Modeling Survival Outcomes Based On Repeated Measurements

February 21, 2020

1 Executive Summary

Vitals, CBCs, and pre-existing health conditions are intuitively a good indicator of one's health, even more so when that person is in critical condition. Measurements of these variables were obtained for 112 independent patients for both the initial time point the patient was admitted to the hospital/treatment provider and the time point just before that patient passed or that patient was discharged. The main purpose of this analysis was to uncover any useful information through appropriate exploratory data analysis and statistical modeling that explores the relationship between the outcome of survival (binary factor variable) and predictor variables (continuous physiological and categorical demographic/status). Several models were fit and compared; these models only utilized data from the initial observations as there were concerns of "leakage" from the response variable to the final observations. Exploratory data analysis proved effective as the predictor variables graphically shown to be correlated to the response variable were used in the initial model (Urine Output, Patient in Shock, Mean Arterial Pressure, Body Surface Index, and Mean Central Venous Pressure). Prediction accuracies were not assessed as there were limitations with the dimension of this dataset, but the final model was proven to be a good model through diagnostic testing.

2 Introduction

Physiological elements such as vitals and CBCs in addition to medical conditions are recorded and measured when patients are admitted to the hospital. These signs, levels, and factors are closely monitored if the patients are critically ill; abnormal levels can signify a decline in the patient's health. Treatments are also given depending on the particular illness or ailment. This dataset contains 112 paired observations of demographic/health status data, initial physiological measurements, and final physiological measurements (the patient either passed or was discharged shortly after).

There's a lack of consistency in this data as the patients are only described as "critically ill" and could be seen for a variety of illnesses. In addition, the time between the initial measurements and the final measurements was likely very different between each patient. Each patient also likely received different treatments between the initial measurements and the final measurements, so essentially this analysis becomes a problem of the profiles of each patient and their physiological elements determining if they will survive or not.

The data was originally displayed long rather than wide, so before any analysis was done, initial and final variables were created for each of the 14 measurements and the observations were flattened into one vector per patient (so instead of matched observations, we have two different sets of measurements for analysis). This turns the data into a repeated measures format, giving option to simultaneously include both variables in a model. Intuitively, it's hypothesized the final observations are more related to the response variable compared to the initial observations. A more interesting question from this data would be to predict the survival outcome based on the initial observations only, since final observations taken just before death/discharge may tell too much information (possibility of response variable leakage). Due to limitations of this data, prediction is not the main focus of this analysis, but a logistic regression model will be fit to explore the relationships between the initial variables and the response. The response variable is a binary factor, the initial instances are independent of each other, multicollinearity will be addressed, and the number of variables included in the model will be minimized due to the limited amount of data. Ultimately a simpler more interpretable model is preferred over something more complex that may violate assumptions.

3 Exploratory Data Analysis

Table 2 in Appendix A displays a brief summary of the variables in this dataset (21 variables total), and summary statistics of the continuous variables are listed at the beginning of Appendix B. There is a response variable that is a binary variable describing if a patient survived, several demographic/health descriptive variables such as age, height, and what kind of shock that patient was in, and 14 continuous physiological elements such as vitals and CBCs. The survival rate was 61.6 percent, so a fair distribution of classes between the response. As noted previously, the data was made wide according to the record (by initial or final measurement). Variables such as ID and Record were dropped. EDA was completed on the initial variables to show relationships between those variables and assess collinearity. Plots were also created that showed a comparison of the measurements between the first and final observations and the outcome of survival.

A correlation plot of the initial variables is displayed on the next page. Obviously, variables such as SBP, DBP, and MAP are correlated. To address this issue, MAP was chosen as the variable to be used in the model as a boxplot shows a significant relationship between MAP and the survival response variable (Figure 2). In Table 3 in Appendix A, a Welch's two-sample-t-test yielded a low p-value and shows that there is enough evidence that a difference in means exists between the Mean Arterial Pressure and Survival outcomes; this means there is an association between these variables.

There's collinearity between the variables CI, AT, MCT, HG, and HCT. Welch's two-sample-t-test was performed on each of these variables and it was determined a difference in means likely didn't exist between these variables and the response; there's not enough evidence there's a significant association between the two since the p-values were too high (not listed in the table). In cases of high collinearity, often times there requires a significant degree of domain knowledge or business side intervention to influence which variables should be kept; you can also look at graphical plots to determine behavior with the response. Boxplots for these initial variables didn't show a significant relationship between these variables and Survival, so for variable comprehension CI and HG were chosen for model fitting. The results of the t-tests and chi-square tests that tested for differences in

means for association with the response concluded that from the initial variables, MAP, UO, BSI, MCVP, and IsShock have a relationship with the Survival outcome based on really low p-values.

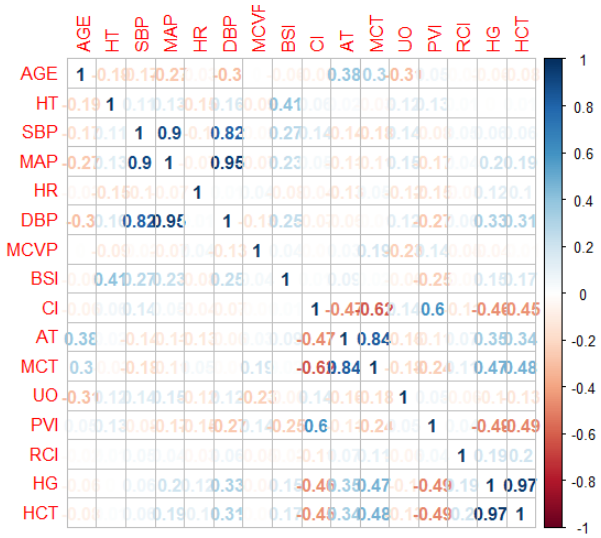


Figure 1: The above plot shows correlation values between all the continuous variables. All variables with correlations greater than .5 were reviewed and the appropriate variable was picked for model building.

Shock type is the only variable that possibly differentiates the medical ailment between each patient. A grouped bar chart in the appendix (Figure 4) illustrates the relationships between the shock type and survival; it's clear there's a significant relationship between the shock type and survival. Notice there are very few deaths among those with no shock, but the survival rate hovers around 50 percent for the other shock types. Two of the shock types have slightly higher rates of mortality than others, but since all of the shock types (besides non-shock) have 10 or less instances per survival outcome, a better relationship can be created that more easily separates survival outcome and shock type. A chi-square test was performed and yielded a small p-value of .00076, confirming association between the shock types and survival. Given the limited data and that a logistic model will be fit, the shock type variable is transformed into a binary variable that describes if the patient was in shock or not. A more granular analysis of each shock type would be beneficial if there was more data, but we want to reduce the dimension. Transforming the variable may lose too information for bigger data sets, but in this instance will help make the model a bit simpler.

Exploratory analysis also showed a difference in measurements and how they relate to the response variable. The above boxplots show a significant difference between measurement 1 (initial measurement) and measurement 2 (final measurement). The pink plots represent the range of values for those who survived, while the blue plots represent the range of values for those who died. The left plot shows a clear negative relationship between MAP and Survival; the relationship is even more clear for the 2nd measurement where the median MAP drops about 10 points so it's below 50. The right plot shows a negative relationship between HG and Survival; we can see the median HG slightly dips for the 2nd measurement to a median HG of about 100. These decreases and clear relationships for the 2nd measurement are a reason why the final variables

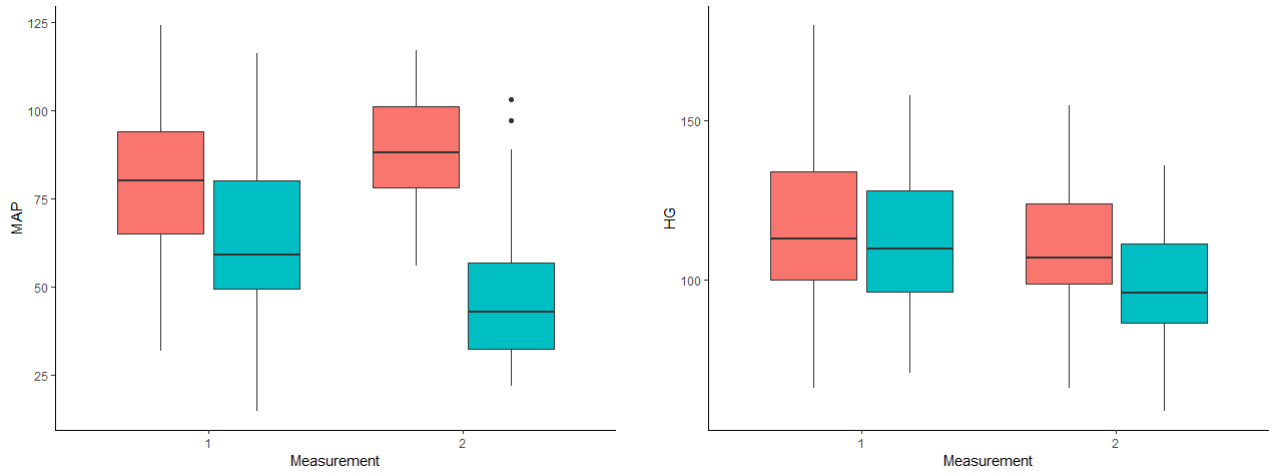


Figure 2: Left: MAP and Survival. Right: HG and Survival.

weren't included in the model; they would be too influential compared to the initial variables. Table 4 in Appendix A displays variables from the final variable subset that had low p-values from the t-test and chi-square tests, indicating there is a relationship between the predictor variables and the response variables listed. The variables listed were 5 additional variables that indicated a relationship with the response. The final variable subset that indicated a relationship also included the same variables that were included in the initial variables subset.

4 Statistical Analysis

Model Selection

Stepwise selection (both directions) was used to determine the best model fit given the variables considered in the model. All variables included in the model were significant at the 90 percent significance level and yielded an AIC of 116.02, which is lower than the AIC of 127 from the model before stepwise selection. The variables included in the model are listed below, along with their associated coefficient values and p-values.

Analysis of variance tests to determine if other models are statistically different were performed by fitting a model similar to the above except with the variable Sex, as that variable from the initial variable subset was statistically significant at the .1 level in exploratory analysis. The anova test yielded a high p-value of .65, indicating this model wasn't statistically different than the model performed after stepwise selection. Another model was fit utilizing the variable MCT instead of CI (MCT was significant at the .1 level) to see if the model was statistically different or had a better AIC. The p-value from this test was also high at .45 and had an AIC of 117. Therefore, the model picked after stepwise selection is still the best model fit.

Diagnostics

Diagnostics for the above selected model were done by transforming the covariates into binomial explanatory variables patterns (EVP) and then modeling the probability of survival by the

Table 1: Best Model Coefficients and Odds Ratios

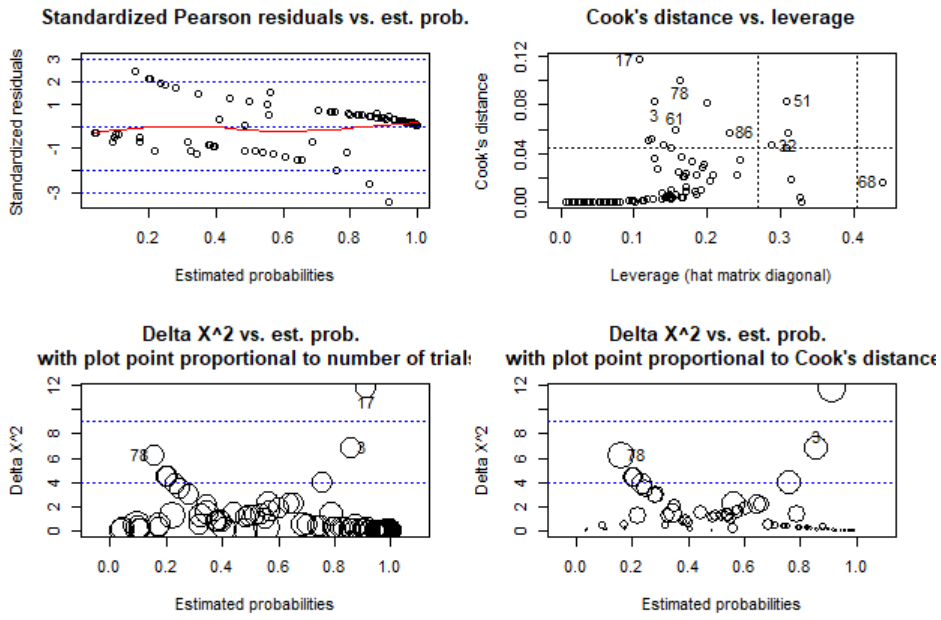
	Coefficient	Odds Ratio	Std. Error	P-value	95% CI
Intercept	-3.5700	0.028	2.6200	0.1722	(0.00013 , 4.06000)
MAP	0.0289	1.030	0.0126	0.0218	(1.01000 , 1.06000)
MCVP	-0.0123	0.988	0.0046	0.0075	(0.97800 , 0.99600)
BSI	0.0261	1.030	0.0143	0.0681	(0.99900 , 1.06000)
UO	0.0072	1.010	0.0042	0.0862	(1.00000 , 1.02000)
IsShock1	-1.9100	0.148	0.7150	0.0076	(0.03000 , 0.53400)

transformed covariates. The continuous covariates are reduced to smaller dimensions by binning; the number of bins selected was chosen to reduce the amount of EVPs while also maintaining an informative enough shape of the distribution of each continuous covariate. The reduced dimensions still allow enough generality of the data to explore any potential outliers. Overall, there were 89 EVPs.

The graphs shown on the next page in Figure 3 display little concern for outliers. The standardized residual plot shows some linear pattern but appears to be random and centered around 0. The one point of concern is EVP 17 which has a high residual of -3.42. The 2nd plot shows the relationship between Cook's distance (measure of influence) and leverage (distance from mean predictor variables), and EVP 17 is shown as having the highest Cook's distance at .12. The leverage is relatively small and hovering above the cloud of points, so not too concerning (same issue for EVP 78, but not considered influential enough). EVP 51 shows a higher leverage and Cook's distance but since the Cook's distance is only .08 it's not too influential. Removal and analysis of these points doesn't significantly change the model coefficients, so no need to remove any points.

Inference

Unsurprisingly, a significant variable in the model is "IsShock1" which is the variable that describes if a patient arrived in shock. The odds ratio of .148 implies that given a patient is in shock upon arrival, they are approximately 85 percent less likely to survive. The confidence interval can be interpreted as being 95 percent confident they are between 46.7 and 97 percent less likely to survive, given they arrived in shock (all other variables held constant). The standard error was a bit higher than the other variables, but the other variables also don't have the same effect. All the variables in the model are significant at at least the 90 (.1) percent level, with Mean Arterial Pressure, Mean Central Venous Pressure, and Shock being significant at the .05 level. For Mean Arterial Pressure, for every one unit increase in either of those units, the patient is 3 percent more likely to survive; there's also 95 percent confidence this probability is between 0 and 6 percent. For every one unit increase in Urine Output, the patient is 1 percent more likely to survive; there's also a 95 percent confidence this probability is between 0 and 2 percent. The inverse is true for Mean Central Venous Pressure, for every one unit increase, the patient is 1 percent less likely to survive; there's also a 95 percent confidence this probability is between 0 and 2.2 percent.



Deviance/df = 1.02; GOF thresholds: 2 SD = 1.32, 3 SD = 1.48

Figure 3: Diagnostic Plots

5 Conclusion

The separation of initial and final observations was deemed necessary for modeling purposes as the assumptions of logistic regression require independent observations. The decision to analyze initial observations stems from final observations being more significantly related to the response variable survival than the initial variables. Exploratory analysis confirmed these suspicions by the use of Welch's t-test, chi-square, and graphical plots; this analysis suggested much more of the measurement variables were significant compared to the initial variables, giving away too much information and perhaps being the result of response variable leakage. These tests also narrowed down the variables that would likely be included in the final model, ultimately choosing a final logistic regression model that included the terms Urinary Output, Mean Arterial Pressure, Mean Central Venous Pressure, Body Surface Index, and Shock. Shock in particular was shown to be a very important variable. No outliers or concerns were apparent while performing the diagnostics.

There were many limitations with this data set due to a lack of consistency and data set attributes. Perhaps better information can be drawn from a data set like this if patients received similar treatments, so instead of vitals and CBCs predicting survival/death by itself, they can be used in conjunction with treatments to determine if a treatment was successful. Another problem with consistency was the lack of consistent time between the initial and final observations; obviously this will differ as patients will die/be released sooner than others, but time series analysis (measurements drawn at regular intervals) may be more interesting as far as information extraction

goes. In addition, there was a definite lack of data with only 112 subjects; this limited amount of data likely barely meets the requirements for logistic regression. This also limited prediction aspirations since more data would be needed to successfully split into training/testing/validation sets and concluding confident performance evaluations (based on initial admittance to the hospital).

Other approaches to this data set included unsupervised learning; clustering was initially implemented to see if there could be a graphical separation between attributes. Another interesting approach (if there was more data) would be to associate each patient to a cluster and see how and if their cluster association changes over time. If a patient was associated with a certain cluster at a particular time series point, intervention could happen to readjust that patient's treatment and prevent them from dying. Repeated measures/conditional logistic regression could also potentially be used for this data structure.

The original approach to this data set was to implement a predictive model on the initial data and unsupervised learning on the final observation data. Association rules were considered for the final observation data, but that would require more domain knowledge to appropriately categorize the continuous variables. Logistic regression on both the initial and final observations was performed in a way to account for multicollinearity (and the same variables weren't chosen for both initial and final), but this ultimately chose a model that was heavily favored towards the final variables and had issues with convergence; penalization parameters would be required and assumptions may have been violated. For a small dataset like this, a simple logistic regression model that fit the initial data extracted a sufficient amount of information.

6 References

Hothorn, T and Everitt, B. *A Handbook of Statistical Analysis Using R.* CRC Press.

Brownlee, J. (2019, May 22). How Much Training Data is Required for Machine Learning? machinelearningmastery.com/much-training-data-required-machine-learning/

Bujang, M. A., Sa'at, N., Sidik, T. M. I. T. A. B., Joo, L. C. (2018, July). Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data. /www.ncbi.nlm.nih.gov/pmc/article

Appendix A: Data Tables and Plots

Table 2: Descriptions of variables in the dataset.

Variable	Description	Value
Age	Age of Patient	Year 0-100
HT	Height	CM
Sex	Male or Female	1 = Male, 2 = Female
Survive	Survival	1 = Survive, 3 = Died,
Shock Typ	Shock Patient Was In	2 = Non-Shock, 3=Hypovolemic shock, 4=Cardiogenic shock, 5=
SBP	Systolic Blood Pressure	mmHg
MAP	Mean Arterial Pressure	mmHg
HR	Heart Rate	beats/min
DBP	Diastolic Blood Pressure	mmHg
MCVP	Mean Central Venous Pressure	cm H2O
BSI	Body Surface Index	m2
CI	Cardiac Index	liters/min m2
AT	Appearance Time	Sec
MCT	Mean Circulation Time	Sec
UO	Urinary Output	ml/hr
PVI	Plasma Volume Index	ml/kg
RCI	Red Cell Index	ml/kg
HG	Hemoglobin	gm/100 ml
HCT	Hematocrit	percent
Record	Card sequence	1=Initial, 2=Final
ID	Patient ID	Numerical

Table 3: Welch's T Test and Chi-Square P-Values for Initial Variables The Chi-Square test was used for factor variables.

	SBP	MAP	DBP	MCVP	BSI	MCT	UO	Shock TYP	Sex
Initial Variables	.000116	9.991e-05	.00034	.004267	.02	.0888	.0004	.0007616	.06376

Table 4: Welch's T Test and Chi-Square P-Values for Final Variables The below variables are variables that now yielded a small p-value where the initial variables had high p-values. The final variables listed in the previous table also have small p-values, so it's redundant to list those as well.

	CI	AT	MCT	HG	HCT
Final Variables	1.516e-07	0.02248	.001412	.0003911	.002484

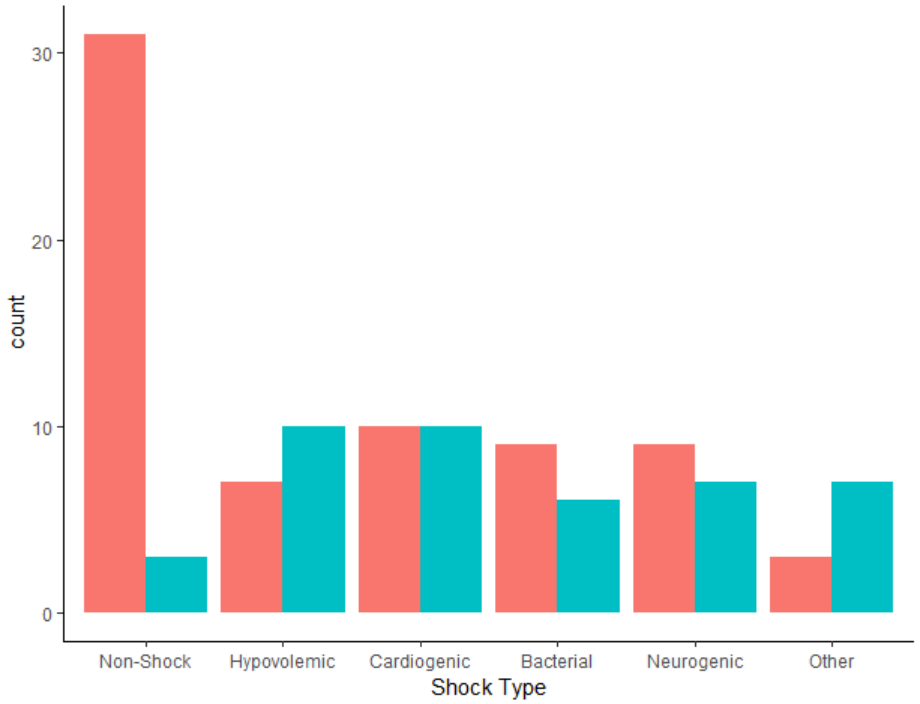


Figure 4: Relationship between the different shock types and the survival outcome. Notice non-shock has a very high survival rate compared to the other shock types. Pink represents survival and blue represents death.

Appendix B: Summary Stats R Code

```
Summary Statistics (Initial Variables)

SBP              MAP               HR               DBP
Min.   : 26.0   Min.   : 15.00   Min.   : 25.00   Min.   : 10.00
1st Qu.: 85.0   1st Qu.: 59.00   1st Qu.: 85.75   1st Qu.: 44.00
Median :104.5   Median : 72.50   Median :101.50   Median : 59.00
Mean   :106.5   Mean   : 73.53   Mean   :104.17   Mean   : 58.61
3rd Qu.:131.0   3rd Qu.: 88.00   3rd Qu.:119.25   3rd Qu.: 71.25
Max.   :171.0   Max.   :124.00   Max.   :217.00   Max.   :108.00
MCVP              BSI               CI               AT               MCT
Min.   :  2.00   Min.   :109.0   Min.   : 17.0   Min.   : 20.00   Min.
: 81.0
1st Qu.: 44.00   1st Qu.:157.2   1st Qu.:140.8   1st Qu.: 64.75   1st Qu.:151.0
Median : 80.00   Median :169.0   Median :226.0   Median : 92.50   Median :194.0
Mean   : 88.32   Mean   :168.2   Mean   :255.1   Mean   :102.46   Mean
:228.4
3rd Qu.:125.00   3rd Qu.:180.2   3rd Qu.:348.2   3rd Qu.:132.25   3rd Qu.:296.2
Max.   :302.00   Max.   :224.0   Max.   :763.0   Max.   :261.00   Max.
:590.0
UO               PVI               RCI               HG
Min.   :  0.00   Min.   : 207.0   Min.   :107.0   Min.   : 66.00
1st Qu.:  0.00   1st Qu.: 377.8   1st Qu.:167.0   1st Qu.: 97.75
Median :  1.00   Median : 461.5   Median :205.0   Median :112.00
Mean   : 54.91   Mean   : 486.0   Mean   :214.7   Mean   :114.79
3rd Qu.: 41.25   3rd Qu.: 563.8   3rd Qu.:241.2   3rd Qu.:133.25
Max.   :510.00   Max.   :1066.0   Max.   :858.0   Max.   :180.00
HCT
Min.   :0.2000
1st Qu.:0.2975
Median :0.3375
Mean   :0.3501
3rd Qu.:0.4113
Max.   :0.5400


Summary Statistics (Final Variables)

     SBP              MAP               HR               DBP
 Min.   : 38.00   Min.   : 22.00   Min.   : 25.00   Min.   : 16.00
 1st Qu.: 78.75   1st Qu.: 48.75   1st Qu.: 78.00   1st Qu.: 39.75
 Median :117.00   Median : 78.00   Median : 94.00   Median : 58.00
 Mean   :110.33   Mean   : 72.82   Mean   : 96.41   Mean   : 55.13
 3rd Qu.:134.00   3rd Qu.: 92.25   3rd Qu.:108.25   3rd Qu.: 72.00
 Max.   :182.00   Max.   :117.00   Max.   :221.00   Max.   :100.00
      MCVP              BSI               CI               AT
```

```
MCT
 Min.   :  1.00   Min.   :109.0   Min.   : 66.0   Min.   : 13.00   Min.
: 71.0
 1st Qu.: 46.00   1st Qu.:160.5   1st Qu.:194.5   1st Qu.: 50.75   1st Qu.:130.
 Median : 82.50   Median :169.0   Median :276.5   Median : 83.00   Median :181.
 Mean   : 85.32   Mean   :168.8   Mean   :292.3   Mean   : 91.24   Mean
:203.0
 3rd Qu.:118.75   3rd Qu.:180.2   3rd Qu.:361.2   3rd Qu.:124.50   3rd Qu.:240.
 Max.   :319.00   Max.   :224.0   Max.   :794.0   Max.   :232.00   Max.
:546.0
       UO               PVI              RCI              HG
HCT
 Min.   :  0.00   Min.   : 333.0   Min.   :105.0   Min.   : 59.0   Min.
:0.1700
 1st Qu.:  1.00   1st Qu.: 443.0   1st Qu.:161.8   1st Qu.: 92.0   1st Qu.:0.26
 Median : 24.00   Median : 521.5   Median :186.5   Median :103.5   Median :0.30
 Mean   : 77.84   Mean   : 540.7   Mean   :206.6   Mean   :105.3   Mean
:0.3048
 3rd Qu.: 88.00   3rd Qu.: 622.0   3rd Qu.:230.2   3rd Qu.:118.2   3rd Qu.:0.34
 Max.   :850.00   Max.   :1066.0   Max.   :858.0   Max.   :155.0   Max.
:0.4500
```

```
setwd("C:/Users/travi/OneDrive/Desktop")
library(tidyverse)
library(data.table)
install.packages("stargazer")
install.packages("corrplot")
library(stargazer)
library(MASS)

#preprocessing
varnames<-c("ID","AGE","HT","Sex","SURVIVE","Shock_TYP","SBP","MAP","HR","DBP",
DAR<-read_csv("DATA-FILEsp2020.csv",col_names = varnames)
DAR$HCT<-DAR$HCT/1000
```

```
DAR$Is_Shock<-ifelse(DAR$Shock_TYP!=2,1,0)
DAR$Is_Shock<-as.factor(DAR$Is_Shock)
DAR$SURVIVE<-as.factor(DAR$SURVIVE)
DAR$Shock_TYP<-as.factor(DAR$Shock_TYP)
DAR$Sex<-as.factor(DAR$Sex)
dar_unique<-DAR[!duplicated(DAR$ID),]
###Summary Statistics
#Demographics
summary(dar_unique[c(2,3)])
#Num Males, Females 58/54
table(dar_unique$Sex)
#Num Survived, Died 69/43
table(dar_unique$SURVIVE)
#shock Types NonShock, Hypovolemic, Cardiogenic, Bacterial, Neurogenic, Other
#34/17/20/15/16/10
dar_unique$Shock_TYP<-as.factor(dar_unique$Shock_TYP)
dar_unique$SURVIVE<-as.factor(dar_unique$SURVIVE)

#####EDA#####
###Shock Analysis
DAR_Shock_Initial<-DAR[which(DAR$Is_Shock==1&DAR$RECORD==1),]
ggplot(DAR_Shock_Initial, aes(x=SURVIVE, y=UO)) +geom_boxplot()
DAR_Shock_Final<-DAR[which(DAR$Is_Shock==1&DAR$RECORD==2),]
ggplot(DAR_Shock_Final, aes(x=SURVIVE, y=UO)) +geom_boxplot()
dar_unique$Shock_TYP<-revalue(dar_unique$Shock_TYP, c("2"="Non-Shock", "3"="Hyp
table(dar_unique$Shock_TYP)
ggplot(dar_unique,aes(fill=SURVIVE,x=Shock_TYP))+geom_bar(position="dodge")+lab
  theme_classic()+theme(legend.position="none")


#Initial Readings
DAR_Initial<-DAR[which(DAR$RECORD==1),]
summary(DAR_Initial[7:20])
colnames(DAR_Initial) <- paste("Initial", colnames(DAR_Initial), sep = "_")
#Final Readings
DAR_Final<-DAR[which(DAR$RECORD==2),]
summary(DAR_Final[7:20])
colnames(DAR_Final) <- paste("Final", colnames(DAR_Final), sep = "_")
DAR_Final<-DAR_Final[-c(1:6)]
#Combined
Wide_Data<-cbind(DAR_Initial,DAR_Final)
Indata<-DAR_Initial[-c(1,3,4,5,6,7,10,14,15,17,20,21,22)]
CorIn<-DAR_Initial[-c(1,4,5,6,21,22)]

#Based on above results, we want to exclude SBP, DBP, AT, MCT, PVI, HCT
AllData<-Wide_Data[-c(1,3,4,5,6,7,10,14,15,17,20,21,22,23,26,28,30,31,34,35,37,
Final_Data<-Wide_Data %>% select(Initial_SURVIVE,Initial_Is_Shock,Initial_MAP,I
```

```
M<-cor(AllData)
M<-cor(CorIn)
#Correlation
library(corrplot)
corrplot(M, method="number")
###T Testing
attach(DAR_Initial)
t.test(SBP[SURVIVE==1],SBP[SURVIVE==3]) #.000116, but highly correlated with MA
t.test(MAP[SURVIVE==1],MAP[SURVIVE==3]) #really small
t.test(HR[SURVIVE==1],HR[SURVIVE==3]) #.3006 too high of a p value, remove
t.test(DBP[SURVIVE==1],DBP[SURVIVE==3]) #.00034, but highly correlated with MAP
t.test(MCVP[SURVIVE==1],MCVP[SURVIVE==3]) #.004267
t.test(BSI[SURVIVE==1],BSI[SURVIVE==3]) #.02
t.test(CI[SURVIVE==1],CI[SURVIVE==3]) #.2308 too high so remove
t.test(AT[SURVIVE==1],AT[SURVIVE==3]) #.3418 too high so remove
t.test(MCT[SURVIVE==1],MCT[SURVIVE==3]) #.0888 too high so remove
t.test(UO[SURVIVE==1],UO[SURVIVE==3]) #.0004
t.test(PVI[SURVIVE==1],PVI[SURVIVE==3]) #.5683 too high so remove
t.test(RCI[SURVIVE==1],RCI[SURVIVE==3]) #.2342 too high so remove
t.test(HG[SURVIVE==1],HG[SURVIVE==3]) #.5729 too high so remove
t.test(HCT[SURVIVE==1],HCT[SURVIVE==3]) #.6216 too high so remove
t.test(AGE[SURVIVE==1],AGE[SURVIVE==3]) #.1882 too high so remove
t.test(HT[SURVIVE==1],HT[SURVIVE==3]) #.06965 remove, BSI is more relevant
chisq.test(x=SURVIVE,y=Shock_TYP) #.0007616
chisq.test(x=SURVIVE,y=Is_Shock) #really small

chisq.test(x=SURVIVE,y=Sex) #.06376 remove
detach(DAR_Initial)

attach(DAR_Final)
t.test(SBP[Initial_SURVIVE==1],SBP[Initial_SURVIVE==3]) #really small, but high
t.test(MAP[SURVIVE==1],MAP[SURVIVE==3]) #really small
t.test(HR[SURVIVE==1],HR[SURVIVE==3]) #.0233
t.test(DBP[SURVIVE==1],DBP[SURVIVE==3]) #really small, but highly correlated wi
t.test(MCVP[SURVIVE==1],MCVP[SURVIVE==3]) #.000158
t.test(BSI[SURVIVE==1],BSI[SURVIVE==3]) #.098 Duplicate
t.test(CI[SURVIVE==1],CI[SURVIVE==3]) #really small
t.test(AT[SURVIVE==1],AT[SURVIVE==3]) #.0224
t.test(MCT[SURVIVE==1],MCT[SURVIVE==3]) #.001412
t.test(UO[SURVIVE==1],UO[SURVIVE==3]) #really small
t.test(PVI[SURVIVE==1],PVI[SURVIVE==3]) #.1887 Remove
t.test(RCI[SURVIVE==1],RCI[SURVIVE==3]) #.3863 Remove
t.test(HG[SURVIVE==1],HG[SURVIVE==3]) #.0003911
t.test(HCT[SURVIVE==1],HCT[SURVIVE==3]) #.002484
t.test(AGE[SURVIVE==1],AGE[SURVIVE==3]) #.1882 too high so remove
t.test(HT[SURVIVE==1],HT[SURVIVE==3]) #.002579 Remove, correlated
```

```
chisq.test(x=SURVIVE,y=Shock_TYP) #.0007616
chisq.test(x=SURVIVE,y=Sex) #.06376 remove
detach(DAR_Final)
###Conditional plots to determine relationships between measurements
Died<-DAR[which(DAR$SURVIVE==3),]
DAR$RECORD[which(DAR$RECORD==1)]<-""
ggplot(DAR, aes(x=RECORD, y=HG, fill = SURVIVE)) +geom_boxplot(aes(fill=SURVIVE
  theme_classic()+theme(legend.position="none")
ggplot(DAR, aes(x=RECORD, y=MAP, fill = SURVIVE)) +geom_boxplot(aes(fill=SURVIV
  theme_classic()+theme(legend.position="none")
#Histograms and boxplots
ggplot(Wide_Data, aes(x=Initial_SURVIVE, y=Final_CI)) +geom_boxplot()
ggplot(Wide_Data, aes(x=Initial_SURVIVE, y=Final_AT)) +geom_boxplot()
ggplot(Wide_Data, aes(x=Initial_SURVIVE, y=Final_MCT)) +geom_boxplot()
ggplot(Wide_Data, aes(x=Initial_SURVIVE, y=Final_HG)) +geom_boxplot()
ggplot(Wide_Data, aes(x=Initial_SURVIVE, y=Final_HCT)) +geom_boxplot()
hist(Data_Initial$Initial_UO)
hist(Data_Initial$Initial_MAP,breaks=5)
hist(Data_Initial$Initial_MCVP)
hist(Data_Initial$Initial_BSI)
##Model Fitting
Data_Initial<-DAR_Initial[-c(1,6,7,10,14,13,17,20,21)]
Data_Final<-Wide_Data[-c(1,3,6,7,10,13,14,15,17,20,21,23,26,27,28,30,31,32,34,3
library(MASS)
install.packages("car")
library(car)
Final_Data<-Wide_Data %>% select(Initial_SURVIVE,Initial_Is_Shock,Initial_MAP,I
fitallvars<-glm(Initial_SURVIVE~.,family=binomial(link=logit),data=Final_Data)
vif(fitallvars)
#fitallvars<-glm(Initial_SURVIVE~.,family=binomial(link=logit),data=DAR_Initial
#Final_Data<-Wide_Data %>% select(Initial_SURVIVE,Initial_Is_Shock,Initial_MAP,

#Using Initial Data Only
Model_Initial<-glm(Initial_SURVIVE~.,family=binomial(link=logit),data=Data_Init
summary(Model_Initial)
vif(Model_Initial)
Initial_fit=stepAIC(Model_Initial,direction="both")
summary(Initial_fit)
Best_Initial<-glm(Data_Initial$Initial_SURVIVE~Initial_MAP+Initial_MCVP+Initial
Plus_Sex<-glm(Data_Initial$Initial_SURVIVE~Initial_MAP+Initial_MCVP+Initial_BSI
Plus_MCT<-glm(Data_Initial$Initial_SURVIVE~Initial_MAP+Initial_MCVP+Initial_BSI
summary(Best_Initial)
summary(Plus_Sex)
summary(Plus_MCT)
anova(Best_Initial,Plus_Sex,test="Chisq")
anova(Best_Initial,Plus_MCT,test="Chisq")
```

```
vif(Best_Initial)
#using Final Data Also
Model_All<-glm(Initial_SURVIVE~.,family=binomial(link=logit),data=Data_Final)
All_Fit<-stepAIC(Model_All,direction="both")
vif(All_Fit)
summary(All_Fit)
All_Final<-glm(Data_Final$Initial_SURVIVE~Initial_AGE+Initial_MCVP+Initial_Is_S
summary(All_Final)
###Diagnostics

one.fourth.root=function(x){
  x^0.25
}
source("examine.logistic.reg.R")
attach(Data_Initial)

#bin continuous covariates to create EVPs
hist(Initial_UO,breaks=2)
breaks <- c(unique(quantile(Initial_UO)))
UO_interval = cut(Initial_UO, breaks, include.lowest = TRUE)
hist(Initial_MAP)
hist(Initial_MAP, breaks=6)
g=4
MAP_interval = cut(Initial_MAP, quantile(Initial_MAP, 0:g/g), include.lowest =
hist(Initial_MCVP)
hist(Initial_MCVP,breaks=3)
g=3
MCVP_interval = cut(Initial_MCVP, quantile(Initial_MCVP, 0:g/g), include.lowest
hist(Initial_BSI)
g=4
BSI_interval = cut(Initial_BSI, quantile(Initial_BSI, 0:g/g), include.lowest =
Data_Initial$Initial_SURVIVE<-as.numeric(Data_Initial$Initial_SURVIVE)
Data_Initial$Initial_SURVIVE[Data_Initial$Initial_SURVIVE==2]<-0

# Diagnostic plots
w <- aggregate(formula = Initial_SURVIVE ~ BSI_interval+MCVP_interval+MAP_inter
n <- aggregate(formula = Initial_SURVIVE ~ BSI_interval+MCVP_interval+MAP_inter

w.n <- data.frame(w, trials = n$Initial_SURVIVE, prop = round(w$Initial_SURVIVE
mod.prelim1 <- glm(formula = Initial_SURVIVE/trials ~ BSI_interval+MCVP_interva
                   family = binomial(link = logit), data = w.n, weights = trial
save1=examine.logistic.reg(mod.prelim1, identify.points=T, scale.n=one.fourth.r

w.n.diag1=data.frame(w.n, pi.hat=round(save1$pi.hat, 2), std.res=round(save1$st
                     cookd=round(save1$cookd, 2), h=round(save1$h, 2))
p=length(mod.prelim1$coef) # number of parameters in model (# coefficients)
ck.out=abs(w.n.diag1$std.res)>2 | w.n.diag1$cookd>4/nrow(w.n) | w.n.diag1$h > 3
```

```
extract.EVPs=w.n.diag1[ck.out, ]
extract.EVPs
#Seeing if removing outliers changes anything. Rerun model after this
#80 BSI Interval >150 <=164 and MCVP_Interval >80 <=111, MAP >50.9<=60 UO Inter
#96 <=109 >=150 >111 <=142 >=15 <=50.9 >41.2 <=510 Is Shock
Data_Initial<-DAR_Initial[-c(1,3,6,7,10,14,15,17,20,21)]

Data_Initial<-Data_Initial[-which(Initial_BSI>150 & Initial_BSI<=164 & Initial_
Data_Initial<-Data_Initial[-which(Initial_BSI>109 & Initial_BSI<=150 & Initial_

####Final Model
betahat = format(signif(Best_Initial$coeff , digits =3),digits=2, format ="f",
OR = format(signif(exp(Best_Initial$coeff), digits =3),digits=2, format ="f", f
SE = format(signif(summary(Best_Initial)$coeff[,2],digits =3),  digits=2, forma
cibounds = format(signif(exp(confint(Best_Initial)), digits =3),digits=2, forma
pval = format(signif(summary(Best_Initial)$coeff[,4], digits =4),digits=2, form

matrix=cbind(betahat , OR, SE, pval ,matrix(paste ("(",  cibounds [,1], ",", ci
colnames(x) = cbind (" Coefficient", "Odds  Ratio","Std. Error", "P-value", "95
```