**Programming Assignment01**

**22101993 Eungyeol Kim**

|  |
| --- |
| **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. |

**[1. Linear regression (78pt)]**

Apply a multiple linear regression on the given dataset.

Target analysis: The given dataset aims to predict "G3" of students using several explanatory variables related to individual students.

[Variables]

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| **G3** | **Dependent variable, which means final grade (numeric from 0 to 20)** |
| School | Student’s school (binary: “GP”-Gabriel Pereira or “MS”-Mousinho da Silveira) |
| Sex | Student’s sex (binary: “F”-female or “M”-male) |
| Age | Student’s age (numeric from 15 to 22) |
| Address | Student’s home address type (binary: “U”-urban, “R”-rural) |
| Famsize | Family size (binary: “LE3”-less or equal to 3, “GT3”-greater than 3) |
| Pstatus | Parent’s cohabitation status (binary: “T”-living together, “A”-apart) |
| Medu | Mother’s education (numeric: 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education, or 4 – higher education) |
| Fedu | Father’s education (numeric: 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education, or 4 – higher education) |
| Mjob | Mother’s job (nominal: “teacher”, “health” care related, civil “services” like administrative or plice, “at\_home”, or “others”) |
| Fjob | Father’s job (nominal: “teacher”, “health” care related, civil “services” like administrative or plice, “at\_home”, or “others”) |
| Reason | Reason to choose this school (nominal: “mother”, “father”, or “other”) |
| Guardian | Student’s guardian (nominal: “mother”, “father” or “others”) |
| Traveltime | Home to school travel time (ordinal: 1 – less than 15 minutes, 2 – 15 to 30 minutes, 3 – 30 minutes to 1 hour, or 4 – more than 1 hour) |
| Studytime | Weekly study time (ordinal: 1 – less than 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours, or 4 – more than 10 hours) |
| Failures | Number of pass class failures (ordinal: as it is for 1 to 3 failures, or 4 – 4 or more failures) |
| Schoolsup | Extra educational support (binary: yes or no) |
| Famsup | Family educational support (binary: yes or no) |
| Paid | Extra paid classes within the course subject (binary: yes or no) |
| Activities | Extra-curricular activities (binary: yes or no) |
| Nursery | Attended nursery school (binary: yes or no) |
| Higher | Wants to take higher education (binary: yes or no) |
| Internet | Internet access at home (binary: yes or no) |
| Romantic | With a romantic relationship (binary: yes or no) |
| Famrel | Quality of family relationship (ordinal: from 1 to 5) |
| Freetime | Free time after school (ordinal: from 1 to 5) |
| Goout | Going out with friends (ordinal: from 1 to 5) |
| Dalc | Workday alcohol consumption (ordinal: from 1 to 5) |
| Walc | Weekend alcohol consumption (ordinal: from 1 to 5) |
| Health | Current health status (ordinal: from 1 to 5) |
| Absences | Number of school absences (numeric: from 0 to 93) |
| Course | Course of the grade (binary: “math” or “Portuguese”) |
| G1 | First period grade (numeric: from 0 to 20) |
| G2 | Second period grade (numeric: from 0 to 20) |

This dataset is a slightly modified version of a publicly disclosed dataset only for the purpose of this assignment.

**Part 1: Preprocessing**

1-(1) This dataset contains many missing values. If the proportion of missing values in a variable exceeds 80%, the variable should be excluded from the dataset. Is there any variable to be excluded? (3pts)

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

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

- Number of samples removed: 283

- Number of samples remaining: 761

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. (6pts)

failures 0.355073

higher\_yes 0.260030

Medu 0.210133

Fedu 0.186575

course\_portuguese 0.182414

studytime 0.142846

school\_MS 0.136915

romantic\_yes 0.128696

age 0.124083

guardian\_other 0.122843

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)

텍스트, 스크린샷, 디스플레이, 번호이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 번호, 라인이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 번호, 라인이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 번호, 라인이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 번호, 디스플레이이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 라인, 번호이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 번호, 라인이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 번호, 디스플레이이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 라인, 번호이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 디스플레이, 번호이(가) 표시된 사진

자동 생성된 설명

1-(5) Calculate the variance inflation factor (VIF) for each explanatory variable and summarize the values in a table. (6pts)

|  |  |
| --- | --- |
| **Variable names** | **VIF** |
| age | 1.388230 |

텍스트, 메뉴, 스크린샷, 흑백이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 메뉴, 폰트이(가) 표시된 사진

자동 생성된 설명

1-(6) According to the results of Question 1-(5), describe your opinion on which variables should be excluded from training a linear regression model. Only in this assignment, if the VIF is 4 or greater, the multicollinearity is considered to be severe. (6pts)

* All variables have VIF values below 5 and are in the range of 1-2, indicating low correlation among independent variables and a low likelihood of multicollinearity issues. Therefore, the provided variables are deemed suitable for regression coefficient estimation.
* Variables with maximum VIF: ['Fjob\_other']
* Selected variables for training the linear regression model:
* ['age', 'Medu', 'Fedu', 'traveltime', 'studytime', 'failures', 'famrel', 'freetime', 'goout', 'Dalc', 'Walc', 'health', 'absences', 'school\_MS', 'sex\_M', 'address\_U', 'Pstatus\_T', 'Mjob\_health', 'Mjob\_other', 'Mjob\_services', 'Mjob\_teacher', 'Fjob\_health', 'Fjob\_teacher', 'reason\_home', 'reason\_other', 'reason\_reputation', 'guardian\_mother', 'guardian\_other', 'schoolsup\_yes', 'famsup\_yes', 'paid\_yes', 'nursery\_yes', 'higher\_yes', 'internet\_yes', 'romantic\_yes', 'course\_portuguese']
* Variables whose VIF is larger than 4:
* ['Fjob\_other', 'Fjob\_services']
* Ommited variable:
* ['Fjob\_other']

텍스트, 메뉴, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

**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)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Coefficient () |  |  | p-value |
| Intercept |  |  |  |  |
|  |  |  |  |  |

텍스트, 메뉴, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 메뉴, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

2-(2) Train a linear regression model (**M2**) using the selected variables in Question 1-(3) and fill the following table (You should add more rows to include the selected variables in the following table). (5pts)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Coefficient () |  |  | p-value |
| Intercept | 12.066150 | 1.953246 | 6.177487 | 1.067500e-09 |

텍스트, 스크린샷, 메뉴, 폰트이(가) 표시된 사진

자동 생성된 설명

2-(3) Train a linear regression model (**M3**) using the selected variables in Question 1-(6) and fill the following table (You can add more rows to include the selected variables in the following table). (5pts)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Coefficient () |  |  | p-value |
| Intercept |  |  |  |  |

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

텍스트, 메뉴, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 메뉴, 폰트이(가) 표시된 사진

자동 생성된 설명

**Part 3: Assessment**

3-(1) Apply the F-test on M1, M2, and M3 and explain the results. In addition, fill the following tables. (7pts)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| M1 | SS | Degree of freedom | MS | F | p-value |
| Model |  |  |  |  |  |
| Residual |  |  |  |  |  |
| Total |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| M2 | SS | Degree of freedom | MS | F | p-value |
| Model |  |  |  |  |  |
| Residual |  |  |  |  |  |
| Total |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| M3 | SS | Degree of freedom | MS | F | p-value |
| Model |  |  |  |  |  |
| Residual |  |  |  |  |  |
| Total |  |  |  |  |  |

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

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

Adjusted R-Squared Values

M1: 0.252

M2: 0.227

M3: 0.253

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

To choose the better explanatory variables for modeling, consider Combining the following methods which can help in selecting the most suitable variables

1. Compare model evaluation metrics: Compare the Adjusted R-Squared values of each model to assess their explanatory power.
2. Statistical significance of coefficients: Verify the statistical significance of coefficients for each variable to identify important predictors.
3. Account for multicollinearity: Check for correlations among variables to address multicollinearity issues.
4. Utilize domain knowledge: Use expertise in the field to guide variable selection.

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)

텍스트, 스크린샷, 라인, 도표이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

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? (7pts)

Jarque-Bera test for residuals of M1:

JB statistic: 325.9060118162957

p-value: 1.699842661005773e-71

Residuals follow normal distribution: False

Jarque-Bera test for residuals of M2:

JB statistic: 237.93177897283337

p-value: 2.1566052660092452e-52

Residuals follow normal distribution: False

Jarque-Bera test for residuals of M3:

JB statistic: 326.40582936712224

p-value: 1.3239595676451135e-71

Residuals follow normal distribution: False

If the residuals of models are not following a normal distribution according to the Jarque–Bera test, it may due to missing important variables or couldn’t captured enough nonlinear relationships between model variables and response.

Sample size might be also the factor as small sample sizes makes the Jarque–Bera test to be less reliable in detecting departures from normality.

**[2. Logistic regression (22pt)]**

Using the cardiovascular disease dataset, build a classifier using logistic regression.

The included variables in this dataset are as follows.

[Variables]

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| **Cardio** | **Dependent variable, indicating the presence or absence of cardiovascular disease (binary: yes – presence, or no – absence)** |
| Age | Age of the individuals in years |
| Gender | Gender of the individuals (binary: “F”-female or “M”-male) |
| Height | Height of the individuals in centimeters (cm) |
| Weight | Weight of the individuals in kilograms (kg) |
| Ap\_hi | Systolic blood pressure of the individuals |
| Ap\_lo | Diastolic blood pressure of the individuals |
| Cholesterol | Cholesterol level of the individuals (ordinal: 1 – normal, 2 – above normal, or 3 – well above normal) |
| Gluc | Glucose level of the individuals (ordinal: 1 – normal, 2 – above normal, or 3 – well above normal) |
| Smoke | Indicates whether a person smokes (binary: yes or no) |
| Alco | Indicates whether a person consumes alcohol (binary: yes or no) |
| Active | Indicates whether a person engages in regular physical activities (binary: yes or no) |

(1) pre-process data. Some data samples have outliers for some variables, and they need to be properly addressed for accurate analysis. Implement box-plot-based outlier removal.

After removing outliers, split the whole dataset into training/test sets (8:2). Please maintain the same ratio of classes in both the training and test set. (3pts)

텍스트, 그래프, 번호, 라인이(가) 표시된 사진

자동 생성된 설명

텍스트, 그래프, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

(2) Train logistic regression models using the training samples (in x\_train) of each the explanatory variable (this will result in a total of 11 logistic regression models). After training, calculate the accuracy of each model using the validation set (in x\_valid). Which variable is the most important according to the accuracy? (5pts)

Most important variable according to highest accuracy: ap\_hi

Two most important variables according to accuracy: ['ap\_hi', 'ap\_lo']

(3) Using the top two variables with the highest accuracy obtained for Question (1), train a logistic regression model on the training samples (in x\_train). 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). (9pts)

스크린샷, 텍스트, 다채로움, 마조렐 블루이(가) 표시된 사진

자동 생성된 설명

(4) Using the cardiovascular disease dataset, calculate accuracy with varying cutoff for the final decision (if the probability of 1 >= cutoff, the predicted target is 1). cutoff ∈{0.05, 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 x\_train including all explanatory variables and accuracy is calculated using x\_valid. (5pts)

텍스트, 라인, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명