Programming Assignment01

16102275 Park Hyun Woo

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.

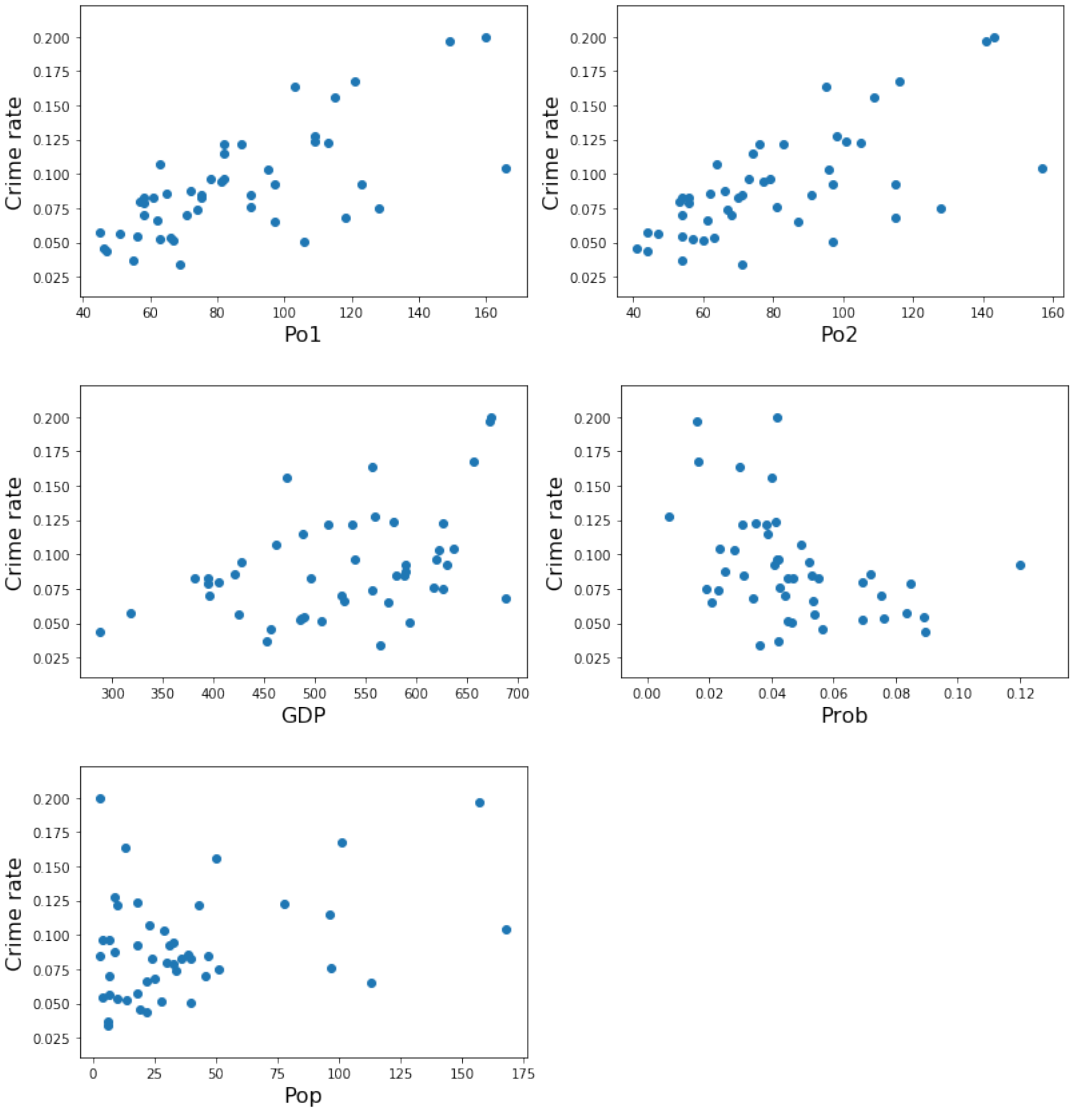
[INPUT]
- M: percentage of males aged 14-24
- So: whether it is in a Southern state. 1 for Yes, 0 for No.
- Ed: mean years of schooling
- Po1: police expenditure in 1960
- Po2: police expenditure in 1959
- LF: labour force participation rate
- M.F: number of males per 1000 females
- Pop: state population
- 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)

5 input variables that high linear correlation with target 'y' are 'Po1' , 'Po2' , 'GDP', 'Prob and 'Pop'

(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)

According to coding on Assginment, I make this table dataframe

	Coefficient	se_beta	t	p-value
intercept	0.092288	0.034898	2.644528	0.011541
Po1	0.002625	0.001213	2.164665	0.036288
Po2	-0.001490	0.001294	-1.151786	0.256081
GDP	-0.000154	0.000075	-2.042297	0.047589
Prob	-0.413348	0.218330	-1.893227	0.065400
Pop	-0.000139	0.000130	-1.066865	0.292275

(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)

Before remove anything

	VIF
Intercept	2.257469
Po1	80.348322
Po2	80.975316
GDP	3.264732
Prob	1.524780
Pop	1.517425

After remove 'Po2'

	VIF
Intercept	2.186715
Po1	3.438328
GDP	3.187330
Prob	1.524200
Pop	0.000001

VIF value of Po1 and Po2 are larger than 10
multicollinearity is severe.
On t-test table p-value of Po2 is large and VIF is also large
So remove 'Po2' variable because of multicollinearity
After remove 'Po2' VIF of all variable is lower than 10
so M2 is trained by 'Po1', 'GDP', 'Prob' and 'Pop'.

(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)

	Coefficient	se_beta	t	p-value
intercept	0.095805	0.034899	2.745225	0.008861
Po1	0.001258	0.000252	4.997391	0.000011
GDP	-0.000167	0.000075	-2.237750	0.030599
Prob	-0.418255	0.219135	-1.908660	0.063155
Pop	-0.000130	0.000130	-0.999677	0.323192

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

M1 have 5 explanatory variables 'Po1', 'Po2', 'GDP', 'Prob', 'Pop'
M2 have 4 explanatory variables 'Po1', 'GDP', 'Prob', 'Pop'

Multicollinearity of M1 is larger than M2 (Between variables, correlation is big in M1)
In M1, There are 2 variables which have significance
Because Po2 and PoP p-values are large on t-test table
In M2, only 1 variable have significance
Because Pop p-value is large on t-test table

(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	0.03048071	5	0.007665712	10.3112485	1.8875e-06
Residual	0.038328562	41	0.000743432		
Total	~~0.6880927~~ 0.06880927	46			

M2	SS	Degree of freedom	MS	F	p-value
Model	0.0314669	4	0.00933557	12.4605	9.0715e-07
Residual	0.0373423	42	0.0007492132		
Total	~~0.6880927~~ 0.06880927	46			

F test for regression model

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_a: not all β_i is not zero

p-value for F-test of M_1 & M_2 are both close to 0

So H_0 is rejected each two model ⇒ accept H_a

So, Both model is overall significant for predicting Target.

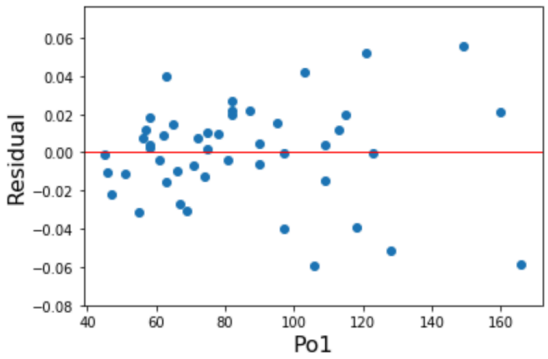
(8) Calculate R^2 and adjusted R^2 for M1 and M2. Then, compare two models. (7pts)

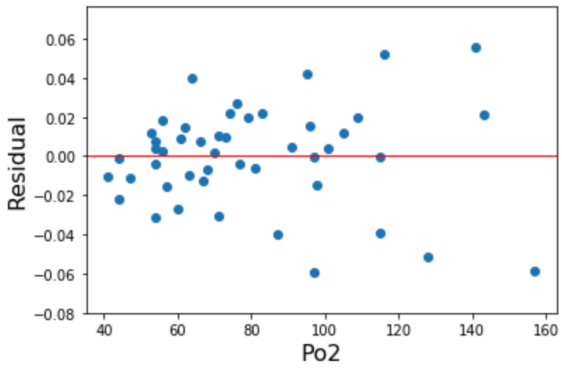
```
----------Model 1----------

R^2 : 0.5570260971334449
adjust_R^2 : 0.5030048894667918
----------Model 2----------

R^2 : 0.5426930814254629
adjust_R^2 : 0.49914004156122127
```

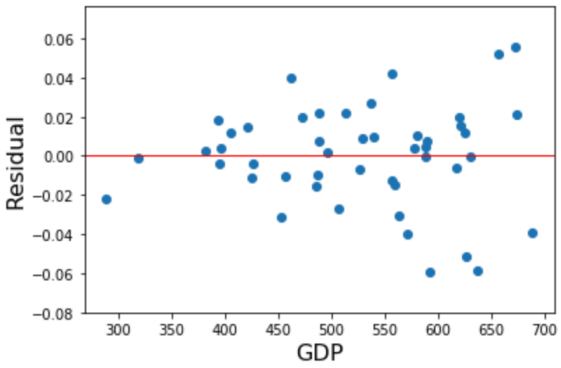
R^2 & adjust R^2 values of M1 > that of M2

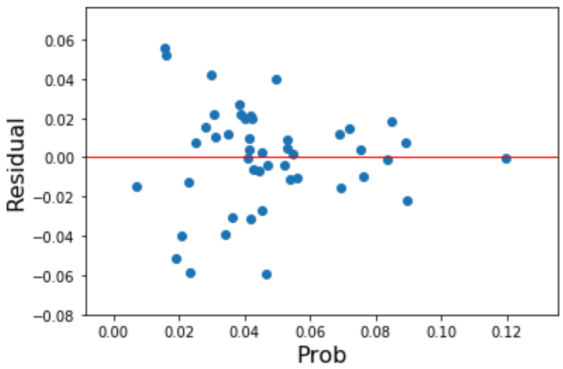
This mean Model M1 have stronger correlation with target value than Model M2

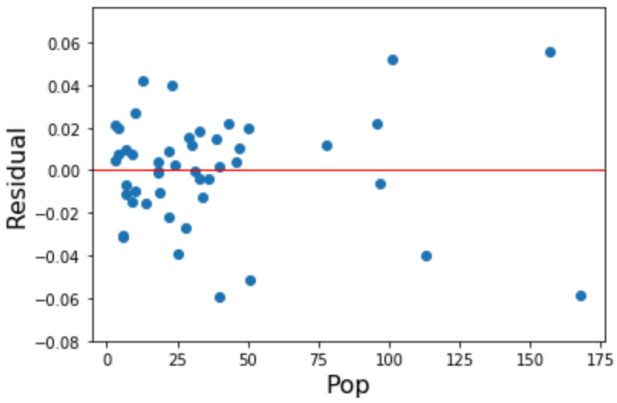
(9) Calculate residuals of M1 and draw scatter plots to show relationship between one of the input variables and residuals. (8pts)











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

p-value of M1 : 0.7079936135927738
p-value of M2 : 0.42260579692833145

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

p-value of M1 : 2.209001834097002e-05
p-value of M2 : 2.7261145944201814e-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)