**Programming Assignment01**

|  |
| --- |
| **Submission guide**  1. Write answers for following individual questions in this word 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. |

**[Linear regression]**

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=1mA0pIwPDq1jdXBUu6sf5BrixTQXeiczW') |

The given dataset aims to predict "TARGET\_deathRate" of counties using several explanatory variables related to individual counties.

[Variables]

**TARGET\_deathRate: Dependent variable. Mean per capita (100,000) cancer mortalities**

medIncome: Median income per county

binnedInc: Median income per capita binned by decile

popEst2015: Population of county

povertyPercent: Percent of populace in poverty

studyPerCap: Per capita number of cancer-related clinical trials per county

MedianAge: Median age of county residents

MedianAgeMale: Median age of male county residents

MedianAgeFemale: Median age of female county residents

Geography: County name

AvgHouseholdSize: Mean household size of county

PercentMarried: Percent of county residents who are married

PctNoHS18\_24: Percent of county residents ages 18-24 highest education attained: less than high school

PctHS18\_24: Percent of county residents ages 18-24 highest education attained: high school diploma

PctSomeCol18\_24: Percent of county residents ages 18-24 highest education attained: some college

PctBachDeg18\_24: Percent of county residents ages 18-24 highest education attained: bachelor's degree

PctHS25\_Over: Percent of county residents ages 25 and over highest education attained: high school diploma

PctBachDeg25\_Over: Percent of county residents ages 25 and over highest education attained: bachelor's degree

PctEmployed16\_Over: Percent of county residents ages 16 and over employed

PctUnemployed16\_Over: Percent of county residents ages 16 and over unemployed

PctPrivateCoverage: Percent of county residents with private health coverage

PctPrivateCoverageAlone: Percent of county residents with private health coverage alone (no public assistance)

PctEmpPrivCoverage: Percent of county residents with employee-provided private health coverage

PctPublicCoverage: Percent of county residents with government-provided health coverage

PctPubliceCoverageAlone: Percent of county residents with government-provided health coverage alone

PctWhite: Percent of county residents who identify as White

PctBlack: Percent of county residents who identify as Black

PctAsian: Percent of county residents who identify as Asian

PctOtherRace: Percent of county residents who identify in a category which is not White, Black, or Asian

PctMarriedHouseholds: Percent of married households

BirthRate: Number of live births relative to number of women in county

In this assignment, the variable, “Geography” is omitted in the modeling and the variable “binnedInc” is converted to the mid values of ranges.

**Part 1: Preprocessing**

1-(1) Some input variables contain missing values. If the proportion of missing values in a variable exceeds 70%, the variable should be excluded from the dataset. Is there any variable to be excluded? (3pts)

* Yes. [‘PctSomeCol18\_24’] should be excluded.

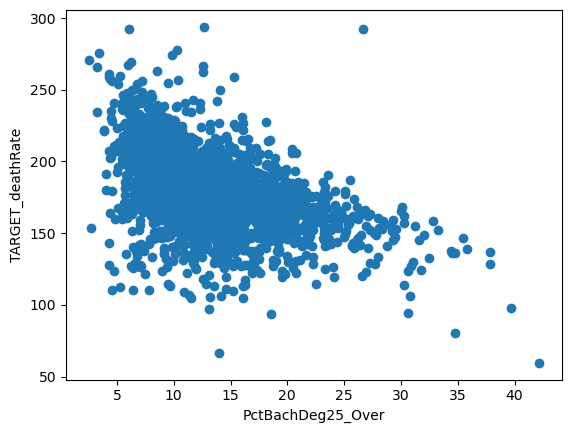
1-(2) After excluding some variables, if necessary, some rows containing missing values should be deleted from the dataset. After removing any sample with missing values, how many samples remain in the dataset? (3pts)

* 2,332 samples left.

1-(3) Using the preprocessed data, find the top 10 input variables that show the high linear correlation with the target based on the correlation coefficients. (5pts)

* PctBachDeg25\_Over 0.476887
* PctPublicCoverageAlone 0.431260
* medIncome 0.408499
* povertyPercent 0.406637
* PctEmployed16\_Over 0.401630
* PctHS25\_Over 0.398854
* PctPublicCoverage 0.390347
* binnedInc 0.372913
* PctPrivateCoverage 0.372887
* PctUnemployed16\_Over 0.367271

1-(4) Draw pairwise scatter plots – one scatter plot illustrates the pairwise relationship between one of input variables selected in Question 1-(3) and output target. Paste figures here. (5pts)



1-(5) Calculate variance inflation factor (VIF) for each explanatory variable and summary the values using a table. (5pts)

Variable VIF

0 TARGET\_deathRate 1.65

1 medIncome 13.97

2 popEst2015 1.42

3 povertyPercent 8.86

4 studyPerCap 1.04

5 binnedInc 7.41

6 MedianAge 1.03

7 MedianAgeMale 9.18

8 MedianAgeFemale 10.32

9 AvgHouseholdSize 1.34

10 PercentMarried 11.47

11 PctNoHS18\_24 1.67

12 PctHS18\_24 1.63

13 PctBachDeg18\_24 1.96

14 PctHS25\_Over 3.76

15 PctBachDeg25\_Over 5.64

16 PctEmployed16\_Over 6.58

17 PctUnemployed16\_Over 2.70

18 PctPrivateCoverage 61.91

19 PctPrivateCoverageAlone 80.94

20 PctEmpPrivCoverage 11.17

21 PctPublicCoverage 48.74

22 PctPublicCoverageAlone 39.14

23 PctWhite 7.51

24 PctBlack 5.53

25 PctAsian 2.01

26 PctOtherRace 1.60

27 PctMarriedHouseholds 9.65

1-(6) According to the results of Question 1-(5), describe your opinion on how to select which variables should be excluded from training a linear regression model. In this assignment, if the VIF is 10 or greater, the multicollinearity is considered to be severe. (5pts)

Variable VIF

1 medIncome 13.97

8 MedianAgeFemale 10.32

10 PercentMarried 11.47

18 PctPrivateCoverage 61.91

19 PctPrivateCoverageAlone 80.94

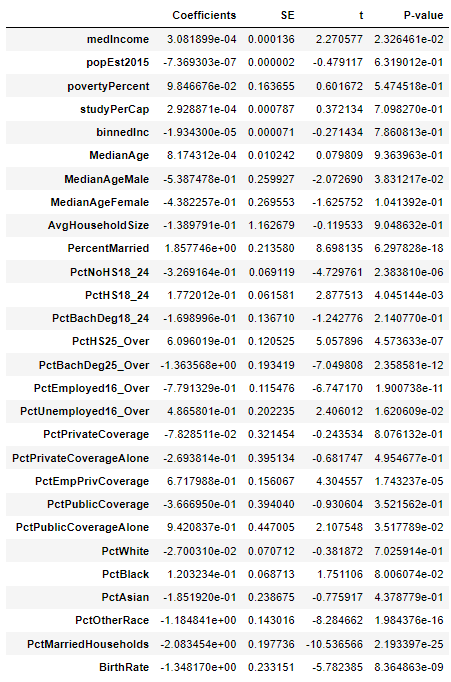
20 PctEmpPrivCoverage 11.17

21 PctPublicCoverage 48.74

22 PctPublicCoverageAlone 39.14

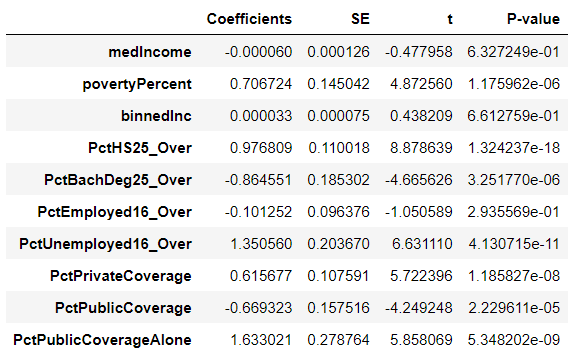
**Part 2: Modeling**

2-(1) Train a linear regression model (**M1**) using all variables and fill the following table (You should add more rows to include all variables in the following table). (5pts)



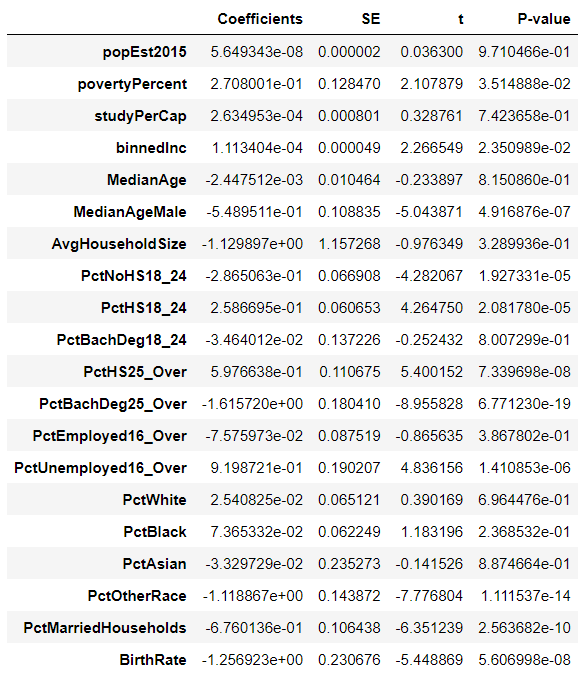
intercept: 247.0187677579828

2-(2) Train a linear regression model (**M2**) using the selected variables in Question 1-(1) and fill the following table (You should add more rows to include the selected variables in the following table). (5pts)



intercept: 93.41651878779794

2-(3) Train a linear regression model (**M3**) using the selected variables in Question 1-(4) and fill the following table (You can add more rows to include the selected variables in the following table). (5pts)



intercept: 228.27322357641185

2-(4) Describe difference between M1, M2, and M3 based on the tables of Question 2-(1), (2), and (3). (5pts)

M1 is modeled by all the variables in ‘data’.

M2 is modeled by the extracted top ten variables that are ‘**medIncome’, ‘povertyPercent’, ‘’binnedInc’, ‘PctHS25\_Over’, ‘PctBachDeg25\_Over’, ‘PctEmployed16\_Over’, ‘PctUnemployed16\_Over’, ‘PctPrivateCoverage’ ,’ PctPublicCoverage’ ,’ PctPublicCoverageAlone’.**

M3 is modeled by the extracted less than 10 VIF variables. The Variables whose VIF is over 10 are 'TARGET\_deathRate', 'medIncome', 'MedianAgeFemale', 'PercentMarried', 'PctPrivateCoverage', 'PctPrivateCoverageAlone', 'PctEmpPrivCoverage', 'PctPublicCoverage', 'PctPublicCoverageAlone'.

**Part 3: Assessment**

3-(1) Apply the F-test on M1, M2, and M3 and explain the results. In addition, fill the following tables. (6pts)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| M1 | SS | Degree of freedom | MS | F | p-value |
| Model | 711047.0467409284 | 28 | 25394.537383604587 | 55.722597050254166 | 1.1102230246251565e-16 |
| Residual | 1049549.423220478 | 2303 | 455.731403916838 |
| Total | 1760596.4699614064 | 2331 |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| M2 | SS | Degree of freedom | MS | F | p-value |
| Model | 560295.1992128154 | 10 | 56029.519921281535 | 108.34322924292972 | 1.1102230246251565e-16 |
| Residual | 1200301.270748591 | 2321 | 517.1483286292938 |
| Total | 1760596.4699614064 | 2331 |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| M3 | SS | Degree of freedom | MS | F | p-value |
| Model | 657319.2417587207 | 20 | 32865.96208793604 | 68.8432938192273 | 1.1102230246251565e-16 |
| Residual | 1103277.2282026857 | 2311 | 477.4025219397169 |
| Total | 1760596.4699614064 | 2331 |  |

3-(2) Calculate and for M1, M2 and M3. Then, compare these models. (4pts)

r2\_m1: 0.40386713189111134

adj\_r2\_m1: 0.3966193158654714

r2\_m2: 0.3182416918199873

adj\_r2\_m2: 0.3153043445206336

r2\_m3: 0.37335031222295345

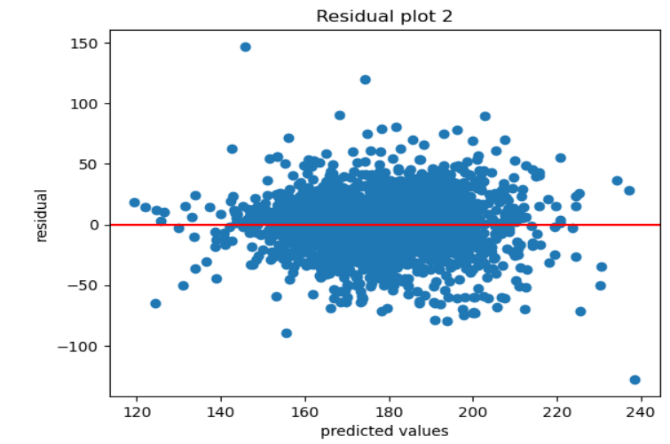
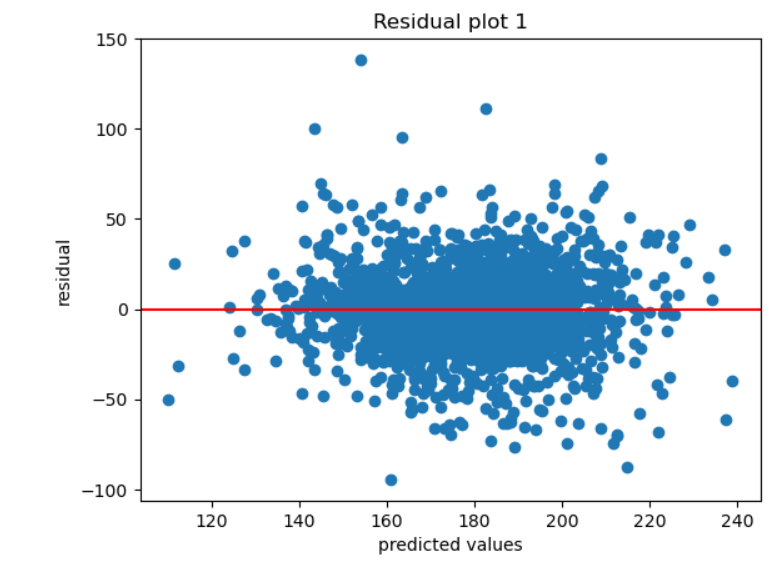
adj\_r2\_m3: 0.36792712150225204

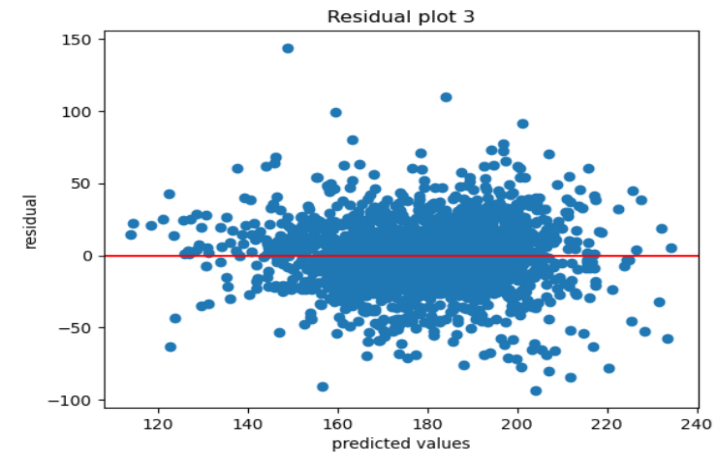
* As you can see the results, M1 has biggest r2 value among these 3 models. M3 and M2 are following sequentially.

3-(3) Considering the results of previous questions, suggest a better approach to select explanatory variables for the modeling. (5pts)

* Based on the provided information, it appears that the models are not performing well, with R-squared values ranging from 0.318 to 0.404. Moreover, the adjusted R-squared values are not much better, suggesting that the models are likely overfitting the data.
* To select better explanatory variables for the modeling, one approach could be to use a feature selection method such as backward elimination, forward selection, or stepwise regression. These methods can help identify the most relevant variables for the model and eliminate those that do not contribute significantly.
* Another approach could be to use a regularization method such as Lasso or Ridge regression, which can help avoid overfitting and improve the model's performance. These methods penalize the coefficients of the variables that are not significant and can help select the most important variables for the model.

3-(4) Obtain the residual plots of M1, M2, and M3 (x-axis=predicted target values, y-axis=residuals) and assess the assumptions related to errors based on the plots. (6pts)





* M1: The residual plot of M1 appears to have a random pattern without any discernible trend, indicating that the assumption of linearity may hold. However, the spread of residuals seems to increase as the predicted values increase, which suggests heteroscedasticity. Also, there are a few outliers with large residuals, which may have a significant impact on the model's performance.
* M2: The residual plot of M2 shows a random pattern without any obvious trend, indicating that the assumption of linearity may hold. The spread of residuals is also relatively constant across the predicted values, suggesting homoscedasticity. However, there are a few outliers with large residuals, which may have a significant impact on the model's performance.
* M3: The residual plot of M3 appears to have a random pattern without any discernible trend, indicating that the assumption of linearity may hold. The spread of residuals is also relatively constant across the predicted values, suggesting homoscedasticity. However, there are a few outliers with large residuals, which may have a significant impact on the model's performance.

3-(5) Do residuals of M1, M2, and M3 follow the normal distribution based on the Jarque–Bera test? (significance level is 0.05). If residuals do not follow the normal distribution, what might be the cause? (6pts)

<M1>

Skewness: 0.09019713089705159

Kurtosis: 5.5163861365487685

JB: 611.0151626786953

cv : 5.991464547107979

p-value : 0.0

<M2>

Skewness: -0.055006134279318074

Kurtosis: 5.3537020186013935

JB: 537.1575375124637

cv : 5.991464547107979

p-value : 0.0

<M3>

Skewness: 0.05151474206134811

Kurtosis: 5.439673569333103

JB: 574.399271885274

cv : 5.991464547107979

p-value : 0.0

: the result of the Jarque-Bera test suggests that the data does not follow a normal distribution.

In this case, the JB statistic is 611.015, 537.157, 574.399 respectively, which is much larger than the critical value for a significance level of 0.05. This means that the p-value is very small, and we can reject the null hypothesis. Therefore, we can conclude that the data does not follow a normal distribution.

3-(6) Do residuals of M1, M2, and M3 satisfy homoskedasticty based on the Breusch–Pagan test? (significance level is 0.05) (5pts)

* The p-values are all below 0.05. Therefore all the models satisfy heteroskedasticity

**[Logistic regression]**

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

The included variables in this dataset are as follows. “class” is the target (dependent) variable in this data set.

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]

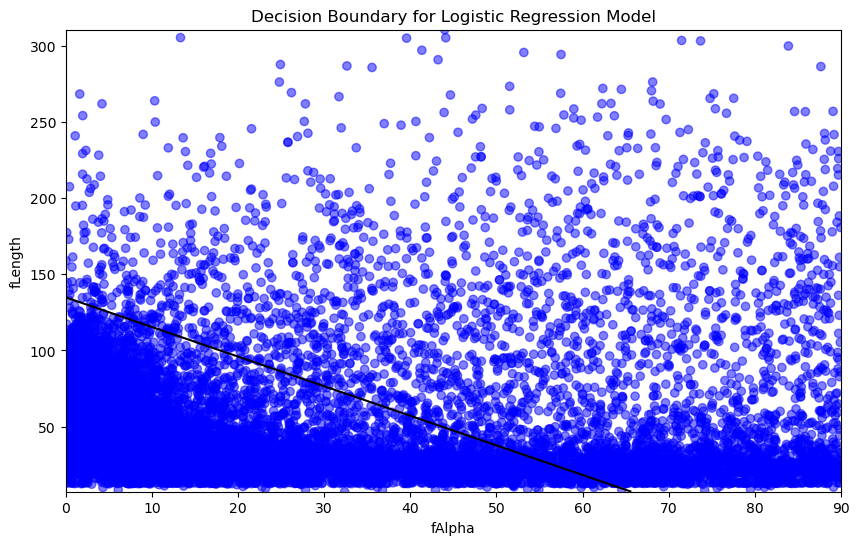
**11. class: g,h # gamma (signal), hadron (background)**

**- g is treated as class 1 while h is treated as class 0**

(1) Train logistic regression models using one of the explanatory variables using trnX (train total 10 logistic regression models) and then calculate accuracy of the models using valX. Which variable is the most important according to the accuracy? (6pts)

* The most important variable is fAlpha with an accuracy of 0.736

(2) Using the top two variables with the highest accuracy obtained for Question (1), train a logistic regression model on trnX. Draw the decision boundary showing equal probability values for classes 0 and 1 (x-axis=the variable with the highest accuracy, y-axis = the variable with the second highest accuracy) with scatter plots of the samples used for the training (assign different colors depending on the predicted class for the scatter plot). (10pts)



(3) Using MAGIC Gamma Telescope data set, calculate accuracy with varying cutoff for the final decision (if the probability of 1 >= cutoff, the predicted target is 1). 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 including all explanatory variables and accuracy is calculated using valX. (6pts)

