**Performance Assessment: D208 – Logistic Regression**

**A. Research Question**

**1.** For this assessment, the research question is at follows: using logistic regression, can we predict the readmissions variable based on the condition variables?

**2.** The goal of this data analysis is to see if we can predict whether or not a patient was readmitted into the hospital within one month of release based on whether or not the patient has a condition (asthma, diabetes, etc.).

**B. Method Justification**

**1.** The following are 4 assumptions about logistic regression:

* The response variable is binary: the target variable must be binary, i.e., 0, 1 or No/Yes
* The observations are independent: the observations should not be related in any way to one another
* There is no multicollinearity: the predictor variables are assumed to have no multicollinearity as it can affect the fitting of the models and how it is interpreted
* There are no extreme outliers: extreme outliers can heavily influence the model so they should be removed beforehand (Zach 2020).

**2.** One benefit of using R is that R is a free and open-sourced platform that anyone can download and use. Another benefit of using R is that R has numerous free packages to install that allows the user to do more with the data than the original code allows. There are also great visualization tools, and it is easy to pick up and learn and apply numerous statistical methods to your data set.

**3.** Since the target variable (ReAdmis) is a categorical/binary variable, then logistic regression is an appropriate way to analyze this research question.

**C. Data Preparation**

**1.** All of my variables (both dependent and independent) for this analysis are categorical variables. Because of this, my data cleaning goals are to remove all the variables that will not be used for this analysis and then to check for missing data in the remaining variables. Any missing data will be replaced with a “No” since we cannot assume a patient has one or more of the condition variables. Since these are categorical variables, there is no need to check for outliers since Yes/No variables cannot contain any outliers. The code has been attached as an RScript file alongside this written assessment.

**2.** Below is a screenshot of the summary statistics for each variable used in this analysis. This screenshot shows how many patients answered “Yes” and “No” to the readmission variable and each of the condition variables.

A black text on a white background

Description automatically generated

* ReAdmis (target variable): categorical, 6331 No and 3669 Yes
* Soft\_drink: categorical, 7425 No and 2575 Yes
* HighBlood: categorical, 5910 No and 4090 Yes
* Stroke: categorical, 8007 No and 1993 Yes
* Overweight: categorical, 2906 No and 2094 Yes
* Arthritis: categorical, 6426 No and 3574 Yes
* Diabetes: categorical, 7262 No and 2738 Yes
* Hyperlipidemia: categorical, 6628 No and 3372 Yes
* BackPain: categorical, 5886 no and 4114 Yes
* Anxiety: categorical, 6785 No and 3215 Yes
* Allergic\_rhinitis: categorical, 6059 No and 3941 Yes
* Reflux\_esophagitis: categorical, 5865 No and 4135 Yes
* Asthma: categorical, 7107 No and 2893 Yes

**3.** Univariate distributions of each variable as well as bivariate distributions using the target ReAdmis variable have been generated in R. The RScript file showing these distributions has been attached alongside this written assessment.

**4.** All of the variables for this analysis are character classes. In order to use logistic regression, I will convert each variable to a factor. This is the only step for data transformation. The code has been attached as an RScript file alongside this written assessment.

**5.** A copy of the cleaned data set has been attached alongside this written assessment.

**D. Model Comparison**

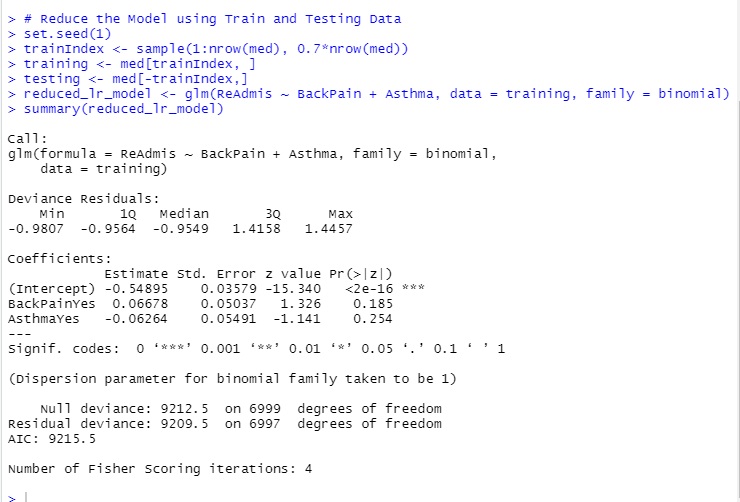
**1.** Using R, an initial model using all the predictor variables was created using all of the predictor variables as described previously. The following screenshot showcases the summary of the model:

A screenshot of a computer program

Description automatically generated

**2.** To evaluate the model and to determine which features to select in order to reduce the model, I will look at the p-values for each variable, as seen from the screenshot above. P-values that are less than 0.05 are seen as statistically significant, however, all of these variables have p-values that are higher than 0.05. Some variables are much closer to 1 than they are to 0, however both BackPain and Asthma are the closest the 0.05. BackPain has a p-value of 0.1763 and Asthma at 0.0862. While these are not less than 0.05, since they are the closest to 0, they are the most significant within this model. Based on these p-values, I am reducing the model to include just these two variables against the target ReAdmis variable. In order to reduce the model using this model evaluation metric, I will start by creating a training index on a 70/30 ratio which will be used to create training and testing data sets. The model will be reduced by using the training data set, and the testing data set will be used to make our predictions for the confusion matrix, as will be demonstrated in part **E2**.

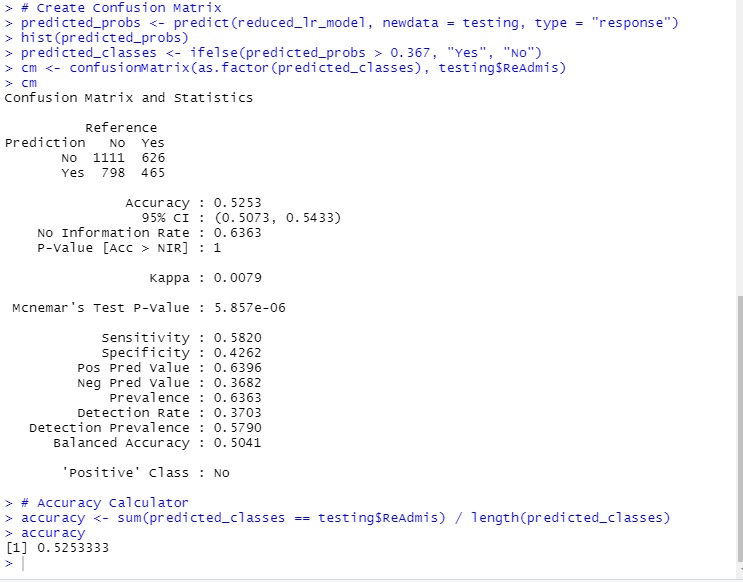
**3.** After reducing the model to just BackPain and Asthma against the ReAdmis variable, using the model evaluation metric as described in the previous step, the following screenshot showcases the output of that reduced model:



**E. Analysis**

**1.** As discussed above, the initial model with all predictors was reduced to include just BackPain and Asthma, as they had the p-values closest to 0.05. When creating the reduced model using these predictors, their p-values increased to 0.185 and 0.254 respectively. Since they were already above 0.05, the increase in their p-values with the reduced model again showcases that these variables are not statistically significant but there are other ways we can evaluate the model. By looking at AIC of each model, we can interpret how good of a fit each model is, with a lower AIC score suggesting a better fit. The initial model has an AIC of 13164 and the reduced model has an AIC of 9215.5. The reduced model is much lower, so even with less predictors, the reduced model is a better fit despite the fact that the predictors are not statistically significant.

**2.** The following screenshot showcases the confusion matrix and the accuracy score, as well as the code in order to create them:



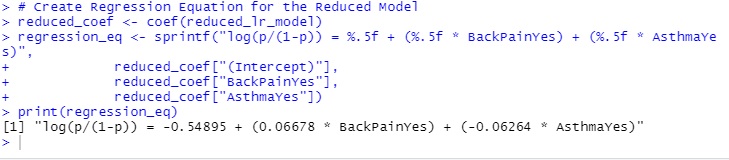
As showcased in the code above, the testing data set that we created earlier was used to create our predicted probabilities with a response type. Using a histogram to determine the center of the predicted probabilities, I then created a set of predicted classes which was then used to predict ReAdmis from the testing data set and create a confusion matrix and accuracy score. As seen in the screenshot, both the confusion matrix and accuracy score give us an accuracy of 0.5253 or 52.53%. That means that our model was able to correctly predict the ReAdmis variable a little over half the time, which suggests that this model is no better than random guessing, since there are only two responses, so everyone has a 50/50 chance to correctly predict.

**3.** The code used to perform the analysis has been attached as an RScript file alongside this written assessment.

**F. Data Summary and Implications**

**1.** The following list is a discussion of the results of the data analysis:

* This screenshot showcases the code in R used to develop the regression equation in R:



As we can see, the regression equation is as follows: the logit expression is equal to -0.54895 (the intercept) + (0.06678 \* BackPainYes) + (-0.0624 \* AsthmaYes).

* The following are interpretations of the coefficients from the equation above:
  + -0.54895: this is the intercept of the equation. When neither BackPain nor Asthma are present, then the chances of being readmitted to the hospital decrease by 0.54895.
  + 0.06678: When a patient has back pain then the chances of being readmitted increase by 0.06678.
  + -0.06264: When a patient has asthma then the chances of being readmitted decrease by 0.06264.
* As stated before, the p-values of BackPain and Asthma in the reduced model increased to 0.185 and 0.254 respectively. Neither predictor has a p-value less than 0.05 before and their respective p-values have only increased with the reduced model. This indicates that neither predictor is statistically significant in predicting the readmissions variable, but even though they are not statistically significant, they could still be practically significant. Based on the coefficient of 0.06678, a patient having back pain does seem to have some effect on increasing the odds of readmission, even if it a very small increase. In a real medical environment, even the smallest percent chance should be taken seriously, but more testing will need to be done to truly determine the practical significance of the predictors.
* The data analysis does suffer with some limitations. For starters the reason for original admittance and readmission is not mentioned in the data. A patient could have been admitted for an injury which has nothing to do with their condition variables but later readmitted for a different medical reason, and the data does not explore this. The data analysis also assumes no multicollinearity but if the predictors have multicollinearity, then this would affect the coefficients of the regression equations. P-values also do not capture practical significance, as stated above, so further testing in a real medical environment is needed in order to truly ascertain the significance of these predictors.

**2.** Based upon the complete data analysis, I cannot recommend that the hospital treat patients more thoroughly or any differently based on if that patient has one or more of the condition variables. I recommend that the hospital continue treating everyone as they have been while further testing is done to determine the conditions significance.

**G. Sources**

Zach. “The 6 Assumptions of Logistic Regression (with Examples).” Statology, 13 Oct. 2020, www.statology.org/assumptions-of-logistic-regression/.