**D208 Performance Assessment**

**NBM2 Task 2**

**Multiple Logistic Regression**

**for Predictive Modeling**

**Western Governors University**

**By: Michael Rupert**

**May 2022**

Contents

[A1: RESEARCH QUESTION 3](#_Toc104797407)

[A2: OBJECTIVES AND GOALS 3](#_Toc104797408)

[B1: SUMMARY OF ASSUMPTIONS 3](#_Toc104797409)

[B2: TOOL BENEFITS 4](#_Toc104797410)

[B3: APPROPRIATE TECHNIQUE 4](#_Toc104797411)

[C1: DATA GOALS 4](#_Toc104797412)

[C2: SUMMARY STATISTICS 5](#_Toc104797413)

[C3: STEPS TO PREPARE THE DATA 6](#_Toc104797414)

[C4: VISUALIZATIONS 13](#_Toc104797415)

[C5: PREPARED DATA SET 20](#_Toc104797416)

[D1: INITIAL MODEL 20](#_Toc104797417)

[D2: JUSTIFICATION OF MODEL REDUCTION 22](#_Toc104797418)

[D3: REDUCED LOGISTIC REGRESSION MODEL 22](#_Toc104797419)

[E1: MODEL COMPARISON 23](#_Toc104797420)

[E2: OUTPUT AND CALCULATIONS 23](#_Toc104797421)

[E3: CODE 23](#_Toc104797422)

[F1: RESULTS 29](#_Toc104797423)

[F2: RECOMMENDATIONS 29](#_Toc104797424)

[G: PANOPTO DEMONSTRATION 30](#_Toc104797425)

[H: SOURCES OF THIRD-PARTY CODE 30](#_Toc104797426)

[I: SOURCES 30](#_Toc104797427)

# 

# A1: RESEARCH QUESTION

Can we determine using the provided data set which patients are likely to be readmitted?

# A2: OBJECTIVES AND GOALS

The focus of this is to define which independent variables affect readmission of a patient. Providing this information would be critical to a facility in lowering readmission rates and possibly reducing cost through potential fines that would be levied. This is of value to both administration, government and care givers, to help reduce cost, improve care and come up with strategies that focus on the key factors.

# B1: SUMMARY OF ASSUMPTIONS

Logistical Regression makes different assumptions than Linear and General Linea Regression models that are based on Least Squares, specially dealing with homoscedasticity, linearity, measurement level and normality.

Logistic Regression has no requirement for linear relationship between the independent and dependent variables. Residuals or error terms have no need of normal distribution. Logistical Regression has no need for homoscedasticity. The target variable is not measured on a ratio or interval.

**Requirements of Logistical Regression**

* The target or dependent variable must be binary and ordinal Logistical Regression the target variable must be ordinal.
* The observations must be independent of each other. The measurements should not be matched data or repeated measurements.
* Multicollinearity between the independent variables must be kept to a minimum or have no multicollinearity. A low correlation between independent variables.
* There are no extreme outliers
* There is a Linear relationship between explanatory variables and the Log.
* The sample size is sufficiently large. Generally calculated as (number of Records) \*(number of explanatory variables)/ (least frequent outcome) (ZACH, 2020)

# B2: TOOL BENEFITS

Python will be used due to its robust packages and ability to handle large datasets.   
Python is platform agnostic and can be used by any computers. Some of the libraries that will be used in the analysis: Numpy, Pandas, ScikitLearn, Seaborn, Matplotlib, Statsmodel as well as many others. With these packages, rich visualizations and powerful analysis can be performed.

# B3: APPROPRIATE TECHNIQUE

Since readmission is a simple binary response, yes or no, Logistical Regression is suitable for modeling this set of data.

# C1: DATA GOALS

As done previously in Data cleaning many of the missing variables were replaces with mean values. This would not be suitable for Logistical Regression, since it violates one of the assumptions:” observations must be independent of each other. The measurements should not be matched data or repeated measurements.” In the case of missing continuous values, a strategy other than mean replacement must be employed. The strategy that will be employed here will be removal of records with nulls or incorrect information i.e., Population areas report as 0. Other possible ways to handle missing or erroneous information is; Deletion (columns, rows, pairwise), Imputation (Mean, Median, sample, Linear Interpolation, seasonal adjustment, create a new level such as N/A, Multiple Imputation and Linear Regression). (BILOGUR, 2018)

The data must be examined for extreme outliers and these either need to be removed or normalized.

Check the linear relations between the data the Log.

# C2: SUMMARY STATISTICS

The first is to remove move data that is obviously unneeded, 'CaseOrder', 'Customer\_id', 'Interaction', 'UID', 'County','Timezone', 'Job', 'Education', 'Employment', Children are the first to me dropped from the data set. This leaves us with the following summary shown below:

Graphical user interface

Description automatically generated

The above data output shows there are several values missing for certain fields, specifically The fields of Age, Income, Overweight, Anxiety, Initial\_days and if we further check the field of population, we find 109 zero values which is obvious erroneous since if a person lives in that area, then the population is at least 1. Each of these fields will be handles differently for imputation or deletion of records.

In the case of population, the 109 represents only .1% of the total records and those records will simple be removed. In the case of Overweight and Anxiety a new categorical value of N/A will be added. For initial\_days and Age the missing values will be randomly filled with an integer between min and max. For income missing values will be replaced randomly with values between min and max. by randomly replacing the values.

We must check for duplicates. This was previous checked on the full dataset in D206, and no duplicates were found.

We must use various means to determine if there are extreme outliers and these must be handled. We can use boxplots and histograms to do this.

We also will take categorical information and code according as in the case of Booleans 1/0 and if ordinal we will use an appropriate integer scale. The encode dictionary will be represented in code.

We are left with the following variables: 'Lat', 'Lng', 'Population', 'Age', ‘Household\_Income', 'ReAdmis', 'VitD\_levels', 'VitD\_supp', 'Soft\_drink', 'HighBlood',

'Stroke', 'Complication\_risk', 'Overweight', 'Arthritis', 'Diabetes', 'Hyperlipidemia', 'BackPain', 'Anxiety', 'Allergic\_rhinitis', 'Reflux\_esophagitis', 'Asthma', 'Daily\_Average\_Charges', 'Additional\_charges', 'Timely\_Admission', 'Timely\_Treatment',

'Timely\_Visits', 'Reliability', 'Hours\_Treatment', 'Courteous\_Staff', 'Active\_Listening'.

# C3: STEPS TO PREPARE THE DATA

1. Import the data set to a Data frame.
2. Remove unneeded columns.
3. Check for duplicates and remove.
4. Encode Categorical data.
5. Look at the description of the data.
6. Determine what fields are missing values or have erroneous information.
7. Handle missing values through imputation or deletion of records.
8. Look for extreme outliers and normalize or remove.
9. Rename misleading or ambiguous columns

Import Data frame

Text

Description automatically generated

Remove unneeded columns

Text

Description automatically generated

Check for duplicates and remove

Text

Description automatically generated

Look at the description of the data

Graphical user interface, text

Description automatically generated

Check for Duplicates

Graphical user interface, text, website

Description automatically generated

Remove unlikely data, specifically population data of zero.

Text

Description automatically generated with medium confidence

Remove the records that contain a population of zero.

A screenshot of a computer

Description automatically generated with medium confidence

Encode the Categorical data

A picture containing text

Description automatically generated

Find the columns with NaN values.

Text

Description automatically generated

Replace the NaN values with the appropriate typed and range of values.



Text

Description automatically generated

Examine the data description once more. And we can see by the counts and the minimum values the data is now complete.

Graphical user interface, application

Description automatically generated

Lastly, Rename misleading and ambiguous columns.

Graphical user interface

Description automatically generated

If we inspect the Means, Quartiles and standards deviation all seem to be within reasonable limits. The Income at the low end seems low but could be explained as part time workers, students or other explainable reasons.

Graphical user interface

Description automatically generated with low confidence

# C4: VISUALIZATIONS

Univariate Visualizations

Chart

Description automatically generatedChart, histogram

Description automatically generated

Above are the output histograms of continuous variables selected from this dataset. Here we can see the probability distribution for the continuous values we selected. There appears to be some skew in the probabilities in some areas, but this is not an unacceptable set of data.

For the Categorical and Ordinal data variables below is a series of Bar charts that easily allow us to see the distributions. All of the Visualizatiosn below seem reasonable with no unexpected extremes.

The code to produce this is below.

Text

Description automatically generated

Bar chart

Description automatically generated with medium confidence Graphical user interface, chart, application, Excel, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Bivariate Visualizations

Continuous Variables

This are shown with scatterplots showing the frequency in points, of continuous variables, placed in prespective to whether readmitted or not.

Text

Description automatically generated

Chart, bar chart

Description automatically generated Timeline

Description automatically generated with medium confidence

Categorical Values

Stacked bar graph represent these Bivariate comparisons of categorical varaibles from out selected data.

A screenshot of a computer

Description automatically generated with medium confidence

Chart, bar chart

Description automatically generated Chart, bar chart, box and whisker chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart, box and whisker chart

Description automatically generated Chart, bar chart, box and whisker chart

Description automatically generated

Chart, bar chart, box and whisker chart

Description automatically generated Chart, bar chart

Description automatically generated

# C5: PREPARED DATA SET

*\*\*\*Prepared as a CSV file.\*\*\**

The code to generate this file is shown below.



# D1: INITIAL MODEL

Using Logit build an initial model. Below is the code and the output of the initial model.

Text

Description automatically generated

Params output:

Text

Description automatically generated with medium confidence

Logit Regression Results below:

Text

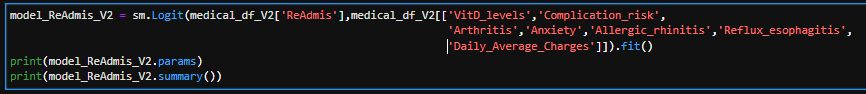
Description automatically generated with medium confidence

# D2: JUSTIFICATION OF MODEL REDUCTION

It can be seen from the above summary results that many of the p-values are above 0.05. Very few are 0.05 or below. We will keep these variables below 0.05 as being of interest in a reduced regression model. It is also noted that the pseudo-R-squared value is 0.6746, this indicates a poor fit for this model. Of the 30 exploratory variables in the data set 23 of them are above 0.05. it is justified that the reduced model contains the remaining seven.

# D3: REDUCED LOGISTIC REGRESSION MODEL

The reduced model consists of 7 variables and the target variable. Below is the code and the output of this model:



Text

Description automatically generated with medium confidence

# E1: MODEL COMPARISON

The Reduced model compared to the initial model is not much of an improvement. Both models have an accuracy of the Initial model is 0.9060762309169953 or 91% accuracy, the reduced model accuracy is 0.9057729248812051 slightly less but still 91% if we round. Similarly, sensitivity of initial and reduced are respectively, 0.9054836042987049 and 0.9038302562689446, 91% and 90%. The specificity is similarly close, initial = 0.6333407721490739, reduced= 0.6338877106819958 of both 63%. Either model is a good fit with this data. It would probably be easier to use the reduced model since the accuracy and other calculations are similar.

# E2: OUTPUT AND CALCULATIONS

Below is the output code and confusion matrices and visualizations

Initial model outputs and visualizations

Text

Description automatically generated with medium confidence

Reduced model outputs and visualizations

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

# E3: CODE

The code will be included here, but also a separate file will be included with this submission.

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

Text

Description automatically generatedText

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

# F1: RESULTS

The regression equation that is yield from this analysis is:

ReAdmis= -0.636669(VitD\_levels) + -0.259786(Complication\_risk) + -0.204376(Arthritis) + -0.110353(Anxiety) + -0.248485(Allergic\_rhinitis) + -0.154311(Reflux\_esophagitis) + 0.001852(Daily\_Average\_Charges)

The coefficients are exactly that. These respective values negative trending, more towards not being readmitted and positive trending towards readmission. Daily\_Average\_Charges seems to be the biggest impact on chances of readmission and VitD\_levels levels seem to impact it in the opposite direction.

Limitation of this analysis, as with any set of data, there is a risk that the whole population is not fully represented. There is a chance that this model simply fits just this data and not the general population. There could also be substitution made for the missing data that cause some sort of skew. The large set of data of the initial model and the reduce model being so closely alike in the scoring results seem to indicate that either model seems to be a very good fit.

# F2: RECOMMENDATIONS

As with any medical facility patients will become ill or infirm. They will seek care and if the situation persists, they will return. However, this analysis shows that there are areas that can be given attention and perhaps reduce the likelihood of readmission of patients. Why is the average amount of money spent each day, a factor for readmission? Perhaps these patients are much more ill, or perhaps they are released too soon due to insurance limitation by their medical insurer? It would seem that higher Vitamin D levels are a factor in preventing readmission in some patients. Some of the other factors can be explored, Anxiety, complication risk and some of the others are treatable while still in the 30-day period through outpatient agencies. There are areas that can help reduce readmission in patients.

# G: PANOPTO DEMONSTRATION

# H: SOURCES OF THIRD-PARTY CODE

# <https://app.datacamp.com/learn/courses/intermediate-regression-with-statsmodels-in-python>

# <https://seaborn.pydata.org/generated/seaborn.catplot.html>

# I: SOURCES

# <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/>

team, C. c. (n.d.). *Histograms and Box Plots*. Retrieved from CIToolkit: https://citoolkit.com/articles/histograms-and-boxplots/

The Trustees of Princeton University. (2007). *Interpreting Regression Output*. Retrieved from Princeton University Library: https://dss.princeton.edu/online\_help/analysis/interpreting\_regression.htm#:~:text=a%20miniscule%20effect.-,Coefficients,the%20direction%20of%20the%20effect.

Zach. (2021, November 16). *Assumptions of Multiple Linear Regression*. Retrieved from statology: ps://www.statology.org/multiple-linear-regression-assumptions/

# Bibliography

BILOGUR, A. (2018, April 28). *Simple techniques for missing data imputation*. Retrieved from Kaggle: https://www.kaggle.com/code/residentmario/simple-techniques-for-missing-data-imputation/notebook

team, C. c. (n.d.). *Histograms and Box Plots*. Retrieved from CIToolkit: https://citoolkit.com/articles/histograms-and-boxplots/

The Trustees of Princeton University. (2007). *Interpreting Regression Output*. Retrieved from Princeton University Library: https://dss.princeton.edu/online\_help/analysis/interpreting\_regression.htm#:~:text=a%20miniscule%20effect.-,Coefficients,the%20direction%20of%20the%20effect.

ZACH. (2020, October 13). Retrieved from Statology: https://www.statology.org/assumptions-of-logistic-regression/

Zach. (2021, November 16). *Assumptions of Multiple Linear Regression*. Retrieved from statology: ps://www.statology.org/multiple-linear-regression-assumptions/