LETTERKENNY INSTITUTE OF TECHNOLOGY

ASSIGNMENT COVER SHEET

Lecturer’s Name: Dr James Connolly

Assessment Title: CA2 – predictive modelling of stroke dataset

Work to be submitted to: Blackboard

Date for submission of work:

Place and time for submitting work:

To be completed by the Student

Student’s Name: Danfeng Wang

Class:

Subject/Module: Data Science

Word Count (where applicable): 2838

I confirm that the work submitted has been produced solely through my own efforts.

Student’s signature: Date: 30th May 2021

|  |
| --- |
| **Notes**  **Penalties:** The total marks available for an assessment is reduced by 15% for work submitted up to one week late. The total marks available are reduced by 30% for work up to two weeks late. Assessment work received more than two weeks late will receive a mark of zero. [Incidents of alleged plagiarism and cheating are dealt with in accordance with the Institute’s Assessment Regulations.]  **Plagiarism:** Presenting the ideas etc. of someone else without proper acknowledgement (see section L1 paragraph 8).  **Cheating:** The use of unauthorised material in a test, exam etc., unauthorised access to test matter, unauthorised collusion, dishonest behaviour in respect of assessments, and deliberate plagiarism (see section L1 paragraph 8).  **Continuous Assessment:** For students repeating an examination, marks awarded for continuous assessment, shall normally be carried forward from the original examination to the repeat examination. |

### Abstract

Age can be used to evaluate the health status of individuals and predict age-related diseases. In this report, the dataset, consists a total of 4906 patients aged 0.08 to 82 years old over 5 years, is used to test whether underlying disease, health status, and life styles can predict the age of patient. MLR (Multi Linear Regression) algorithm is used to do age prediction. The assumptions of MLR model have been checked, and a best model with suitable independent variables is selected by backward stepwise regression approach.

At the initial stage, the model shows 0.6682 R-squared value based on the 70% randomly selected testing observations. It represents the ability of predicting patient age by underlying diseases and health related indicators. However, at the MLR assumptions check stage, all of the five MLR assumptions except homoscedasticity was accepted. The detailed process will be described in the next sections.

The related work has been committed to the below repository, <https://github.com/rachel0614/stroke.git> (see *ml.R*)

### Research question

This dataset consists patients over 5 years, in which the lifestyle and underlying diseases information of these patients were recorded.

* Life style: marital status, smoking status, living in the suburb or city, the bmi level, and overweight or not, work type, average glucose level, gender, bmi and etc.
* Diseases: if the patient has had heart disease, hypertension, stroke

The stroke dataset will be used to build a predictive model for forecasting the patient age. Lifestyle, underlying diseases have an impact to age. However, can it be predicted by these factors? In this study, an exploration of this dataset will be done for building a predictive model for forecasting the patients age.

### Building the predictive model

#### Data preparation

From the dataset summary, it can be seen that there are some cleansing and transformation work need to be done to make the dataset prepared for the data analysis. There are many categorical variables in the dataset which will be dummy encoded into numeric values.

The related works are represented in the below table.

* Table1 shows the strategy of dealing with missing values.
* Table2 shows the dichotomous variables which will be converted from labelled factor to numeric variable with values of 0 or 1.
* Table 3 shows the nominal and ordinal categorical variables which will be converted to numeric values and dummy encoded to multiple variables. The original variables will be removed and replaced by the new dummy variables. The variables with an asterisk are new created which might be helpful for the regressive model.

The evaluation of these variables will be described in next section.

Table 1 data cleansing

| **Name** | **Values** | **Method** |
| --- | --- | --- |
| ID | unique id | Drop this meaningless variable |
| date | time string | Drop this unused variable |
| Gender | Male/Female/Other | Drop rows which gender==’Others’ |
| Bmi | Continuous variable | Drop total 201 rows which bmi == Na |

Table 2 dichotomous to binary

| **Name** | **Values** | **Method** |
| --- | --- | --- |
| Gender | Male/Female/Other | Convert Male/Female to 1/0 respectively |
| Ever\_married | Yes/No | Convert to 1/0 respectively |
| Residence\_type | Rural/ Urban | Convert to 1/0 respectively |

Table 3 nominal categorical variables transformation

| **Name** | **Values** | **Method** |
| --- | --- | --- |
| work\_type | children/Govt\_job/  Never\_worked/Private/  Self-employed | Convert to 1/2/3/4/5 respectively, and then dummy encode it |
| Smoking\_status | formerly smoked/never  smoked/smokes/Unknown | Convert to 1/2/3/4 respectively, and then dummy encode it |
| Work\_type | children/Govt\_job/Never\_worked/  Private/Self-employed | Convert to 1/2/3/4/5 respectively, and then dummy encode it |
| Smoking\_status | formerly smoked/never smoked/smokes/Unknown | Convert to 1/2/3/4 respectively, and then dummy encode it |
| \*overweight | Derived from continuous variable *bmi* | Convert to 0 / 1 respectively based on the below formula:  0 – normal <25  1 - overweight >=25 |

#### Variables of interest

This dataset described the health status of patients, the underlying diseases such as cardiac disease, hypertension, average glucose level, and so forth, are indicated by different variables. Each of these patient’s data are independent.

It is widely known that many diseases are age related, which means they will be more common as people get older. Some variables relating personal information represent the patient’s life style which might also implies the patients age, such *ever\_married, work\_type, smoking\_status*. For this reason, they will be considered as independent variables of this model as well.

With this background, *age* is selected as the dependent variable of the regression model, and the independent variables will be determined from the other variables.

Among the candidate independent variables, the *residence\_type, gender* of this dataset doesn’t seem to be correlated with age based on common sense, and also, they are not really related to health status. As to smoking status, 1544 rows of smoking\_status are *unknown,* which accounts for a great part of the population with over 30%. It will be treated as a category rather than being ignored.

#### Training and testing dataset

The dataset will be used for training and test so that it will be split into two group with 70% for training and 30% for testing. *Scorecard* library is used to provide the dataset. The same split group will be used for further testing in order to evaluate the model at the same scale.

#### Building the MLR model

Table 4 build model

|  |  |
| --- | --- |
| **Initial model** | |
|  | The predictors of the initial model consists disease related variables, and life style elated variables. |

Table 5 model summary

|  |  |
| --- | --- |
| **Model summary** | |
|  | As the image shows, most of the predictors show a significant different coefficient except bmi, work\_type\_5, smoking\_status\_2, smoking\_status\_3, smoking\_status\_4.  The R-square value is 0.6682 which means the age can be predicted, and the predictors involved in this model account for 66% of the variance in age across the dataset. |

Table 6 conflint of the model

|  |  |
| --- | --- |
| **Confint of the model** | |
|  |  |

### Model validation

#### MLR assumptions

Below sections described the checklist for assumptions of this Multi-Linear Regression model.

##### Correct model

In this regression model, the dependent variable is age, which is a continuous interval.

The independent variables consist continuous variables, such as *bmi, avg\_glucose\_level*, and binary variables such as *hypertension, heart\_disease, ever\_married, stroke, work\_type, smoking\_status.* Among these independent variables, *work\_type, smoking\_status* has been converted into binary dummy variables before creating the regression model. Therefore, it has met the assumption of correct way to perform a Multi Linear Regression.

##### Outliers check

Outliers are believed influential factors to the model.

Because there is no need to check categorical variables, the other continuous variables including *bmi, avg\_glucose\_level* will be checked by qqplot.

Although there is no unique rule that in what extent to remove the outliers, it is necessary to verify them with some statistical tests. For some outliers which are extremely distant from the observations, will be removed, while other outliers will be evaluated carefully.

Table 7 bmi qqplot

| **QQplot of the model** | |
| --- | --- |
|  | From the qqplot, it can be seen that there are some points obviously skewed top left, but most of the them still keep reasonable distance.  The qqplot, by default listed the top 4 outliers, from, here it can be seen that two data points on the top right corner are labelled but the distance is not significant. |

Table 8 distribution of errors

|  |  |
| --- | --- |
| **Distribution of errors** | |
|  | The histogram overlaid with kernel density curve shows the errors likely normally distributed. It is slightly convex at bottom right area.  The row of patient id 40704 of the training set is listed as an outlier, the p-value shows it is less than significant level 0.05, the decision is to keep the outlier exist. |

Table 9 influential observations

|  |  |
| --- | --- |
| **Influential observations** | |
|  | There are three observations seem to be the most influential observations. |

Table 10 outliers before and after

|  |  |
| --- | --- |
| **Before and after summary of the initial models** | |
| **Before**    **After** | From the before and after of the same models, it can be seen that the R-squared value and the residual standard error have been improved slightly from 0.6682 to 0.6698 after removing one outlier which indicated by *outlierTest* function. However, the p-value of the outlierTest is less than 0.05 which means it is a extreme sample rather than outlier.  Finally, the code of removing this outlier this step has been commented. |

Table 11 plot bmi

| **Plot bmi value** | |
| --- | --- |
|  | The plot shows there are two outliers which are over 90 of bmi variables.  Based on common sense, BMI greater than or equal to 40 are considered extremely obese, it makes no sense that the values are two more times higher than 40. Therefore, these values will be regarded as collection errors and removed later. |

Table 12 dealing with outliers

|  |  |
| --- | --- |
| **Before and after summary of the initial models** | |
| **Before**    **After**    **After removing wrong value** | From the before and after of the same models, it can be seen that the R-squared value and the residual standard error have been improved slightly from 0.6682 to 0.6698 after removing one outlier which indicated by *outlierTest* function. However, the p-value of the outlierTest is less than 0.05 which means it is a extreme sample rather than outlier.  Finally, only two rows with suspected wrong value removed, and the R-squared value improved from 0.6698 to 0.677. |

##### Assumption – linearity and multicollinearity check

The linearity between continuous independent variables and age and the correlation between dummy independent variables and age were check as image shown below.

According to the plots, none of these dependent variables show strong correlation with each other so that the data met the assumption that independent variables should not show multicolinearity.

Table 13 check correlation of continuous variables

| **Correlation co-efficiency for continuous variables** | |
| --- | --- |
|  | Pearson method is applied in this scatterplot to indicate the correlation between continuous variables. As the graph shows, bmi and avg\_glucose\_level has low correlation (0.18 only).  Avg\_glucose\_level and ags is 0.24, which is commonly regarded to indicate the absence of correlation  Bmi and age has a positive association but it’s very weak with only 0.33. |

Table 14 correlation of binary variables and age

|  |  |
| --- | --- |
| **Correlation matrix** | |
|  | As the correlation matrix shows, most of the variables have a correlation with *age,* but the colour indicates that the correlation co-efficient is quite weak (0.4 or lower).  *Gender* and *residence type*, which is exactly matched the assumptions based on common sense in previous section.  *Ever\_married* has the highest correlation with age but it’s under 0.7 which belongs to moderate correlation.  As described in **Table 3 nominal categorical variables transformation** *work\_type\_1* represents children, so that it shows a correlation with age as well, this is also the reason why it shows a positive correlation with *ever\_married*. |

Table 15 influential observations

| **Influential observations** | |
| --- | --- |
|  | The graph indicates the influential residuals |
|  |

Table 16 Standardized residual

| **Standardized residual** | |
| --- | --- |
|  | The last outlier indicated  Below (patient id =13861) has is extreme obese and blood sugar is too high according to the values. Though it is considered to be reasonable value, the model expects average patients rather than an extreme sample.  Remove this patient where id is 13861 later to see if the performance improved. |

##### Assumption – homoscedasticity check

Table 17 non-constant error variance test

|  |  |
| --- | --- |
| **Test for Non-Constant Error Variance** | |
|  | The left image shows the homoscedasticity assumption is violated (p < .05), and this conclusion is bolstered by what we saw in plots 1 and 3 from above. |

Table 18 residuals & fitted values

| **Plot residuals** | |
| --- | --- |
|  | From the sub graph 1 (top left), it can be seen that the redline is approximately horizontal, and the standardized residuals isn’t changing much as the fitted values.  However, the Breusch Pegan test shows the null hypothesis is rejected with a p-value less than 0.05.  The Scal-Location subplot suggests some non-linearity. |

Table 19 Suggested power transformation

|  |  |
| --- | --- |
| **Suggested power transformation** | |
|  | The spread level plot gives a power transformation of 0.53 approximately. According to the left table , a power transformation might effectively stabilize the spread of the data. |

##### Assumption - autocorrelation check of residuals

Table 20 Durbin Watson test

| Use Durbin Watson test to check if the residuals are independent | |
| --- | --- |
|  | *Durbin Watson test* assesses the autocorrelation of residuals of the regression model. The null hypothesis of this test is that there is no autocorrelation. The test result indicates a failure to reject null hypothesis, and the D-W value is 2.004 which is approximately a value of 2, so that there is no autocorrelation. |

Table 21 Variance inflation factor

|  |  |
| --- | --- |
| **Variance inflation factor** | |
|  | The VIF score is used for diagnosing multicollinearity between the predictors. As the data shows, scores are all close to 1, smoking\_status related variables hs value of close to 1.5, but they are under 5 which means these two predictors are good enough. |

##### Assumption– multivariate normality check

Table 22 distribution of errors

| **Distribution of errors** | |
| --- | --- |
|  | The qqnorm combined with qqline shows the residuals deviates from the normal line on the upper end and the lower end, which means the model does not predict very well at lower and higher age patients. However, the other part of the plot approximately followed a straight line which seems normally distributed. |

Table 23 check normality

|  |  |
| --- | --- |
| **Normality assumption check** | |
|  | As the histogram shows, the residuals looks roughly normally distributed but left skewed. |

#### Transformation variables

Table 24 transformation

| **Compare models by using AIC** | |
| --- | --- |
|  | The AIC result shows the transformed age dependent variable improved. These two model involved the same terms. |

#### Stepwise regression

In order to achieve a balance by including a proper number of independent variables in this regression model, stepwise regression methods, including backward and forward, will be used.

Table 25 stepwise backward

| **Stepwise backward** | |
| --- | --- |
|  | Backward stepwise starts with maximum predictor variables, and being evaluated at each step, the less contribution predictors will be removed until the model achieved a stable quality.  Work\_type\_3 is dropped finally with a very slight drop of AIC value. |

Table 26 leaps

| **leaps** | |
| --- | --- |
|  | As the image shows, overweight has an adjusted R-square of 0.25, compares to the contribution of bmi, the overweight is the best. This variables is derived from bmi, and the previous correlation test shows it has higher correlation with age than bmi. |

#### Global validation

The below table shows the test result of MLR assumptions. The normality assumption.

From the plots in previous section, It can be seen that the relationship is not linear, a linear regression model should be applied or the transforming the variables to check the model again following the same process. Owing to time constraints and the limitation of the dataset, the research will be stopped here and to conclude all the related works in previous stage.

Table 27 global validation of assumptions

|  |  |
| --- | --- |
| **Global validation of linear regression model** | |
|  | The assessment of the linear assumptions shows that only Heteroscedasticity assumption is satisfied. The other p-value of assumption test is significantly lower than 0.05 which means the assumptions were violated. |

### Model forecasting

Three models have been created to compare the performance.

In model3 (see below table), *smoking\_status* was picked out and the continuous variable BMI was replaced by dichotomous term overweight. It achieved a relatively good prediction result than the other models. Simpler predictors make the model easy to use for end user, so that the model supposed to be the best one in this research. On the other hand, the research question of this study is to predict the patient age according to the health status of the patients, compare to bmi value, the overweigh is exactly a simple indicator among the health status. Additionally, the smoking\_status of this dataset consists about 1/3 of unknown observations, though it is regarded as a separate category initially, it is still lack of reliability. After removing smoking\_status, the model still gives a similar performance. Similarly, the avg\_glucose\_level can also indicate whether the patients have diabetes or not. This can be tested in further investigation.

As a small dataset with only five thousand or so patient’s data over 5 years, the prediction ability is still need to be checked in practice. The dropped bmi as a continuous variable, usually believed to be able to give more implication than the dichotomous variable, the real effect of dropping this predictor require further assessment in practice.

Table 28 model1

|  |
| --- |
| **Model1 – including smoking status predictor** |
|  |

Table 29 model2

|  |
| --- |
| **Model2 – including sqr age including smoking\_status precdictor** |
|  |

Table 30 model3

|  |
| --- |
| **Model3 –sqr age and simpler precdictors** |
|  |

### Conclusion

The research is based on the *stroke* dataset which consist of patient’s information gathered from 2000 to 2021. There are 4906 observations with 13 variables after cleaning.

In this report, a model of predicting patients age by health status was constructed. Most of the independent variables of this dataset is categorical variables, including dichotomous and nominal. In order to meet the assumption of building a correct MLR model, the categorical variables have all been converted to dummy binary.

As a linear regression, the five key assumptions were checked at the first stage after building the model. The linearity, multivariate, multicollinearity, autocorrelation and homoscediscity tests are described and plotted in model validation section. The result shows the model violated all the MLR assumptions except homoscediscity.

In order to achieve a balance of selection independent variables and prediction accuracy, the backward stepwise regression was executed. After comparing the ACI values, smoking\_status variables were discarded.

For practical purpose, the overweight variable was added to replace bmi as a predictor. With simpler predictor involved, even though the smoking\_status has been removed from the same model, it gives a good performance.

Since the model violated most of the MLR assumptions, the research question cannot be answered because of lacking further attempts to meet the assumptions.

### Appendix

[Table 1 data cleansing 3](#_Toc73310879)

[Table 2 dichotomous to binary 3](#_Toc73310880)

[Table 3 nominal categorical variables transformation 3](#_Toc73310881)

[Table 4 build model 4](#_Toc73310882)

[Table 5 model summary 4](#_Toc73310883)

[Table 6 conflint of the model 5](#_Toc73310884)

[Table 7 bmi qqplot 6](#_Toc73310885)

[Table 8 distribution of errors 6](#_Toc73310886)

[Table 9 influential observations 6](#_Toc73310887)

[Table 10 outliers before and after 7](#_Toc73310888)

[Table 11 plot bmi 8](#_Toc73310889)

[Table 12 dealing with outliers 8](#_Toc73310890)

[Table 13 check correlation of continuous variables 9](#_Toc73310891)

[Table 14 correlation of binary variables and age 9](#_Toc73310892)

[Table 15 influential observations 10](#_Toc73310893)

[Table 16 Standardized residual 11](#_Toc73310894)

[Table 17 non-constant error variance test 11](#_Toc73310895)

[Table 18 residuals & fitted values 12](#_Toc73310896)

[Table 19 Suggested power transformation 12](#_Toc73310897)

[Table 20 Durbin Watson test 13](#_Toc73310898)

[Table 21 Variance inflation factor 13](#_Toc73310899)

[Table 22 distribution of errors 14](#_Toc73310900)

[Table 23 check normality 14](#_Toc73310901)

[Table 24 transformation 15](#_Toc73310902)

[Table 25 stepwise backward 16](#_Toc73310903)

[Table 26 leaps 17](#_Toc73310904)

[Table 27 global validation of assumptions 17](#_Toc73310905)

[Table 28 model1 18](#_Toc73310906)

[Table 29 model2 18](#_Toc73310907)

[Table 30 model3 19](#_Toc73310908)