**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

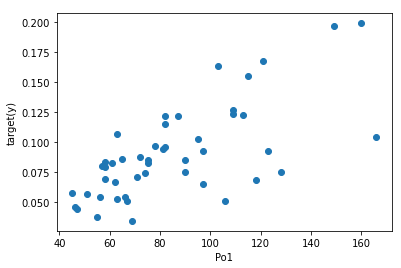
[OUTPUT]

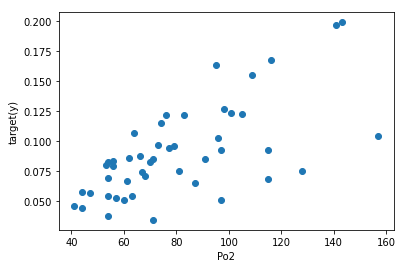
- y: crime rate in an unspecified unit region

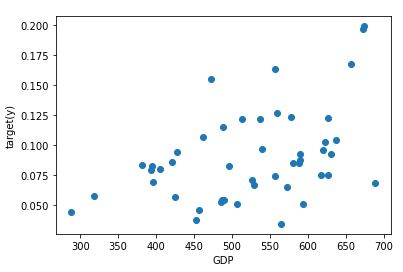
(1) Find the top 5 input variables that show the high linear correlation with the target based on the correlation coefficient. (5pts)

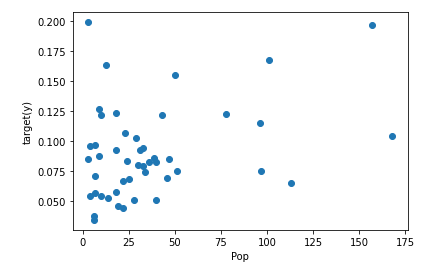
**Po1 Po2 GDP Pop Prob**

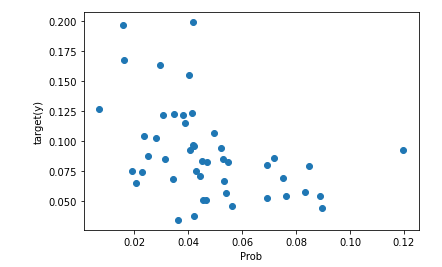
(2) Draw pairwise scatter plots – one scatter plot illustrates the relationship between the input variable selected in Question (1) and output target (Paste figures here) (5pts)











(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 | 9.22875679e-02 | 1.21783956e-03 | 2.644527949497874 | 0.01154114868544065 |
| Po1 | 2.62474439e-03 | 1.47025492e-06 | 2.1646650026677166 | 0.036287873149948524 |
| Po2 | -1.49015707e-03 | 1.67386644e-06 | -1.151785604371229 | 0.2560814022759317 |
| Pop | -1.38774123e-04 | 1.69199035e-08 | -1.0668652044513292 | 0.2922747607879401 |
| GDP | -1.53743853e-04 | 5.66706129e-09 | -2.0422965537393005 | 0.04758909240270626 |
| Prob | -4.13347529e-01 | 4.76678432e-02 | -1.893226588215596 | 0.06540028146571997 |

(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)

**Vifs = [Po1, Po2, Pop, GDP, Prob] ordered.**

**Vifs = [1.5247804645690084,**

**80.34832161088916,**

**80.97531630356094,**

**1.5174252113126923,**

**3.264732309739124]**

**-> Po2 and Pop is higher than 10.**

**For more explanatory, I drop Po2.**

* **[Po1,Pop,GDP,Prob]**

(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). (5pts)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Coefficient () |  |  | p-value |
| Intercept | 9.58045701e-02 | 1.24761871e-03 | 2.7452253646382085 | 0.008861159785890083 |
| Po1 | 1.25836611e-03 | 6.49520308e-08 | 4.9973913784108985 | 1.0736168190916828e-05 |
| Pop | -1.30331717e-04 | 1.74119047e-08 | -0.9996769797297884 | 0.3231917927962804 |
| GDP | -1.67094553e-04 | 5.71172093e-09 | -2.2377497645961713 | 0.03059852337481761 |
| Prob | -4.18254557e-01 | 4.91914623e-02 | -1.9086598210846213 | 0.06315545118196675 |

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

**M1’s Vifs = [Po1, Po2, Pop, GDP, Prob] ordered.**

**Vifs = [1.5247804645690084,**

**80.34832161088916,**

**80.97531630356094,**

**1.5174252113126923,**

**3.264732309739124]**

**M2’s vifs = [Po1, Pop , GDP, Prob] ordered,**

**Vifs= [1.5241998666647474, 3.438327737189884, 1.5126068580843406, 3.187329898579843]**

**In general, the p-value of M2 is greater than M1.**

(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 | 0.03832856278869418 | 5 | 0.007665712557738837 | 10.311248511335304 | 1.8875302661980342e-06 |
| Residual | 0.030480713807041326 | 41 | 0.0007434320440741787 |
| Total | 0.06880927659574468 | 46 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| M2 | SS | Degree of freedom | MS | F | p-value |
| Model | 0.0373423183464016 | 4 | 0.0093355795866004 | 12.460509831623249 | 9.071188887821435e-07 |
| Residual | 0.03146695824934301 | 42 | 0.000749213291651024 |
| Total | 0.06880927659574468 | 46 |  |  |  |

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

**M1**

**R^2: 0.5570260971334449**

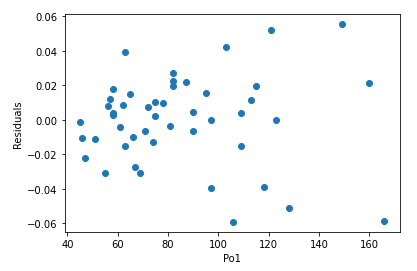
**Adjusted R^2: 0.5030048894667918**

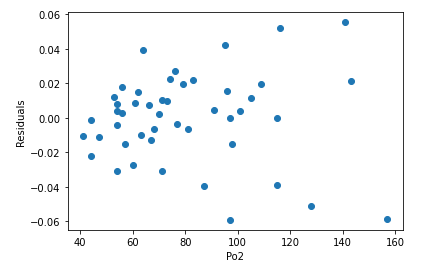
**M2**

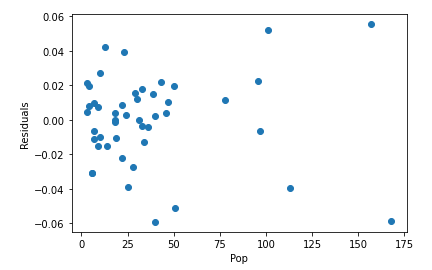
**R^2: 0.5426930814254628**

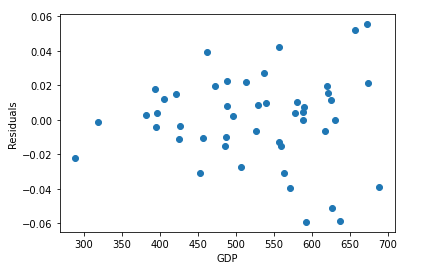
**Adjusted R^2: 0.49914004156122116**

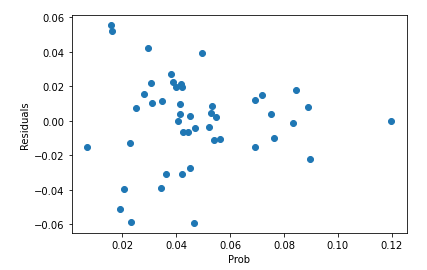
(9) Calculate residuals of M1 and draw scatter plots to show relationship between one of the input variables and residuals. (8pts)











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

**M1’s residuals p-value by Jarque–Bera test = 0.7079936135930361**

**M2’s residuals p-value by Jarque–Bera test = 0.42260579692838063**

**So, both are follow the normal distribution.**

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

**M1’s residuals p-value by Breusch–Pagan test = 8.278723675925193e-06**

**M2’s residuals p-value by Breusch–Pagan test = 2.7261145944201814e-05**

**These two values are too small, so we have to reject Ho.**

**That means, they satisfy homoskedasticty.**

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)

