1710203 Lee yoo soek

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.

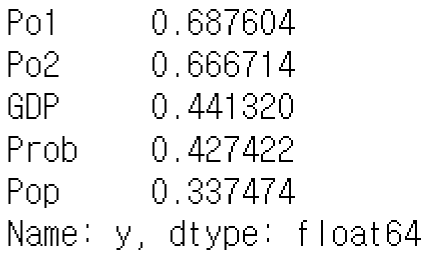
```
data=pd.read_csv('https://drive.google.com/uc?export=download&id=1ssBNxmds4zmmJbAHzJUB0_UyyfyMtoHT')
```

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

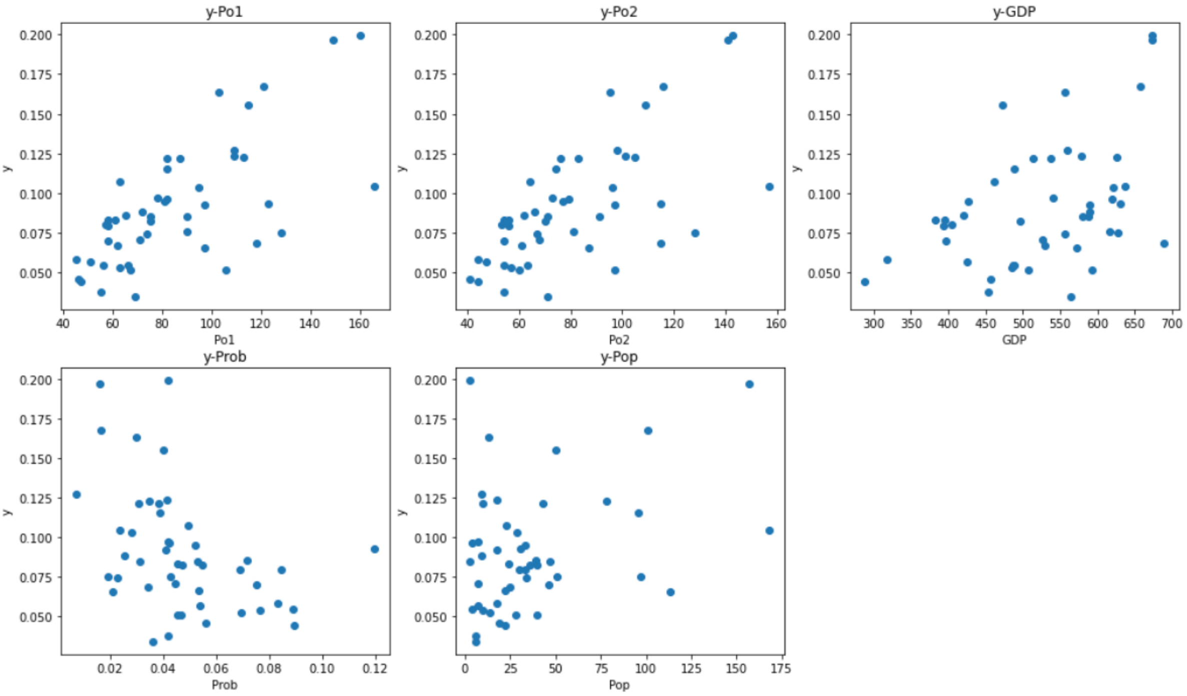
[OUTPUT]
- y: crime rate in an unspecified unit region

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



```
Po1      0.687604
Po2      0.666714
GDP      0.441320
Prob     0.427422
Pop      0.337474
Name: y, dtype: float64
```

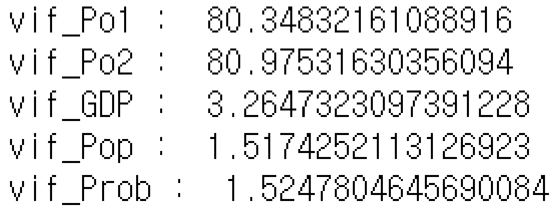
(2) Draw pairwise scatter plots – one scatter plot illustrates the relationship between the input variable selected in Question (1) and output target (Paste figures here) (5pts)



(3) Train a linear regression model (**M1**) using the selected variables in Question (1) and fill the following table. (10pts)

Variable	Coefficient (β_i)	se(β_i)	t	p-value
Intercept	0.09228756786185388	0.034898	2.644527949497416	0.011541148685453972
Po1	2.62474439e-03	0.001213	2.164665002667829	0.03628787314993942
Po2	-1.49015707e-03	0.001294	-1.1517856043711845	0.2560814022759499
GDP	-1.53743853e-04	0.000075	-2.0422965537392863	0.04758909240270759
Prob	-4.13347529e-01	0.218330	-1.8932265882155301	0.06540028146572907
Pop	-1.38774123e-04	0.000130	-1.66865204451320	0.2922747607879439

(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)



```
vif_Po1 :  80.34832161088916
vif_Po2 :  80.97531630356094
vif_GDP :  3.2647323097391228
vif_Pop :  1.5174252113126923
vif_Prob :  1.5247804645690084
```

Because VIF of Po1 and Po2 are is greater than 10, one of Po1 and Po2 is needed to be deleted.

(I decided that po2 would be removed because the VIF of Po2 is higher than VIF of Po1.)

(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 (β_i)	se(β_i)	t	p-value
Intercept	0.09580457009157052	0.034899	2.74522536463845	0.008861159785884531
Po1	1.25836611e-03	0.000252	4.997391378410947	1.0736168190916828e-05
GDP	-1.67094553e-04	0.000075	-2.2377497645964146	0.030598523374800735
Prob	-4.18254557e-01	0.219135	-1.9086598210849088	0.06315545118192856
Pop	-1.30331717e-04	0.000130	-0.9996769797298682	0.32319179279624244

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

P-value and t of Po1 in M2 is higher than in M1. So the beta_value of M2 is more meaningful than beta_value of 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	0.038328562788707624	5	0.007665712557741525	10.311248511338919	1.8875302661980342e-06
Residual	0.030480713807041326	41(=47-5-1)	0.0007434320440741787		
Total	0.06880927659574468	46			

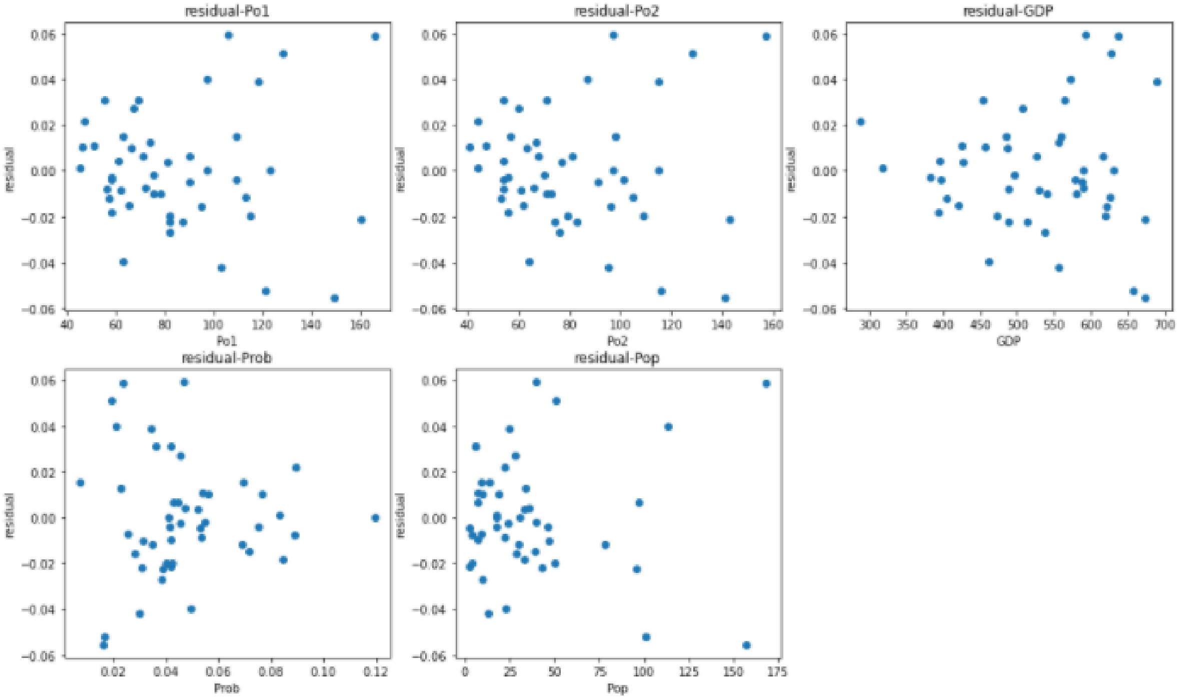
M2	SS	Degree of freedom	MS	F	p-value
Model	0.037342318346401476	4	0.009335579586600369	12.460509831623202	9.071188887821435e-07
Residual	0.03146695824934302	42(=47-4-1)	0.0007492132916510243		
Total	0.06880927659574468	46			

(8) Calculate R^2 and $\text{adjusted}\ \mathrm{R}^2$ for M1 and M2. Then, compare two models. (7pts)

	R2	Adjusted R2
M1	0.5570260971334451	0.503004889466792
M2	0.5426930814254629	0.49914004156122127

M1 is higher than M2 at R2 and Adjusted R2. M1 is a more accurate model than 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 with Jarque–Bera test of M1 :   0.7079936135930536

p_value with Jarque–Bera test of M2 :   0.4226057969283279

Both of them are bigger than 0.05 (significance level). So accept null hypothesis.

Both of them 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

- p_value of f-test statistics :   2.209001834108104e-05
- p_value of chi-test statistics :   0.00028991282719703637

M2

- p_value of f-test statistics :   2.7261145944201814e-05
- p_value of chi-test statistics :   0.0002475056888057159

Both of them are lower than 0.05 (significance level). So reject null hypothesis.

Both of them don't satisfy homoscedasticity.

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]

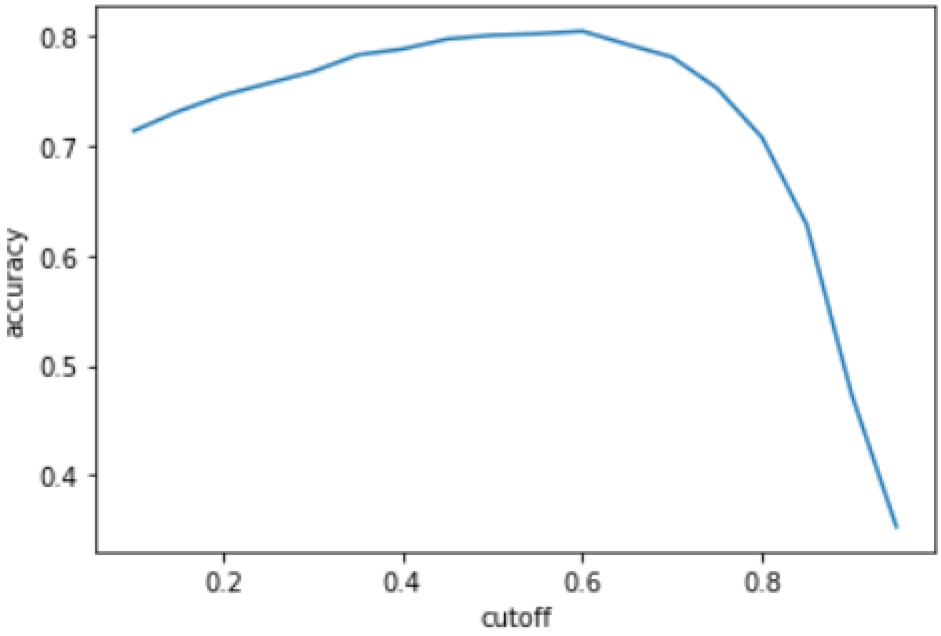
8. fM3Trans: continuous # 3rd root of third moment along minor axis [mm]

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
cutoff