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):

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, \_\_\_\_ patients aged 0.08 to 82 years old over \_\_\_ years is used to test whether underlying disease, health status, and living styles can predict the age of patient. Multi Linear Regression algorithm are used in this study to select specific variables to do age prediction.

The related work has been committed to the below repository <https://github.com/rachel0614/stroke.git>

参考

<https://training.galaxyproject.org/training-material/topics/statistics/tutorials/age-prediction-with-ml/tutorial.html>

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1599-6

### 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 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 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 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 |
| \*Diabetes | Derived from continuous variable  avg\_glucose\_level variable | Convert to 1/2/3 respectively based on the below formula:  1 - normal <= 140  2 - pre-diabetes <= 199  3 - diabetes >199  , and then dummy encode it |
| \*Bmi\_level | Derived from continuous variable bmi | Convert to 1/2/3/4 respectively based on the below formula:  1 - underweight < 18.5  2 - normal >=18.5  3 - overweight >=25  4 - obese >= 30 |

#### Initial stage

Select constant terms, no variable terms yet

#### Test linear terms

If it’s necessary, add linear terms.

#### Test cross-product terms

If necessary, add cross-product terms

#### Test univariate 2nd, 3rd 4th order terms

If necessary, higher order terms added

#### Test sufficient

#### How it fits with your dataset

##### 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.

Before constructing the model, some statistical tests and plots will be applied to identify the most important independent variables in this regression model.

例如 Fit the regression model using the standardized independent variables and compare the standardized coefficients. how much does the [R-squared](https://statisticsbyjim.com/regression/interpret-r-squared-regression/) increase for each variable when you add it to the model last。When an independent variable is the last one entered into the model, the associated change in R-squared represents the improvement in the goodness-of-fit

##### Examine the variables of interest for linearity, outliers, and normality

|  |  |
| --- | --- |
| **Normality of age** | |
|  |  |

Table 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 4 correlation of dichotomous variables

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

#### Decision

* Ignore *residence\_type*, *gender* variable because the correlation co-efficiency with age is only -0,03 and 0.
* Keep *bmi* and *avg\_glucose\_level* as predictors at the initial stage. As continuous variables, they represent more information than dichotomous, and replacing it just because of less correlation than dichotomous variable might bring some disadvantages[[1]](#footnote-1).
* Keep *avg\_glucose\_level* as predictor, but it need to be evaluated carefully because of the low correlation co-efficiency with age.
* *Diabetes will be as a substitute predictor of avg\_glucose\_level.*
* *Although overweight* or *bmi\_level* has relatively high correlation with *age,* they will be candidate variables and evaluated later.
* *Work\_type* and *smoking\_status*, as nominal variables, will be replaced by the relevant dummy variables.
* The other variables including *hypertension, heart\_disease, stroke* will be included in the model as predictors.

In order to achieve a balance by including a proper number of independent variables in this regression model, R-squared, p-value and other measures will be used to evaluate the model

|  |  |
| --- | --- |
|  | |
|  | The initial model involved all the independent variables except *gender* and *residence\_type*. |

|  |  |
| --- | --- |
| **Variance inflation factor** | |
|  | The VIF score is used for diagnosing multicollinearity between the predictors. As the data shows, scores are all close to 1 except work\_type\_1 and work\_type\_4, but they are under 5 which means these two predictors are good enough. |

|  |  |
| --- | --- |
| **Normality assumption check** | |
|  | The residuals looks roughly normally distributed. |

##### Outliers

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 bmi qqplot

|  |  |
| --- | --- |
| **QQplot of bmi variable** | |
|  | Qqplot bmi value, it can be seen that there are some points obviously skewed top left, but most of the them still keep reasonable distance.  While the 2020 and 4030 (row index) are far higher than the other data points. |

Table outliers

|  |  |
| --- | --- |
| **Standardized residual** | |
|  | The 4030 & 2020 point indicate the row number of outliers. It can be seen that the corresponding bmi values of these two rows are 92 and 97.6 respectively.  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.  The outlier of row 181 has a bmi value of 40 and avg\_glucose\_level of 210.40, though both are extremely high, it is still considered to be a reasonable value, and will be kept. |

Actually, bmi >60 kg/m2 is defined as super-super obesity, the health-risks 10-folder greater mortality than normal population, also heart disease, hypertension, and many other diseases.

Table 9 distribution of errors

|  |  |
| --- | --- |
| **Distribution of errors** | |
|  | The histogram overlaid with kernel density curve shows the errors are almost normally distributed. It is slightly convex at bottom right area.  Row 433 is listed as an outlier in the outlierTest. The p-value of this test shows it is significant so that it will be deleted later. |

Table outliers before and after

|  |  |
| --- | --- |
| **Before and after summary of the initial models** | |
|  | 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. |

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

### Model forecasting

##### Influential measures

The observations that have 4 times greater distance than mean are regarded as influential.

### Conclusion

The research is based on the *stroke* dataset which consist of patient’s information gathered from 2000 to 2021. There are 3426 observations with 11 variables after cleaning. 180 of them have ever had stroke. The variables applied in this investigation include stroke, age, smoking, average glucose value, etc.

The study mainly focussed on age, smoking and the relationship with stroke. Due to space and time limitation, we have ignored the testing of some other important factors of stroke, such as hypertension, heart disease, and etc.

In conclusion, increasing age and smoking are both associated with stroke. Among the stroke patients, smoker shows a significant difference in age than non-smokers, they are more likely to have stroke in younger age than non-smokers.

1. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006 Jan 15;25(1):127-41. doi: 10.1002/sim.2331. PMID: 16217841. [↑](#footnote-ref-1)