D208: Predictive Modeling

Task 2

Lucas Benish

Western Governors University

# Part 1: Research Question

# Describe the purpose of this data analysis by doing the following:

## Summarize **one** research question that is relevant to a real-world organizational situation captured in the data set you have selected and that you will answer using logistic regression.

Using the dataset we have gathered can we predict what patients, based on comorbidities and/or underlying factors, may readmit to our organization in the future?

## Define the objectives or goals of the data analysis. Ensure that your objectives or goals are reasonable within the scope of the data dictionary and are represented in the available data.

The goal of this data analysis is to create a logistic regression equation to predict patient readmission status as ReAdmis (Yes) or ReAdmis (No), using as few variables as possible while still providing a quality prediction. This readmission status is noted for all patients within the current dataset, as are patients’ demographic information, socioeconomic status, and medical history. These may all be variables that play a role in predicting readmission status, and will be considered within the initial logistic regression.

# Part II: Method Justification

# Describe logistic regression methods by doing the following:

## Summarize the assumptions of a logistic regression model

The assumptions needed for a logistic regression are different from those of a linear regression model. Logistic regression models are based on the following assumptions (Statology, 2020):

* The dependent variable is binary, e.g. Yes/No (as the ReAdmis variable is)
* The measurements gathered are independent from one another, such that we are not utilizing repeated measures
* The independent variables are not highly correlated to each other
* Outliers are minimal/not extreme
* A linear relationship can be seen between the logit of the response variable and the independent variable(s)
* The sample is large

## Describe the benefits of using the tool(s) you have chosen (i.e., Python, R, or both) in support of various phases of the analysis

One of the benefits to utilizing R once again is that I am becoming much more comfortable coding in R over python. R has various packages that lend themselves well to data preparation/cleaning, data visualization, and logistic and linear regression modeling. These packages include, but are not limited to: naniar, ggplot, tidyr, and broom.

## Explain why logistic regression is an appropriate technique to analyze the research question summarized in Part 1.

Logistic regression is an appropriate technique to answer the question, “Can we predict which patients will readmit,” because the answer to the question of ReAdmis is categorized as a “yes” or “no.” This type of response variable lends itself to logistic regression due to only having the two responses, whereas if the response variable were continuous then linear regression would be more appropriate. Categorizing patients based on potential for readmission will allow providers to educate patients before they leave the hospital after an initial stay to try and reduce readmission risk.

# Part III: Data Preparation

# Summarize the data preparation process for logistic regression by doing the following:

## Describe your data preparation goals and the data manipulations that will be used to achieve the goals.

Like the preparation required of the dataset prior to use in linear regression, I want to ensure that there are no null values, duplicate values, nor outliers within the dataset. The dataset is also the same as the linear regression dataset, thus I am confident claiming that there are no null values, no duplicates, and 467 outliers that will be removed rather than imputed, as they make up less than 5% of the data.

The next goal of the data preparation phase is to visualize the potential explanatory variables with the response variable, ReAdmis, or the numeric version, ReAdmisNumeric. Each variable will be displayed in a univariate visualization, such as histogram for numeric variables and bar graph for categorical variables. Furthermore, each potential explanatory variable will be displayed with the response variable in bivariate visualizations. Categorical variables and the response variable will be displayed in bar graphs, while numeric variables and the response variable will be displayed in scatter plots.

## Discuss the summary statistics, including the target variable and *all* predictor variables that you will need to gather from the data set to answer the research question.

The variable that I am exploring as my response variable is Readmis, which is readmission status and is either “Yes” or “No,” which will be converted to 1 or 0, respectively, for the sake of analysis.

To begin, I will consider all potential explanatory variables from the following list, which includes almost all other variables except for survey items, TotalCharge, and Additional\_charges. The latter two of these variables will be addressed below. Explanatory variables that will be considered are as follows:

Children, the number of children that a patient has (numeric)

Age, a patient’s age (numeric)

Income, a patient’s income (numeric)

Marital, a patient’s marital status (categorical)

Gender, a patient’s gender identity (categorical)

ReAdmis, whether the hospital stay was a readmission (categorical)

VitD\_levels, a patient’s Vitamin D levels (numeric)

Doc\_visits, a patient’s number of doctor visits during a stay (numeric)

Full\_meals\_eaten, number of meals eaten in entirety during stay (numeric)

vitD\_supp, number of Vitamin D supplements given (numeric)

Soft\_drink, whether the patient consumes soda (categorical)

Initial\_admin, how the patient was initially admitted to the hospital (categorical)

HighBlood, whether the patient has high blood pressure or not (categorical)

Stroke, whether the patient has suffered a stroke or not (categorical)

Complication\_risk, the patient’s level of complexity (categorical)

Overweight, whether the patient is considered overweight or not (categorical)

Arthritis, whether the patient has arthritis or not (categorical)

Diabetes, whether the patient has diabetes or not (categorical)

Hyperlipidemia, whether the patient has hyperlipidemia or not (categorical)

BackPain, whether the patient has hyperlipidemia or not (categorical)

Anxiety, whether the patient has anxiety or not (categorical)

Allergic\_rhinitis, whether the patient has allergic rhinitis or not (categorical)

Reflux\_esophagitis, whether the patient has reflux esophagitis or not (categorical)

Asthma, whether the patient has asthma or not (categorical)

Services, what services the patient received during their stay (categorical)

Initial\_days, how many days a patient initially spends in the hospital for their stay (numeric)

TotalCharge is being withheld because this was determined during the cleaning and exploration process to be highly correlated to Initial\_days. The correlation was nearly 1.0, meaning that only one of these variables would be necessary for our regression model.Furthermore, TotalCharge is determined based on a patient’s hospital stay, and would therefore make more sense as a response variable, rather than helping to predict if a patient will readmit.

Additional\_charges are being withheld because this is also a variable that would be determined in response to a hospital stay, similarly to TotalCharge, rather than assisting in determining the type of hospital stay predicted.

## Explain the steps used to prepare the data for the analysis, including the annotated code.

As I stated previously, I will initially be cleaning my dataset to ensure there are no null values, duplicates, nor outliers within the data. Prior to beginning this process, I familiarize myself with the data via summary() and create an extra dataframe in case something goes awry with the cleaning process.

#general glimpse of dataset

summary(medical\_clean)

#copy data set for outliers if needed

medical\_clean2 <- medical\_clean

Next, I visualize the missingness of my data using vis\_miss() from the Naniar package, and assess for duplicate rows with the following code:

#assessing missingness and duplicates

vis\_miss(medical\_clean)

str(medical\_clean)

sum(duplicated(medical\_clean))

Following this, I need to prepare the data further by treating any outliers that may be present. To first determine the presence of outliers, I create z-score columns with the nomenclature “variable\_z” and add them to my extra dataset, medical\_clean2.

#z-score columns

medical\_clean2$children\_z <- scale(x=medical\_clean2$Children)

medical\_clean2$age\_z <- scale(x=medical\_clean2$Age)

medical\_clean2$income\_z <- scale(x=medical\_clean2$Income)

medical\_clean2$vitd\_levels\_z <- scale(x=medical\_clean2$VitD\_levels)

medical\_clean2$doc\_visits\_z <- scale(x=medical\_clean2$Doc\_visits)

medical\_clean2$full\_meals\_eaten\_z <- scale(x=medical\_clean2$Full\_meals\_eaten)

medical\_clean2$vitd\_supp\_z <- scale(x=medical\_clean2$vitD\_supp)

medical\_clean2$initial\_days\_z <- scale(x=medical\_clean2$Initial\_days)

medical\_clean2$totalcharge\_z <- scale(x=medical\_clean2$TotalCharge)

medical\_clean2$additional\_charges\_z <- scale(x=medical\_clean2$Additional\_charges

These columns of z-scores are then filtered for values greater than 3 or less than -3 standard deviations from the mean. The output of these filters are stored as vectors, which will later be used to treat all rows with outliers.

#outlier vectors

children\_outliers <- which(medical\_clean2$children\_z >3 | medical\_clean2$children\_z < -3)

age\_outliers <- which(medical\_clean2$age\_z >3 | medical\_clean2$age\_z < -3)

income\_outliers <- which(medical\_clean2$income\_z >3 | medical\_clean2$income\_z < -3)

vitd\_levels\_outliers <- which(medical\_clean2$vitd\_levels\_z >3 | medical\_clean2$vitd\_levels\_z < -3)

doc\_visits\_outliers <- which(medical\_clean2$doc\_visits\_z >3 | medical\_clean2$doc\_visits\_z < -3)

full\_meals\_eaten\_outliers <- which(medical\_clean2$full\_meals\_eaten\_z >3 | medical\_clean2$full\_meals\_eaten\_z < -3)

vitd\_supp\_outliers <- which(medical\_clean2$vitd\_supp\_z >3 | medical\_clean2$vitd\_supp\_z < -3)

initial\_days\_outliers <- which(medical\_clean2$initial\_days\_z >3 | medical\_clean2$initial\_days\_z < -3)

total\_charge\_outliers <- which(medical\_clean2$totalcharge\_z >3 | medical\_clean2$totalcharge\_z < -3)

additional\_charges\_outliers <- which(medical\_clean2$additional\_charges\_z >3 | medical\_clean2$additional\_charges\_z < -3)

After these vectors are created, another vector containing all unique rows with outliers is created which will be used to remove the outliers from the dataset. Following this removal, another copied, clean dataset will be created under the name, medical\_clean3.

#treating outliers

unique\_outliers <- unique(c(children\_outliers, doc\_visits\_outliers, full\_meals\_eaten\_outliers, income\_outliers, vitd\_levels\_outliers, vitd\_supp\_outliers))

medical\_clean3 <- medical\_clean2[-unique\_outliers, ]

medical\_clean3 <- subset(medical\_clean3, select = -c(age\_z, income\_z, vitd\_levels\_z, doc\_visits\_z, full\_meals\_eaten\_z, vitd\_supp\_z, initial\_days\_z, totalcharge\_z, additional\_charges\_z, children\_z))

This completes the cleaning portion of the task, but one more column was created prior to further exploring the dataset and the relationship between variables. ReAdmis is a categorical variable, hence the use of logarithmic regression to predict whether a patient will readmit. For the purposes of some visualizations, and creation of the model requiring responses between 0 and 1, ReAdmisNumeric was created as another column with the following code:

#creating numeric response variable

medical\_clean3$ReAdmisNumeric <- as.numeric(as.factor(medical\_clean3$ReAdmis))

summary(medical\_clean3$ReAdmisNumeric)

medical\_clean3$ReAdmisNumeric <- medical\_clean3$ReAdmisNumeric-1

hist(medical\_clean3$ReAdmisNumeric)

## Generate univariate and bivariate visualizations of the distributions of variables in the cleaned data set. Include the target variable in your bivariate visualizations.

Univariate and bivariate visualizations of all potential variables were created. For univariate visualizations, histograms were created for numeric variables and bar graphs created for categorical variables. For bivariate visualizations, scatter plots were created for numeric variables on the x-axis with ReAdmisNumeric, the created numeric response variable, on the y-axis. Furthermore, bivariate bar graphs were created with the previously noted categorical variables on the x-axis, with the fill of the bars color coded based on ReAdmis status. Examples of these will be noted below, including a ‘histogram’ of the response variable, ReAdmisNumeric.

The code utilized to create the univariate visualizations is as follows:

#univariate visualizations

ggplot(medical\_clean3, aes(x=ReAdmisNumeric)) + geom\_histogram()

ggplot(medical\_clean3, aes(x=Children)) + geom\_histogram(binwidth=1)

ggplot(medical\_clean3, aes(x=Age)) + geom\_histogram(binwidth=1)

ggplot(medical\_clean3, aes(x=Income)) + geom\_histogram(bins=60)

ggplot(medical\_clean3, aes(x=Marital)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Gender)) + geom\_bar()

ggplot(medical\_clean3, aes(x=VitD\_levels)) + geom\_histogram(bins=30)

ggplot(medical\_clean3, aes(x=Doc\_visits)) + geom\_histogram(binwidth=1)

ggplot(medical\_clean3, aes(x=Full\_meals\_eaten)) + geom\_histogram(binwidth=1)

ggplot(medical\_clean3, aes(x=vitD\_supp)) + geom\_histogram(bins=3)

ggplot(medical\_clean3, aes(x=Soft\_drink)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Initial\_admin)) + geom\_bar()

ggplot(medical\_clean3, aes(x=HighBlood)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Stroke)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Complication\_risk)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Overweight)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Arthritis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Diabetes)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Hyperlipidemia)) + geom\_bar()

ggplot(medical\_clean3, aes(x=BackPain)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Anxiety)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Allergic\_rhinitis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Reflux\_esophagitis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Asthma)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Services)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Initial\_days)) + geom\_histogram(binwidth=1)

ggplot(medical\_clean3, aes(x=TotalCharge)) + geom\_histogram(bins=60)

ggplot(medical\_clean3, aes(x=Additional\_charges)) + geom\_histogram(bins=60)

The code to create the bivariate visualizations is as follows:

#bivariate visualizations

ggplot(medical\_clean3, aes(x=Children, y=ReAdmisNumeric)) + geom\_point() + geom\_smooth(method='glm', se=FALSE)

ggplot(medical\_clean3, aes(x=Age, y=ReAdmisNumeric)) + geom\_point() + geom\_smooth(method='glm', se=FALSE)

ggplot(medical\_clean3, aes(x=Income, y=ReAdmisNumeric)) + geom\_point() + geom\_smooth(method='glm', se=FALSE)

ggplot(medical\_clean3, aes(x=Marital, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Gender, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=VitD\_levels, y=ReAdmisNumeric)) + geom\_point() + geom\_smooth(method='glm', se=FALSE)

ggplot(medical\_clean3, aes(x=Doc\_visits, y=ReAdmisNumeric)) + geom\_point() + geom\_smooth(method='glm', se=FALSE)

ggplot(medical\_clean3, aes(x=Full\_meals\_eaten, y=ReAdmisNumeric)) + geom\_point() + geom\_smooth(method='glm', se=FALSE)

ggplot(medical\_clean3, aes(x=vitD\_supp, y=ReAdmisNumeric)) + geom\_point() + geom\_smooth(method='glm', se=FALSE)

ggplot(medical\_clean3, aes(x=Soft\_drink, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Initial\_admin, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=HighBlood, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Stroke, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Complication\_risk, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Overweight, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Arthritis, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Diabetes, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Hyperlipidemia, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=BackPain, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Anxiety, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Allergic\_rhinitis, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Reflux\_esophagitis, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Asthma, fill=ReAdmis)) + geom\_bar()

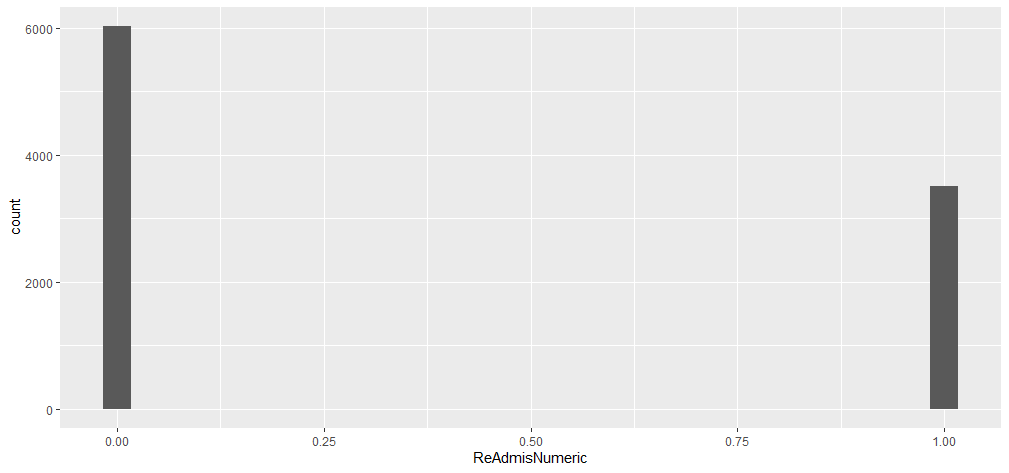
ggplot(medical\_clean3, aes(x=Services, fill=ReAdmis)) + geom\_bar()

ggplot(medical\_clean3, aes(x=Initial\_days, y=ReAdmisNumeric)) + geom\_point() + geom\_smooth(method='glm', se=FALSE)

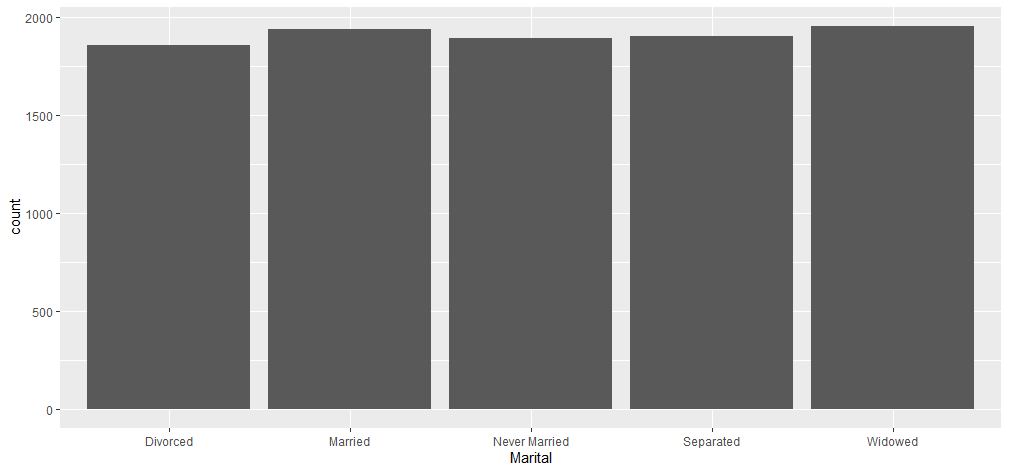
ggplot(medical\_clean3, aes(x=TotalCharge, y=ReAdmisNumeric)) + geom\_point() + geom\_smooth(method='glm', se=FALSE)

ggplot(medical\_clean3, aes(x=Additional\_charges, y=ReAdmisNumeric)) + geom\_point() + geom\_smooth(method='glm', se=FALSE)

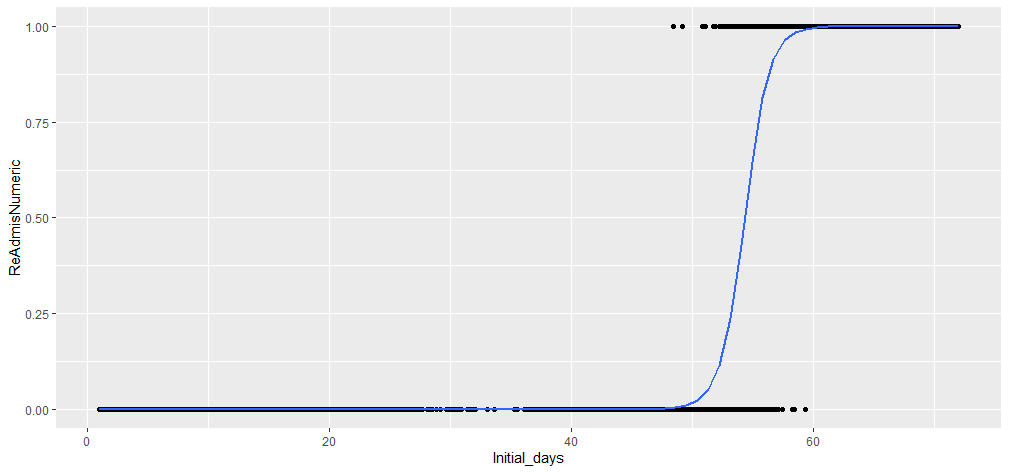
* The following is a ‘histogram’, or univariate visualization, of ReAdmisNumeric, the response variable of interest



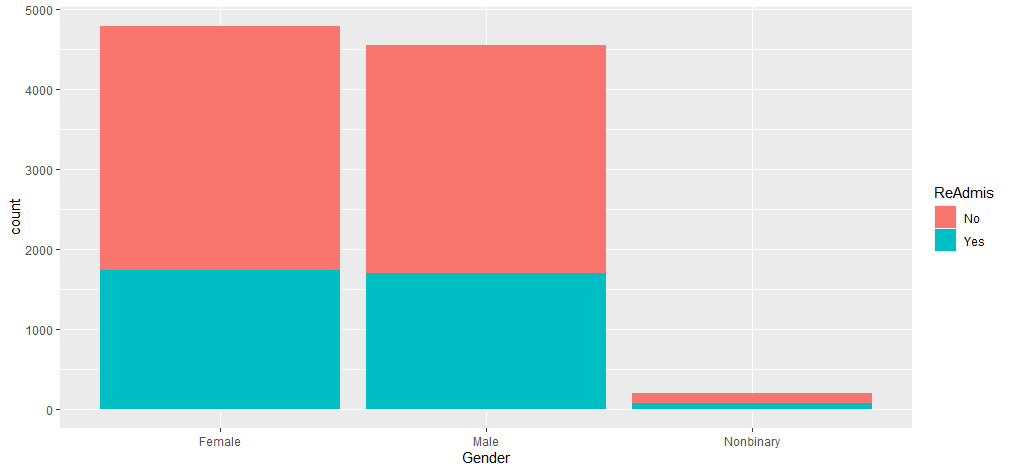
* The following is a bar graph, or univariate visualization, of Marital, a categorical explanatory variable



* The following is a scatterplot, a bivariate representation of the relationship between Initial\_days, a numeric explanatory variable, and ReAdmisNumeric, the numeric version of the response variable



* The following is a bar graph of a categorical explanatory variable, Gender, with var fill based on the status of ReAdmis, the categorical version of the response variable



After performing the visualizations, it was noted that the relationships between Initial\_days and ReAdmisNumeric and TotalCharge and ReAdmisNumeric was nearly identical. For this reason, a scatterplot of Initial\_days and TotalCharge was visualized, and correlation was calculated as well. This was done to see if it was worth including both variables or removing one.

The following is the code utilized for this:

#noticed that Initial\_days and TotalCharge nearly identical when plotted with ReAdmisNumeric

ggplot(medical\_clean3, aes(x=TotalCharge, y=Initial\_days)) + geom\_point() + geom\_smooth(method='lm')

cor(x=medical\_clean3$TotalCharge, y=medical\_clean3$Initial\_days)

## Provide a copy of the prepared data set.

A copy of the cleaned dataset will be available as an attachment with the upload. The code used to write the clean CSV is as follows:

#writing cleaned csv to folder

write.csv(medical\_clean3, 'C:\\Users\\lgben\\OneDrive\\Desktop\\MSDA\\D208 - Predictive Modeling\\medical\_clean3.csv')

# Part IV: Model Comparison and Analysis

# Compare an initial and a reduced logistic regression model by doing the following:

## Construct an initial logistic regression model from *all* predictors that were identified in Part C2

As discussed above, all previously visualized explanatory variables will be included in the initial model, except for TotalCharge and Additional\_charges. The code to create this model is as follows:

#initial logistic regression model (all)

logmodel\_all <- glm(ReAdmisNumeric ~ Children + Age + Income + Marital + Gender + VitD\_levels + Doc\_visits + Full\_meals\_eaten + vitD\_supp + Soft\_drink + Initial\_admin + HighBlood + Stroke + Complication\_risk + Overweight + Arthritis + Diabetes + Hyperlipidemia + BackPain + Anxiety + Allergic\_rhinitis + Reflux\_esophagitis + Asthma + Services + Initial\_days, data=medical\_clean3, family='binomial')

summary(logmodel\_all)

## Justify a statistically based variable selection procedure and a model evaluation metric to reduce the initial model in a way that aligns with the research question.

The summary call above provided an output where I was able to see what explanatory variables were statistically significant for the model. To reduce the model and remove unneeded explanatory variables, variables with p-values <= 0.05 were chosen to remain in the reduced model. This reduced model would therefore utilize less inputs, while maintaining a relative level of significance.

To evaluate whether the reduced model was an improvement over the original, I utilized the Akaike information criterion (AIC). The AIC is a measure that is calculated as AIC = 2K – 2ln(L) and is based on the number of variables in a model (K) and the log-likelihood (ln(L)) of that same model (Statology, 2021). The lower the AIC, the better the model fit. The AIC is used to find a model that explains as much of the variance within the data as possible, while still not allowing for too many variables so as to not overfit the models.

## Provide a reduced logistic regression model. *Note: the output should include a screenshot of each model*.

The reduced logistic regression model is based on the following code, and an output of a summary() call of both the reduced, and original model, will be provided below the summary() call.

#reduced logistic regression model (only previously noted significant at p <= 0.0.05 plus Age for another continuous variable)

logmodel\_reduced <- glm(ReAdmisNumeric ~ Children + Age + Initial\_admin + HighBlood + Stroke + Complication\_risk + Arthritis + Diabetes + Anxiety + Reflux\_esophagitis + Asthma + Services + Initial\_days, data=medical\_clean3, family = 'binomial')

summary(logmodel\_reduced)

The summary of the reduced model is as follows.

Graphical user interface, application

Description automatically generated

The summary of the initial model including all explanatory variables is as follows.

Graphical user interface, application

Description automatically generated

The continuation of the summary call for the initial model including all explanatory variables is as follows.

Graphical user interface, application

Description automatically generated

# Analyze the data set using your reduced logistic regression model by doing the following:

## Explain your data analysis process by comparing the initial and reduced logistic regression models, including the following elements

## The logic of the variable selection technique

For the initial model that includes all explanatory variables, the only logic that was included was removal of TotalCharge due to its near perfect correlation with Initial\_days, and Additional\_charges were removed since Additional\_charges would be determined based on a patient’s stay in the hospital and would therefore be determined in response to a stay.

The reduced model was built based on the significant variables from the initial model. Any explanatory variable in the initial model that had a p-value less than or equal to 0.05 was included in the reduced model. Age was also included in the model, despite it not being significant, to incorporate a continuous variable in case it was needed for visualizations after the model was built.

## The model evaluation metric

The metric that was utilized to determine which model was a better fit was the AIC. As stated earlier, the AIC is used to find a model that explains as much of the variance within the data as possible, while still not allowing for too many variables so as to not overfit the models. The goal is to have a low AIC, and the reduced model had a lower AIC, 669.58, than the model including all variables, 685.67.

## Provide the output and *any* calculations of the analysis you performed, including a confusion matrix. *Note: The output should include the predictions from the refined model you used to perform the analysis*.

The reduced model was used to demonstrate the increased risk that patients have of readmitting to the hospital based upon certain factors. This will be discussed further in F.1.b, where I describe what the coefficients mean, but the code that was utilized to give an example of this is as follows:

#examples of coefficient explanation

#age is numeric, so risk rises with each additional year of life

age\_risk <- exp(0.003)

age\_risk\*18

age\_risk\*60

#binary variable: either ED admission, or not

increased\_ED\_admis\_risk <- exp(2.58)

increased\_ED\_admis\_risk

#binary variable, either received MRI services or didn't

services\_mri\_risk <- exp(3.08)

services\_mri\_risk

The reduced model was utilized to create a confusion matrix, which demonstrates the relationship of predictions relative to actual status of readmission or not. To do this, the logarithmic model was initially tidied using the augment() function from the broom package. A column for predictions of readmission was then added to the dataframe, where values were rounded to the nearest whole of either 0 or 1, rather than their specific, fitted value. This will be used in the future to determine accuracy, specificity, and sensitivity of the model via a confusion matrix.

The confusion matrix was then created from the same augmented model. The output of the matrix creation can be seen here.

A screenshot of a computer

Description automatically generated with medium confidence

When it came to predicting what individuals would be admitted to the hospital via readmission, the model correctly predicted that 3446 individuals would be readmitted and incorrectly predicted that 71 individuals would not readmitted. Furthermore, with predicting individuals that would not be readmitted to the hospital, the model correctly predicted that 5954 individuals would not be readmitted and incorrectly predicted that 62 would be readmitted.

The matrix was then utilized to determine the values for true negatives (correctly predicting NO readmission), true positives (correctly predicting presence of readmission), false negatives (incorrectly predicting NO readmission, when patient really did readmit), and false positives (incorrectly predicting readmission, when patient DID NOT readmit). From these true and false negatives and positives, specificity, sensitivity, and accuracy of the model were calculated. All 3 of these values were greater than 0.98, indicating that the model may be overfit, as there is usually a trade off between specificity and sensitivity.

## Provide the code used to support the implementation of the logistic regression models.

The logarithmic model was augmented and had supporting columns, such as ReAdmis\_Predicted, added with the following code. Visualization of the fitted values was also performed, to explore the model. Furthermore, a confusion matrix was made and from this accuracy, specificity, and sensitivity of the model were calculated.

#augmenting for predictions

logmodel\_reduced\_aug <- logmodel\_reduced %>%

augment(type.predict = 'response')

logmodel\_reduced\_aug <- logmodel\_reduced\_aug %>%

mutate(ReAdmis\_Predict = round(.fitted))

ggplot(logmodel\_reduced\_aug, aes(x=Initial\_days, y=.fitted)) + geom\_point() + geom\_smooth(method='glm', method.args=list(family='binomial'), se=FALSE)

ggplot(logmodel\_reduced\_aug, aes(x=Initial\_days, y=ReAdmis\_Predict)) + geom\_point()

summary(logmodel\_reduced\_aug)

#confusion matrix

conf\_matrix <- logmodel\_reduced\_aug %>%

select(ReAdmisNumeric, ReAdmis\_Predict) %>%

table()

#calculations

cm\_true\_neg <- conf\_matrix[1, 1]

cm\_true\_pos <- conf\_matrix[2, 2]

cm\_false\_neg <-conf\_matrix[2, 1]

cm\_false\_pos <- conf\_matrix[1, 2]

sensitivity <- cm\_true\_pos / (cm\_true\_pos + cm\_false\_neg)

sensitivity

specificity <- cm\_true\_neg / (cm\_true\_neg + cm\_false\_pos)

specificity

accuracy <- (cm\_true\_pos + cm\_true\_neg)/(cm\_true\_pos + cm\_true\_neg + cm\_false\_neg + cm\_false\_pos)

accuracy

# Part V: Summary and Implications

# Summarize your findings and assumptions by doing the following:

## Discuss the results of your data analysis, including the following elements:

## A regression equation for the reduced model

The regression equation can be determined using the coefficients from either the summary(logmodel\_reduced) call, or from calling logmodel\_reduced$coefficients. The model can be written out as:

ReAdmisNumeric = 0.12\*Children + 0.003\*Age + 2.58\*Initial\_admin[Emergency Admission] + 0.80\*Initial\_admin[Observation Admission] + 0.88\*HighBlood[Yes] + 1.63\*Stroke[Yes] – 1.87\*Complication\_risk[Low] – 0.23\*Complication\_risk[Medium] – 1.44\*Arthritis[Yes] + 0.53\*Diabetes[Yes] – 1.04\*Anxiety[Yes] – 0.42\*Reflux\_esophagitis[Yes] – 1.32\*Asthma[Yes] + 1.6\*Services[CT Scan] + 0.053\*Services[Intravenous] + 3.08\*Services[MRI] + 1.48\*Initial\_days – 81.74

## An interpretation of coefficients of the statistically significant variables of the model

The summary(logmodel\_reduced) call shows us the variables included in the model, as well as which variables are significant. Each coefficient can be interpreted as “the change in log odds of having the outcome per unit change” in the associated variable. The coefficient, when used as the exponent in the formula e^coefficient, produces the odds ratio associated with that coefficient. Positive coefficients will increase the odds of having an outcome, in this case ReAdmission, while negative coefficients will decrease the odds of having an outcome (Choueiry, nd). An example of two different variables will be provided after providing each coefficient.

The following variables are not significant within the model: Age, Complication\_risk[Medium], and Services[Intravenous]. The coefficients of Age and Complication\_risk[Medium] are 0.003, and -0.23, respectively.

The following variable is significant at the p <= 0.1 level: Reflux\_esophagitis[Yes]. The coefficient of Reflux\_esophagitis[Yes] is -0.42.

The following variables are significant at the p <= 0.05 level: Children, Diabetes[Yes]. The coefficients of Children and Diabetes[Yes] are 0.12 and 0.53, respectively.

The following variable is significant at the p <= 0.01 level: Initial\_admin[Observation Admission]. The coefficient of Initial\_admin[Observation Admission] is 0.80.

The following variables are significant at the p <= 0.001 level: The intercept of the model, Initial\_admin[Emergency Admission], HighBlood[Yes], Stroke[Yes], Complication\_risk[Low], Arthritis[Yes], Anxiety[Yes], Asthma[Yes], Services[CT Scan], and Services[MRI], and Initial\_days. The coefficients of Initial\_admin[Emergency Admission], HighBlood[Yes], Stroke[Yes], Complication\_risk[Low], Arthritis[Yes], Anxiety[Yes], Asthma[Yes], Services[CT Scan], Services[MRI], and Initial\_days are 2.58, 0.88, 1.63, -1.87, -1.44, -1.04, -1.32, 1.6, 3.08, and 1.48, respectively. The intercept of the model is -81.74.

An example of calculating the odds ratios of two different variables based on their coefficients can be seen as follows:

Age’s coefficient was 0.003. If we then calculate the odds ratio as e^0.003 = 1.003, meaning that an individual’s risk of readmission at 1 year old is 1.003 times that increase in risk of readmission for each additional year of life.

Services[MRI]’s coefficient was 3.08. We can then calculate the odds ratio as e^3.08 = 21.75, meaning that for an individual that had an MRI completed, they were approximately 22 times more likely to readmit relative to those that did not have MRIs completed.

## The statistical and practical significance of the model

Statistically, the model is significant at the p <= 0.05 level. This means that the model can provide statistically significant predictions as to whether a patient will readmit to the hospital.

Practically, this model allows us to quantify the increased risk of readmission, or reduction of risk of readmission, based upon certain patient variables. This will allow hospital staff to further tailor their provided education with specific patients, such as those with other major health conditions like high blood pressure (HighBlood) and history of strokes (Stroke), prior to discharge as these factors are associated with a greater risk of readmission.

## The limitations of the data analysis

One limitation of this data analysis is that Initial\_days as an explanatory variable may not make much sense when thought about practically. A patient does not get admitted the hospital knowing how many days they will stay, but rather this is typically a product of a patient’s complication status, services required, etc. Therefore, the model may need to be re-evaluated in the future to change this.

Another limitation of this model is that it may be overfit to the current data set, and therefore not as applicable to other data sets. The accuracy, specificity, and sensitivity of the model are all above 98%. Typically, there is a tradeoff between specificity and sensitivity where as one rises, the other falls. This would again be a good reason to revisit the data after gathering more and determine if a new model is more appropriate.

## Recommend a course of action based on your results.

I would recommend continued data gathering and revisiting the model within the next 6-12 months to see if there is any change in the significant variables associated with readmission. There may be a change in factors affecting readmission as seasons change, or potentially following a widespread disease. But for the current time, as I stated above, I believe this model will allow caregivers and staff to appropriately tailor patient education to those individuals that need it most, or may be at greatest risk of future readmission, to try and reduce the chances of that happening.

# Part VI: Demonstration

# Provide a Panopto video recording that includes *all* of the following elements:

## A demonstration of the functionality of the code used for the analysis

## An identification of the version of the programming environment

## A comparison of the **two** multiple regression models you used in your analysis

## An interpretation of the coefficients

# List the web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable.

1. Choueiry, G. (n.d.). *Interpret Logistic Regression Coefficients [For Beginners]*. Quantifying Health. Retrieved August 29, 2022, from <https://quantifyinghealth.com/interpret-logistic-regression-coefficients/>
2. N/A, Z. (2020, October 13). *The 6 assumptions of logistic regression (with examples)*. Statology. Retrieved August 29, 2022, from <https://www.statology.org/assumptions-of-logistic-regression/>
3. N/A, Z. (2022, March 13). *How to calculate AIC in R (including examples)*. Statology. Retrieved August 29, 2022, from <https://www.statology.org/aic-in-r/#:~:text=The%20Akaike%20information%20criterion%20(AIC,The%20number%20of%20model%20parameters>
4. YouTube. (2018). *Logistic regression in R, Clearly Explained!!!!* *YouTube*. Retrieved August 29, 2022, from <https://www.youtube.com/watch?v=C4N3_XJJ-jU>.

# Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

# Demonstrate professional communication in the content and presentation of your submission.