D208: Predictive Modeling

Task 1

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 multiple regression.

The question that I would like to answer is:

“Using the data we have gathered, can we predict how long a future patient may need to stay in the hospital based on their presentation?”

This will be answered using multiple linear regression because the response variable, Initial\_days, is a continuous, numeric variable and can have multiple factors affecting it (explanatory variables).

## 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 objective of this data analysis is to create a linear regression model that can be used in predicting how long a patient may stay at the hospital, based on specific, patient-centered factors that may be present at the time of hospital admission.

# Part II: Method Justification

# Describe multiple regression methods by doing the following:

## Summarize the assumptions of a multiple regression model

The assumptions of a linear model can be outlined as follows:

* The data follow a linear trend
* The observations are independent of one another
* The data points in the model are evenly distributed around the regression line
* There is equal variability around the regression line for all explanatory data points

## 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

I have once again chosen R to perform this assessment. R continues to be a more natural programming language for me to use, and the DataCamp courses have been very insightful.

R has specific packages that lend themselves well to creating and interpreting regression models. Tidy and broom allow for improved functionality of the outputted linear models, and improved ease of model interpretation. Dplyr further improves how the code is written for these models by allowing piping from one function to another, creating cleaner, more readable code.

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

Multiple linear regression is an appropriate technique to analyze Initial\_days, which could be interpreted as length of initial stay in a hospital setting, because the response variable is numeric. Linear regression will allow me to visualize the relationship between our numeric response variable, and other explanatory variables that could be categorical or numeric in nature.

This regression will allow our organization to prepare for future hospital staffing needs by predicting which patients may need prolonged stays in the future.

# Part III: Data Preparation

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

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

One goal for the data preparation phase would be to ensure that null values are either removed or imputed, that duplicates are removed as well, and to consider the best course of action with any outliers that are present. With the medical data set being used in this course, there are no null values nor duplicates present. There are a total of 467 unique outliers found, out of 10,000 total observations.

Another goal would be to visualize the response variable, Initial\_days, independently in a univariate plot to determine what type of distribution it has on its own. Furthermore, it would likely be beneficial to visualize Initial\_days along with all the possible explanatory variables in simple, bivariate scatter plots for numeric variables or box plots for categorical variables to get an overview of the relationships that may or may not exist between explanatory variables and my response variable. This will also allow me to visualize any relationships that may need to undergo any sort of transformation, whether that be natural log, square roots, or raising to an exponential value.

## 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 want to explore as my response variable is:

Initial\_days, the number of days that a patient spent in the hospital for an initial stay

All possible explanatory variables that I am going to consider initially, and later reduce to a more succinct model, 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 has asthma or not (categorical)

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

TotalCharge, the total charges for a patient’s hospital stay (numeric)

Additional\_charges, additional charges for a patient’s hospital stay (numeric)

I will not be considering the survey responses as explanatory variables, as surveys were likely completed following a patient’s stay and should have no bearing on their initial length of stay. Furthermore, it will be seen as a comment within my code snippet during the bivariate visualizations that TotalCharges will also be dropped as an explanatory variable. This is because it is nearly perfectly correlated, or colinear, with Initial\_days.

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

As stated above, I will initially be cleaning the data based on the presence of null values, duplicates, and outliers. Prior to checking for missingness and duplicates I gained a sense of familiarity with my data set using summary() and created an extra data frame for future cleaning purposes. This was done as follows:

#general glimpse of dataset

summary(medical\_clean)

#copy data set for outliers if needed

medical\_clean2 <- medical\_clean

No missing values nor duplicates were found within my dataset. The code used to determine this is as follows:

#assessing missingness and duplicates

vis\_miss(medical\_clean)

str(medical\_clean)

sum(duplicated(medical\_clean))

However, a total of 467 unique, observed outliers were noted from the following columns: Children, Doc\_visits, Full\_meals\_eaten, Income, VitD\_levels, and vitD\_supp. These were all dropped in their entirety, as the number of the outliers accounted for less than 5 percent of the dataset. This also allowed application of the same treatment method to all outliers without detracting from the overall dataset, rather than imputing some, leaving some, and removing others. The code utilized to determine presence of outliers, remove them, and create a clean dataset without them is as follows:

#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)

#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)

#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))

Following this, I will visualize univariate and bivariate plots for each variable. For univariate numeric variable visualizations, I will utilize histograms, and for univariate categorical variable visualizations I will utilize bar charts. For bivariate visualizations, I will utilize scatter plot for each numeric explanatory variable with the response variable, Initial\_days, and boxplots for categorical explanatory variables and the response variable, Initial\_days.

I aim to familiarize myself with the relationships that may exist between the potential explanatory variables and the response variable prior to creating a regression model. I know that when multiple factors are included in a regression model that relationships can change, but this will allow me to begin hypothesizing which explanatory variables may hold the most weight in the model. This will also help me to determine which numeric variables may need a transformation for inclusion within the model. The following is the code utilized for all the visualizations:

#univariate visualizations

ggplot(medical\_clean3, aes(x=Initial\_days)) + geom\_boxplot()

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

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

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

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

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

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

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

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

ggplot(medical\_clean3, aes(x=Doc\_visits)) + geom\_histogram(bins=10)

ggplot(medical\_clean3, aes(x=Full\_meals\_eaten)) + geom\_histogram(bins=6)

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=TotalCharge)) + geom\_histogram(bins=60)

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

#bivariate visualizations

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

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

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

ggplot(medical\_clean3, aes(x=Gender, y=Initial\_days)) + geom\_boxplot()

ggplot(medical\_clean3, aes(x=Marital, y=Initial\_days)) + geom\_boxplot()

ggplot(medical\_clean3, aes(x=ReAdmis, y=Initial\_days)) + geom\_boxplot()

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

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

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

ggplot(medical\_clean3, aes(x=Soft\_drink, y=Initial\_days)) + geom\_boxplot()

ggplot(medical\_clean3, aes(x=Initial\_admin, y=Initial\_days)) + geom\_boxplot()

ggplot(medical\_clean3, aes(x=HighBlood, y=Initial\_days)) + geom\_boxplot()

ggplot(medical\_clean3, aes(x=Stroke, y=Initial\_days)) + geom\_boxplot()

ggplot(medical\_clean3, aes(x=Complication\_risk, y=Initial\_days)) + geom\_boxplot()

ggplot(medical\_clean3, aes(x=Overweight, y=Initial\_days)) + geom\_boxplot()

ggplot(medical\_clean3, aes(x=Arthritis, y=Initial\_days)) + geom\_boxplot()

ggplot(medical\_clean3, aes(x=Diabetes, y=Initial\_days)) + geom\_boxplot()

ggplot(medical\_clean3, aes(x=Hyperlipidemia, y=Initial\_days)) + geom\_boxplot()

ggplot(medical\_clean3, aes(x=BackPain, y=Initial\_days)) + geom\_boxplot()

ggplot(medical\_clean3, aes(x=Anxiety, y=Initial\_days)) + geom\_boxplot()

ggplot(medical\_clean3, aes(x=Allergic\_rhinitis, y=Initial\_days)) + geom\_boxplot()

ggplot(medical\_clean3, aes(x=Reflux\_esophagitis, y=Initial\_days)) + geom\_boxplot()

ggplot(medical\_clean3, aes(x=Services, y=Initial\_days)) + geom\_boxplot()

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

cor(medical\_clean3$Initial\_days, medical\_clean3$TotalCharge)

#TotalCharge needs to be withheld as an explanatory variable, nearly perfectly correlated with Initial\_days

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

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

Included here are an example of one of each of the following for explanatory variables:

* histogram for univariate visualization of numeric explanatory variable

Chart, histogram

Description automatically generated

* bar chart for univariate visualization of categorical explanatory variable

Chart, bar chart

Description automatically generated

* histogram for univariate visualization of response variable, Initial\_days

Chart, histogram

Description automatically generated

* scatter plot for bivariate visualization of numeric explanatory variable with response variable, Initial\_days

Chart, rectangle

Description automatically generated

* box plot for bivariate visualization of categorical explanatory variable with response variable, Initial\_days

Chart, box and whisker chart

Description automatically generated

## Provide a copy of the prepared data set.

Please see the attached dataset, medical\_clean3.

# Part IV: Model Comparison and Analysis

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

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

For comparison and future use in a stepwise selection, a model with only the intercept for Initial\_days was fit with the following code:

#model with intercept only

model\_initial\_only <- lm(Initial\_days ~ 1, data=medical\_clean3)

Again, for comparison of R2 and RMSE values, a simple linear model was fit as well with the following code:

#simple example model

simple\_model <- lm(Initial\_days ~ HighBlood + Income + Gender, data=medical\_clean3)

summary(simple\_model)

rmse\_simple <- sqrt(mean((simple\_model$fitted.values - medical\_clean3$Initial\_days)^2))

hist(residuals(simple\_model))

ggplot(simple\_model, aes(x=.fitted, y=.resid)) + geom\_point() + geom\_hline(yintercept=0)

Finally, the model fitting every explanatory variable was fit using the following code:

#model with all explanatory variables

model\_all <- lm(Initial\_days ~ Children + Age + Income + Marital + Gender + ReAdmis + 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 + Additional\_charges, data=medical\_clean3)

summary(model\_all)

rmse\_all <- sqrt(mean((model\_all$fitted.values - medical\_clean3$Initial\_days)^2))

#significant variables: Intercept, ReAdmis(Yes) Initial\_admin(Observation), Complication\_risk(Low), Arthritis(Yes), Anxiety(Yes), Allergic\_rhinitis(Yes), Services(CT Scan)

## 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 method that I chose to utilize for reducing my model is a “forward-directed” stepwise selection (Statsology, 2019). This step() function (R documentation, n.d.) determines the most significant variables within a dataset to utilize in a multiple regression model via addition, removal, or both addition and removal, of single variables at a time while determining how this addition or subtraction of a variable affects the measured Akaike information criterion (AIC) for the model with each change. 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 (Statsology, 2021). This is meant to help find a model that explains as much of the variance within the data as possible, while still not allowing for too many variables to not overfit.

## Provide a reduced multiple regression model that includes *both* categorical and continuous variables. *Note: the output should include a screenshot of each model*.

The reduced multiple regression model that I selected for use was the model outputted from the stepwise selection process. Relative to the model containing all the variables, this reduced model had only a slightly poorer adjusted R2 value, by 0.0002, as well as slightly worse root mean square error, by 0.02. Furthermore, the number of variables included in the stepwise model was greatly reduced, and the model was made up mostly of factors specific to the patient, such as comorbidities, complication risk, and readmission/initial admission status. One noted improvement between the initial model, model\_all, and the reduced model, better\_fit, is that the AIC of better\_fit was smaller.

Below: summary(model\_all)

Graphical user interface

Description automatically generated

Below: summary(better\_fit2), which was later renamed to simply ‘better\_fit’, followed by better\_fit2$anova where one can see the reduction in AIC values as additional variables are included.

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

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

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

## The logic of the variable selection technique

The variable selection technique that I utilized was the step() function in R. This allows for the addition, subtraction, or both addition and subtraction, of explanatory variables in a stepwise fashion to a model. When each step is taken, the AIC is calculated, and steps are taken until an insignificant change in the AIC is noted. This method therefore only keeps variables that are significant due to their coefficient, or their interaction with another variable already present in the model. This is also a more succinct option than adding/subtracting variables manually. Furthermore, if only the variables that were deemed significant from the original model were chosen, the explanatory variables utilized would have been different, and included Services received in the hospital while excluding HighBlood and Full\_meals\_eaten.

## The model evaluation metric

Multiple metrics were considered for comparison between models. Root mean square error, or RMSE, was calculated for each model, as was AIC. RMSE is a measure of the variability of the residuals of the fitted model and AIC has been discussed previously.

## A residual plot

The residual plots for both the full (model\_all) and reduced (better\_fit) models can be seen below, respectively. Both are noted to be bimodal, which is in line with the original data values from medical\_clean3$Initial\_days. They are very similar, with better\_fit having slightly less horizontal spreadChart, histogram

Description automatically generatedA picture containing text, light

Description automatically generated

## Provide the output and *any* calculations of the analysis you performed, including the model’s residual error. *Note: The output should include the predictions from the refined model you used to perform the analysis*.

The RMSE and AIC for both the full and better\_fit model were both calculated manually, as well as noted with outputs from the models themselves. These will be noted below.

rmse\_all <- sqrt(mean((model\_all$fitted.values - medical\_clean3$Initial\_days)^2))

rmse\_better <- sqrt(mean((better\_fit$fitted.values - medical\_clean3$Initial\_days)^2))

The AIC() calls can be seen below as well.

Graphical user interface, text, application

Description automatically generated

The code utilized to create an explanatory\_data frame, and then incorporate predictions from the better\_fit model, is copied below. This data was then visualized with a histogram to see the distribution of the predictions. The predictions follow a similar, bimodal distribution to the original data, however with less variance.

#predicting

explanatory\_data <- expand\_grid(

ReAdmis = unique(medical\_clean3$ReAdmis),

Initial\_admin = unique(medical\_clean3$Initial\_admin),

Complication\_risk = unique(medical\_clean3$Complication\_risk),

Arthritis = unique(medical\_clean3$Arthritis),

Anxiety = unique(medical\_clean3$Anxiety),

Allergic\_rhinitis = unique(medical\_clean3$Allergic\_rhinitis),

HighBlood = unique(medical\_clean3$HighBlood),

Full\_meals\_eaten = 0:8

)

predicted\_data <- explanatory\_data %>%

mutate(Initial\_days = predict(better\_fit, explanatory\_data))

summary(predicted\_data)

ggplot(predicted\_data, aes(x=Initial\_days)) + geom\_histogram()

Chart, histogram

Description automatically generated

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

The R code file in its entirely will be attached to the submission, please see attached. Beyond that, the remaining code that has not yet been seen here is below:

#checking model\_all assumptions

hist(residuals(model\_all))

ggplot(model\_all, aes(x=.fitted, y=.resid)) + geom\_point() + geom\_hline(yintercept=0)

#it's an 'okay' fit; all .fitted remain bimodal

#stepwise selection for model

better\_fit <- step(model\_initial\_only, direction = 'forward', scope=formula(model\_all), trace=0)

summary(better\_fit)

better\_fit$anova

better\_fit$call

better\_fit$coefficients

hist(better\_fit$residuals)

ggplot(better\_fit, aes(x=.fitted, y=.resid)) + geom\_point() + geom\_hline(yintercept=0)

rmse\_better <- sqrt(mean((better\_fit$fitted.values - medical\_clean3$Initial\_days)^2))

# 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 utilizing the coefficients from the better\_fit model. The code used to determine the coefficients was seen above in the better\_fit$coefficients call. The output of that call is:

Graphical user interface, text, application

Description automatically generated

From this, we can create the regression equation as follows:

Initial\_days = 17.85 + 46.55\*ReAdmis - 1.63\*Initial\_admin(Emergency Admission) - 0.31\*Initial\_admin(Observation Admission) + 1.03\*Complication\_risk(Low) - 0.31\*Complication\_risk(Medium) + 0.79\*Arthritis + 0.61\*Anxiety + 0.49\*Allergic\_rhinitis - 0.48\*HighBlood - 0.22\*Full\_meals\_eaten

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

What these coefficients represent is the change in number of Initial\_days that would be produced, if a 1 unit change in a single variable took place while all other variables remained constant. Furthermore, we can also interpret from the model that if all explanatory variables were 0 or “No,” a patient would be predicted to spend 17.85 days in the hospital.

For example, a patient that is **Readmitted** (ReAdmisYes) from **Observation** (EmergencyAdmissionNo, ObservationAdmissionYes) with **Low** complication risk (Complication\_riskLowYes, Complication\_riskMediumNo), with **Arthritis**, **Anxiety**, **Allergic** **rhinitis**, and **High** **blood** **pressure** that **didn’t eat any meals during a stay** would be predicted to stay 66.53 days.

Initial\_days = 17.85 + 46.55\*(1) - 1.63\*(0) - 0.31\*(1) + 1.03\*(1) - 0.31\*(0) + 0.79\*(1) + 0.61\*(1) + 0.49\*(1) - 0.48\*(1) - 0.22\*(0)

A patient who has all the same comorbidities, but who **was not readmitted** (ReAdmisNo) would be predicted to stay only 19.98 days.

Initial\_days = 17.85 + 46.55\*(0) - 1.63\*(0) - 0.31\*(1) + 1.03\*(1) - 0.31\*(0) + 0.79\*(1) + 0.61\*(1) + 0.49\*(1) - 0.48\*(1) - 0.22\*(0)

## The statical and practical significance of the model

The statistical significance of this model can be seen from the p-value of the better\_fit model (previously called better\_fit2 within my script). The model, as a whole, is statistically significant at the p < 0.05 level. This is noted even though some variables are only significant at a p < 0.1 level (HighBlood), or not considered significant at all (Initial\_admin – Observation Admission).

Practically, this model is helpful because it could assist with improving how the hospital is staffed. Practitioners/leaders will be able to assess staffing needs based on patient presentation, admission location, and readmission status, and predict what level of services may be required based on patients’ medical history and admission status. This model also does not require that leaders make decisions based on every single patient factor but can focus in on a specific few to make decisions more easily.

A screenshot of a computer

Description automatically generated with medium confidence

## The limitations of the data analysis

I would say one limitation of this model is that the variance of the residuals is not quite evenly spread across the two modes. The first mode, around the intercept, has a larder spread of residuals than the 2nd mode around the “readmission” group.

Another limitation is that Full\_meals\_eaten is a variable that would be gathered after a patient has been admitted, rather than prior to admission. This would make it more difficult to utilize this factor in making predictions.

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

I would recommend continuing to gather data from the hospital system, potentially revisiting the gathered metrics quarterly, as well as potentially propose new metrics such as wait time, days since most recent PCP visit, length of stay (in ER), etc. After continued data gathering, it would likely be beneficial to revisit the model and see if the most significant factors remained consistent or if there are new coefficients and variables to consider.

# 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. N/A, Z. (2021, May 20). *How to Calculate AIC in R (Including Examples)*. Statology. Retrieved August 7, 2022, from <https://www.statology.org/aic-in-r/#:~:text=The%20Akaike%20information%20criterion%20(AIC,The%20number%20of%20model%20parameters>
2. N/A, Z. (2019, April 27). *A Complete Guide to Stepwise Regression in R*. Statology. Retrieved August 7, 2022, from <https://www.statology.org/stepwise-regression-r/>
3. *Step: Choose a model by AIC in a stepwise algorithm*. RDocumentation. (n.d.). DataCamp. Retrieved August 7, 2022, from <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/step>

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

Please see above. All in-text citations are provided within the discussion of the step() function for variable selection in E.1.a.

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