.30331717e-04 | 1.31954177e-04 | -0.98770437 | 3.29090780e-01 |

(6) Describe difference between M1 and M2. (5pts)

For question 4, I calculated M1 VIF and remake model M2, because I could find a fector. So I recalculate this, Po1’s T-value is decreased a visible extent. Because elimination of 'po2' also increased the correlation between independent and continuous variables. Furthermore T-value is changed, P-value is also changed.

(7) Apply the F-test on M1 and M2 and explain the results. In addition, fill the following tables. (15pts)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| M1 | SS | Degree of freedom | MS | F | p-value |
| Model | 0.03832856278870676 | 5 | 0.007665712557741353 | 10.31124851133869 | 1.8875302661980342e-06 |
| Residual | 0.030480713807041322 | 41 | 0.0007434320440741786 | Null | Null |
| Total | 0.06880927659574468 | 46 | Null | Null | Null |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| M2 | SS | Degree of freedom | MS | F | p-value |
| Model | 0.03734231834640209 | 4 | 0.009335579586600522 | 12.460509831623407 | 9.071188887821435e-07 |
| Residual | 0.031466958249343015 | 42 | 0.0007434320440741786 | Null | Null |
| Total | 0.06880927659574468 | 46 | Null | Null | Null |

(8) Calculate and for M1 and M2. Then, compare two models. (7pts)

M1: 0.5570260971334449 0.5030048894667918

M2: 0.5426930814254629 0.49914004156122127

M2’s and are smaller than M1. So, M2 better describes the actual data.

(9) Calculate residuals of M1 and draw scatter plots to show relationship between one of the input variables and residuals. (8pts)

M1 Residual

0 0.018151

1 0.042103

2 0.001172

3 0.055426

4 0.003995

5 0.038904

6 0.019784

7 0.019875

8 0.014868

9 0.006580

10 0.052099

11 0.010148

12 0.027077

13 0.008651

14 0.011872

15 0.003787

16 0.009942

17 0.000028

18 0.051145

19 0.011536

20 0.012700

21 0.021857

22 0.022170

23 0.009751

24 0.015143

25 0.021421

26 0.030928

27 0.027074

28 0.058819

29 0.003936

30 0.031041

31 0.006442

32 0.039605

33 0.000125

34 0.039765

35 0.015105

36 0.002769

37 0.010880

38 0.004136

39 0.022243

40 0.007414

41 0.007879

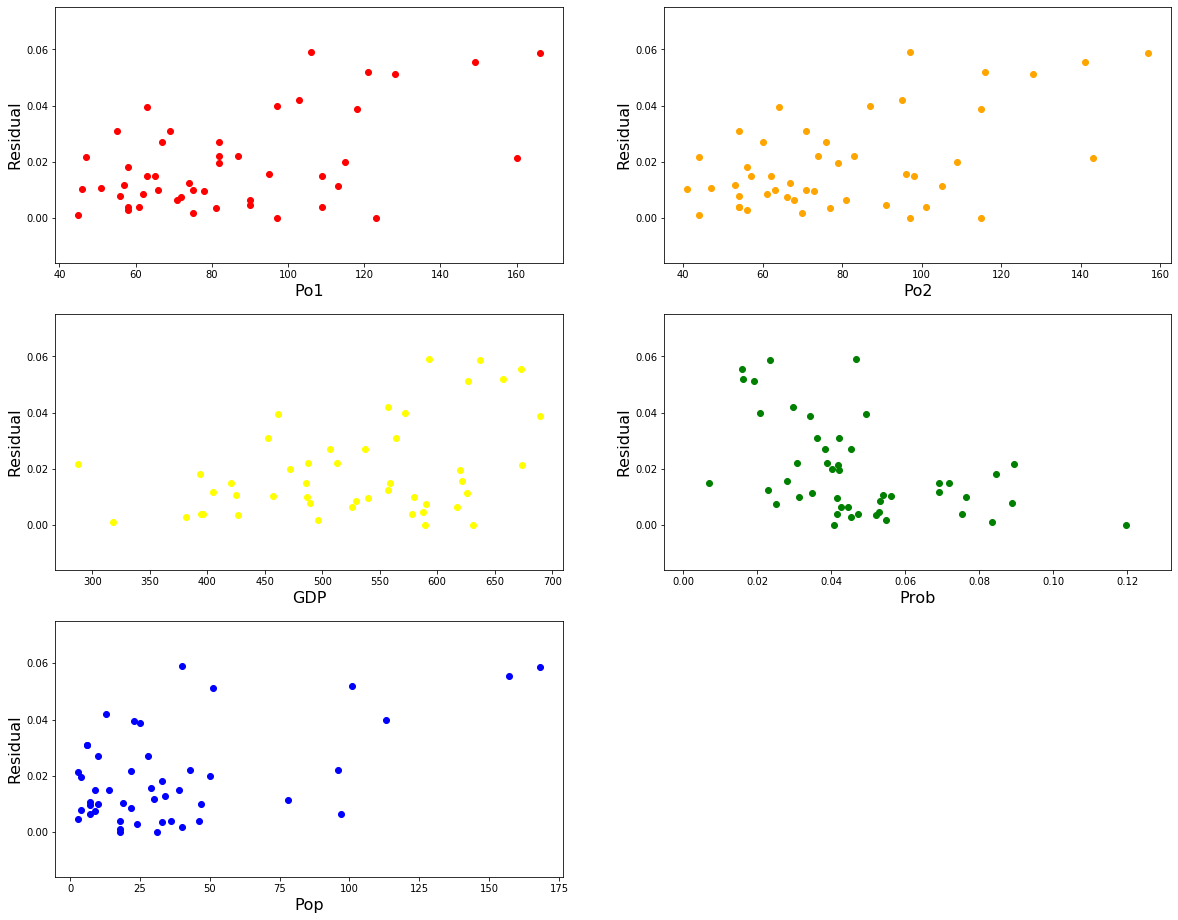
42 0.001969

43 0.015685

44 0.010301

45 0.059183

46 0.004633



(10) Do residuals of M1 and M2 follow the normal distribution based on the Jarque–Bera test? (significance level is 0.05). (10pts)

JB2 test - JB value of M2: 0.690640411356581

JB2 test - chi value of M2: 0.7079936135927738

JB2 test - JB value of M2: 1.722630913201058

JB2 test - chi value of M2: 0.42260579692833145

Yes,

(11) Do residuals of M1 and M2 satisfy homoskedasticty based on the Breusch–Pagan test? (significance level is 0.05) (10pts)

Fvalue for M1 is : 2.209001834097002e-05

Fvalue for M2 is : 8.949839425798078e-05

2. Using the MAGIC Gamma Telescope data set, build a classifier through logistic regression.

The included variabes in this dataset are as follows.

1. fLength: continuous # major axis of ellipse [mm]

2. fWidth: continuous # minor axis of ellipse [mm]

3. fSize: continuous # 10-log of sum of content of all pixels [in #phot]

4. fConc: continuous # ratio of sum of two highest pixels over fSize [ratio]

5. fConc1: continuous # ratio of highest pixel over fSize [ratio]

6. fAsym: continuous # distance from highest pixel to center, projected onto major axis [mm]

7. fM3Long: continuous # 3rd root of third moment along major axis [mm]

8. fM3Trans: continuous # 3rd root of third moment along minor axis [mm]

9. fAlpha: continuous # angle of major axis with vector to origin [deg]

10. fDist: continuous # distance from origin to center of ellipse [mm]

11. class: g,h # gamma (signal), hadron (background)

(1) Using MAGIC Gamma Telescope data set, calculate accuracy with varying cutoff for the final decision. cutoff ∈{0.1,0.15,0.2,0.25,…,0.95}. Draw a line plot (x=cutoff, y=accuracy). For this problem, the model is trained using trnX and accuracy is calculated using valX. (10pts)

accuracy = [0.71346, 0.731598, 0.746057, 0.756835, 0.767876, 0.783123, 0.788118, 0.797319, 0.800736, 0.80205, 0.804416, 0.792324 , 0.780494, 0.752629, 0.707676, 0.628023, 0.474501, 0.353049]

Graph

