Predicting Patient Survival Based on Physiological Variables using Logistic Regression

Joe Sanchez

January 20, 2020

Executive Summary

One of the many problems that hospital staff face is deciding which patients to take care of first. While ethicists and medical professionals don't always agree on who should get treatment first, the common procedure is to treat patients who arrive at a hospital in critical condition first. The goal of this study is to create a logistic regression model which will allow medical staff to quantify the severity of a patient's condition. Patients that are at a higher risk of dying would be given treatment first. The dataset used contains the physiological parameters of 112 critically ill patients which were collected in Southern California. Stepwise model selection was used in order to obtain the final logistic regression model. The model was then trained and tested on a 70/30 split of the data in order to gauge performance. The sensitivity achieved by the model was 0.8 and the specifity obtained was 0.68. The results of the analysis suggest that shock type, mean arterial pressure, mean central venous pressure, and body surface index are significant factors in determining a patient's chance of surival.

1 Introduction

It is a widely known that hospitals in the US face issues due to understaffing of registered nurses. When nurses cannot give patients the care that they require there are serious consequences. Many physicians have reported patient deaths in which they point to severe understaffing as the cause. It is critical for nurses to understand what level of care each patient needs in order to avoid patient deaths.

The purpose of this study is to create a logistic regression model which can accurately predict which patients are at a higher risk of dying. The dataset contains 21 fields and 224 observations, which includes 2 observations for each of the 112 critically ill patients (an initial measurement upon admission and a measurement just before death or discharge). Table 1 provides a brief summary of the variables. An accurate logistic regression model would allow medical staff to quickly identify higher risk patients which need more care than other patients.

Table 1: Descriptions of variables in the dataset.

Variable	Description	Units	Values
ID	ID	None	
AGE	Patient Age	yr	16-90
HT	Patient Height	cm	140-185
Sex		None	1=Male, 2=Female
SURVIVE	Survival	None	1=Survived, 3=Died
SHOCK_TYPE	Shock Type	None	2=Non-shock 3=Hypovolemic shock 4=Cardiogenic shock 5=Bacterial shock 6=Neurogenic shock 7=Other
SBP	Systolic Pressure	mmHg	26-171
MAP	Mean Arterial Pressure	mmHg	15-124
HR	Heart Rate	beats/min	25-217
DBP	Diastolic Pressure	mmHg	10-108
MCVP	Mean Central Venous Pressure	cm H_2O	2-302
BSI	Body Surface ndex	m^2	109-224
CI	Cardiac Index	liters/min*m^2	17-763
AT	Appearance Time	sec	20-261
MCT	Mean Circulation Time	sec	81-590
UO	Urinary Output	ml/hr	0-510
PVI	Plasma Volume Index	ml/kg	207-1066
RCI	Red Cell Index	ml/kg	107-858
HG	Hemoglobin	gm/100 ml	66-180
HCT	Hematocrit	percent	20.0-54.0
RECORD	Card Sequence	None	1=Initial,2=Final

2 Methods

The data set contained two observations for each of the 112 patients. There were no missing values in the dataset, there was however, a discrepancy between the initial and final height measurements for one of the patients. It was assumed that this was a mistake and the initial height of 70 cm was changed to 170 cm to match the final height. Out of the 112 patients 43 (38.4%) patients died.

The goal was to create an accurate model which would allow hospital staff to quickly identify which incoming patients were in more severe condition (predicted to die by model). Therefore, only the initial records were used since, the final measurements may have either been taken shortly before a death or just before discharging a patient in which case they were irrelevant The ID variable was useless for predicting survival therefore it was dropped along with the record variable once all the final measurements were removed. The Survival variable values were changed to 0 = survival and 1 = death. Since the response variable was binary, a logistic regression model was used in order to predict whether a patient in the dataset lived or died.

Exploratory data analysis was performed on the physiological variables and associations between survival and categorical tests were determined using χ^2 tests. Welch's T-test was used to determine differences in the means of continuous variables between the living and dead patients. The variables used in the model were selected by using stepwise model selection, χ^2 tests and T-tests. If the stepAIC function and tests for association both deemed a variable significant the term was kept if not the model was reduced and the process repeated until the final model was obtained. A highly reduced model was sought in order to avoid overfitting. The dataset was then split into training and testing sets using a ratio of 70/30 in order to evaluate the model's sensitivity and specificity. R was used for the analysis in this study.

3 Results

Exploratory Data Analysis

The data set contains both categorical and continuous predictor variables. Table 2a features the counts for survival and deaths among patients. Contingency tables were created for each of the predictor variables and the response and χ^2-tests were used to test association. Table 2b is a contingency table of survival and Sex. A χ^2 -test determined that Sex is not associated with survival, however the P-value was 0.0638, meaning that association between Sex and survival was barely rejected by the test at the 0.05 alpha level. It is possible that the small sample size could have impacted the results. A χ^2-test was also used to determine that there was a significant association shock type and survival.

Table 2: Contingency tables for survive and relation between survive and sex.

(a) Contingency table for survive.

Survived	Died
69	43

(b) Contingency table for survive and sex . A χ^2-test resulted in no association with survive at the 0.05 significance level.

	Female	Male
Survived	28	41
Died	26	17

Figure 1 depicts boxplots of mean arterial pressure (MAP) and mean central venous pressure (MCVP) by survive. Welch's two-sample t-tests were used to determine if there were differences in the means of continuous variables between the patients that survived and those that died. The t-tests determined that there were significant differences in MAP and MCVP between the two groups of patients.

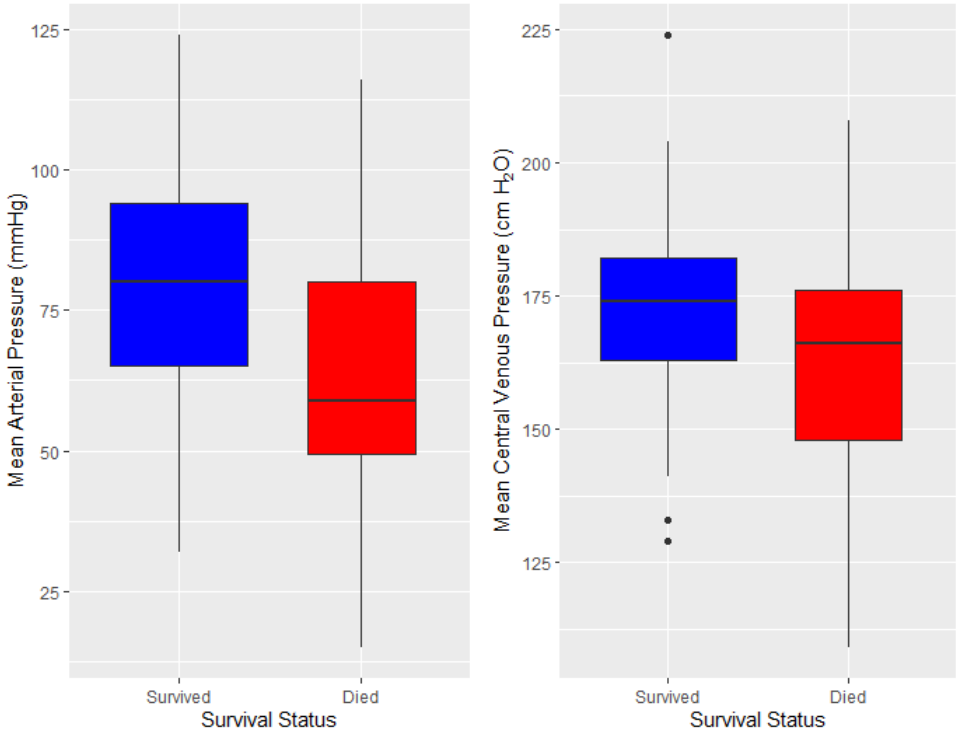

Figure 1: Boxplots for mean arterial pressure (left) and mean central venous pressure (right)

Cohen's d was also calculated for each of the continuous variables. MAP had a value cohen's d estimate of 4.7 indicating roughly a medium level of effect on survival. While MCVP had a cohen's d estimate of 2.2 indicating a small level of effect on survival. Figure 2 shows a correlation plot of the variables. MAP, SBP, and DBP are highly correlated with each other.AT and MCT are also highly correlated with each other and HG and HCT form another duo of highly correlated variables.

Model Selection

In order to obtain the final model, stepwise AIC was used in conjunction with the results from t-tests and χ^2-tests of association. Each variable was used in the initial model, however variables were removed if stepAIC and other tests did not agree that a term was significant. This was done iteratively until the final model was obtained. The results were that shock type, MAP, MCVP and BSI were deemed significant. The same procedure was repeated on model involving an interaction term between sex and shock type (a contingency table showed a possible association with survival) however, the interaction term was deemed insignificant.

Although the first iteration of stepwise AIC initially recommended a model which also included UO and RCI, these were dropped because the results of t-tests in the model summary deemed them

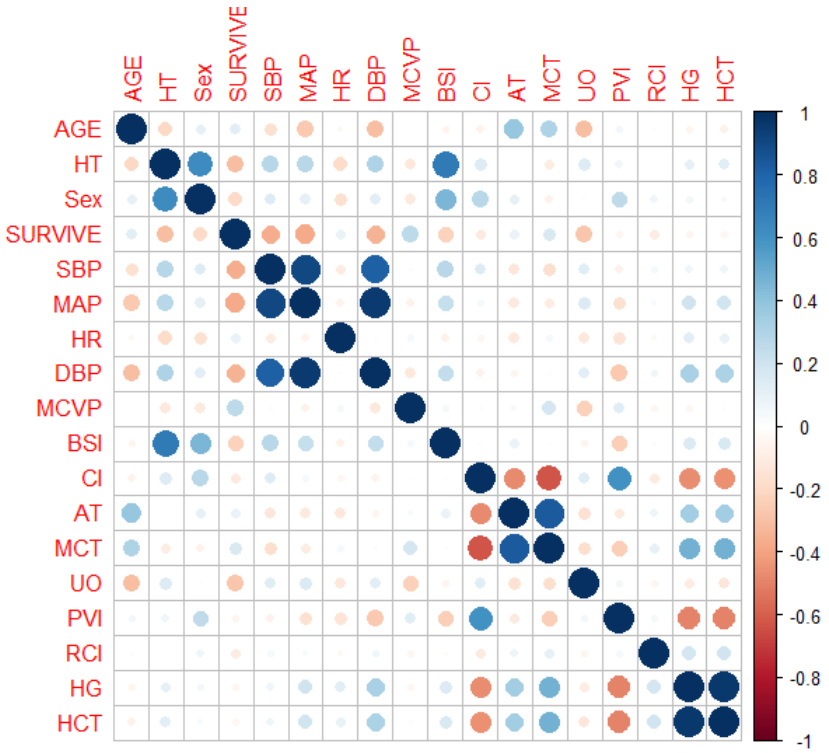

Figure 2: Correlation plot of the variables in the dataset

insignificant at the 0.05 significant level. The AIC score of the model including UO and RCI was 119.1 and the AIC score of the final model was 120.8. The increase in AIC was marginal while the final model benefited from further simplification in order to reduce overfitting.

Diagnostics

Transformation of the covariates into explanatory variable patterns (EVPs) was necessary in order to perform model diagnostics. Due to the small size of the dataset, the continuous variables used in the final model were split into 2 intervals based on percentiles. The data was aggregated into sets of similar covariates and the result of the transformation was reduction of the 112 observations into 39 EVPs.

The studentized residual plot of the EVPS shows two points which are potential outliers since they are just below 3 standard deviations away from 0. The cook's distance vs leverage plot of the EVPs shows four potential outliers. The same two points have a somewhat higher cook's distance along with another point while an additional point has a high leverage compared to all of the other points. These values are displayed in Table 3 and Diagnostic plots are shown in Appendix C. None of the possible outliers were dropped since none of the points had a cook's distance above 4/n (where n is the number of EVPs)while also simultaneously having a high leverage and being over three standard deviations from 0. The Generalized Variance Inflation Factor (VIF) was also calculated for each of the variables included in the model with all results being low, indicating that there was little multicollinearity in the model.

Training and Testing

Table 3: The EVPs which are potential outliers are shown

	SHOCK_TYP	MAP_interval	MCVP_interval	SURVIVE	BSI_interval	std.res	cookd	h
15	5	(15,72.5]	(80,302]	4.00	(109,169]	-0.15	0.00	0.63
16	6	(15,72.5]	(80,302]	1.00	(109,169]	-2.90	0.70	0.43
32	2	(15,72.5]	(80,302]	2.00	(169,224]	2.44	0.28	0.30
38	6	(72.5,124]	(80,302]	2.00	(169,224]	1.59	0.14	0.34

Table 4: Generalized VIF scores and Degrees of freedom

	GVIF	Df	GVIF^(1/(2*Df))
SHOCK_TYP	1.27	5.00	1.02
MAP	1.12	1.00	1.06
MCVP	1.24	1.00	1.11
BSI	1.08	1.00	1.04

The dataset was split into training and testing sets using a 70/30 ratio. The final model was then fit to the training data and predictions were made using the testing data. If the probability predicted was greater than 0.40 a patient was classified as dead if not the patient was classified as a survivor. The cutoff of 0.40 was used since there was an imbalance in the number of patients which lived and died in the dataset with approximately 40% of patients dying.

The results of the training and testing were that a sensitivity of 80% was achieved and a specificity of 68% was also achieved. The results are not quite ideal but they can still be somewhat useful to medical staff. Since the goal is to create a model which can predict which patients are at a higher risk of death, it is better to have higher sensitivity than specificity if one must be lower than the other. As the saying goes "it's better to be safe than sorry" and in this case it is better to give a lower risk patient more care than to give a high risk patient less care in order to avoid patient deaths.

Inferences

In order to make sense of the final model, odds ratios must be used in order to describe relationships between the covariates and survival. Changes in odds ratios specify the effects of an increase or decrease in a variable while all other variables are held constant. According to the model, there is a 1450% increase in the odds of death if a patient has hypovolemic shock. If a patient has cardiogenic shock there is a 827% increase in odds of death. Bacterial shock increases odds of death by 279% and neurogenic shock results in a 598% increase in odds of death. Patients exhibiting "other" shock type have a 2940% increase in odds of death. A one unit increase (mmHg) in MAP reduces odds of death by 2.9%. A one unit increase (mmHg) in MCVP increases odds of death by 2% a one unit increase (m^2) in BSI reduces odds of death by 2.8%. A summary of the model estimates is provided in table 5.

4 Conclusion

The goal of developing a model which could identify patients likely to die was achieved with shock type, MAP, MCVP, and BSI being identified significant predictor variables. The sensitivity

Table 5: Summary of model estimates including odds ratios, standard errors, p-values, and confidence intervals.

	Coefficient	Odds ratio	SE	p-value	95% CI on OR
Intercept	3.10	23.000	2.500	0.21	(0.195 , 3950.000)
Hypovolemic Shock	2.70	15.500	0.935	0.00	(2.750 , 115.000)
Cardiogenic Shock	2.20	9.270	0.835	0.01	(1.980 , 56.100)
Bacterial Shock	1.30	3.790	0.897	0.14	(0.683 , 24.900)
Neurogenic Shock	1.90	6.980	0.871	0.03	(1.360 , 44.400)
Other	3.40	30.400	1.070	0.00	(4.270 , 308.000)
MAP	-0.03	0.971	0.013	0.02	(0.945 , 0.995)
MCVP	0.01	1.020	0.005	0.00	(1.010 , 1.030)
BSI	-0.03	0.972	0.014	0.05	(0.944 , 0.999)

achieved by the final model was 80% and the specificity achieved was 68.4%. While these values could use some improvement, the model could still be used to predict which patients are at a high risk of dying and would allow medical staff to allocate resources and care in order to minimize deaths.

One limitation of this study is that a small sample size was used. This could have affected which terms were deemed significant and could also have affected the coefficients in the model. Additionally, the results from training and testing may be somewhat unreliable because the training set included only a small amount of observations. Another limitation of this study is that nothing is known about a patient's history. Such information could prove useful in assessing whether or not a patient is at a higher risk of dying. Another limitation is that the data was collected in Southern California. It is possible that the model would not be accurate for people in different regions.

In order to produce a more accurate model future studies could include a larger sample size, include a patients history, and use data that was obtained from all over the nation or possibly world. Perhaps researchers could also include variables which describe conditions at the hospital and describe medical staff in order to get a better picture of whether or not a patient will survive.

Appendices

Appendix A: R Code

```
library(dplyr)
library(vioplot)
library(ggplot2)
library(corrplot)
library(MASS)
library(gridExtra)
library(xtable)
library(car)
library(effsize)

column.names<-c('ID','AGE','HT','Sex','SURVIVE','SHOCK_TYP',
'SBP','MAP','HR','DBP','MCVP','BSI',
'CI','AT','MCT','UO','PVI','RCI','HG',
'HCT','RECORD')

physio.data<-read.csv('DATA-FILEsp2020.csv',header=FALSE,
col.names=column.names)
physio.data$SURVIVE<-ifelse(physio.data$SURVIVE==1,0,1) #changing
   survive to 0 and death to 1
physio.data$Sex<-ifelse(physio.data$Sex == 2,0,1) #changed female
   to 0 kept male as 1
physio.data$HCT<-physio.data$HCT/10
#3rd digit in HCT was assumed to specify tenths since it is
#not possible to have an hct above 100
physio.data[physio.data$ID==539 & physio.data$RECORD == 1,'HT
   ']<-170 #error in height measurement fixed
physio.data[physio.data$ID==539,]


#removing ID column
physio.data$ID<-NULL
#getting rid of second observations for each patient
physio.data<-filter(physio.data,RECORD==1)
physio.data2<-physio.data
#removing record column
physio.data$RECORD<-NULL
physio.data2$RECORD<-NULL

#changing categorical columns to factors
physio.data$SHOCK_TYP<-as.factor(physio.data$SHOCK_TYP)
physio.data2$SURVIVE<-as.factor(physio.data2$SURVIVE)
```

```
physio.data2$Sex<-as.factor(physio.data2$Sex)
physio.data2$SHOCK_TYP<- as.factor(physio.data2$SHOCK_TYP)

head(physio.data2)

#112 observations and 19 variables after removing final
   measurements
dim(physio.data2)
sum(is.na(physio.data2)) #no missing values
summary(physio.data2)
attach(physio.data2)



###########variable distributions

table(SURVIVE)
signif(prop.table(table(SURVIVE)),digits=3) #38.4% of patients did
   not survive


#AGE distribution
plot(1, 1, xlim = c(0, 2), ylim = range(AGE), type = 'n', xlab =
   '', ylab = '', xaxt = 'n')
vioplot(AGE, at=1, add=T)
axis(2, at = 50, pos=-0.2, tck=0, labels="Age")

#height distribution
plot(1, 1, xlim = c(0, 2), ylim = range(HT), type = 'n', xlab = '',
    ylab = '', xaxt = 'n')
vioplot(HT, at=1, add=T)
axis(2, at = 165, pos=-0.2, tck=0, labels="Height (cm)")

table(Sex) # 58 Males and 54 Females

table(SHOCK_TYP)
# 2 Non-shock patients
# 3 Hypovolemic
# 20 Cardiogenic
# 5 Bacterial
# 16 Neurogenic
# 7 Other

#SBP distribution
plot(1, 1, xlim = c(0, 2), ylim = range(SBP), type = 'n', xlab =
   '', ylab = '', xaxt = 'n')
vioplot(SBP, at=1, add=T)
```

```
axis(2, at = 100, pos=-0.2, tck=0, labels="Systolic Blood Pressure (mmHg)")

#MAP distribution
plot(1, 1, xlim = c(0, 2), ylim = range(MAP), type = 'n', xlab = '', ylab = '', xaxt = 'n')
vioplot(MAP, at=1, add=T)
axis(2, at = 70, pos=-0.2, tck=0, labels="Mean Arterial Pressure (mmHg)")

#HR distribution
plot(1, 1, xlim = c(0, 2), ylim = range(HR), type = 'n', xlab = '', ylab = '', xaxt = 'n')
vioplot(HR, at=1, add=T)
axis(2, at = 120, pos=-0.2, tck=0, labels="Heart Rate (bpm)")

#DBP distribution
plot(1, 1, xlim = c(0, 2), ylim = range(DBP), type = 'n', xlab = '', ylab = '', xaxt = 'n')
vioplot(DBP, at=1, add=T)
axis(2, at = 60, pos=-0.2, tck=0, labels="Diastolic Blood Pressure (mmHg)")

#MCVP distribution
plot(1, 1, xlim = c(0, 2), ylim = range(MCVP), type = 'n', xlab = '', ylab = '', xaxt = 'n')
vioplot(MCVP, at=1, add=T)
axis(2, at = 150, pos=-0.2, tck=0, labels="Mean Central Venous (cm H2O pressure)")

#BSI distribution
plot(1, 1, xlim = c(0, 2), ylim = range(BSI), type = 'n', xlab = '', ylab = '', xaxt = 'n')
vioplot(BSI, at=1, add=T)
axis(2, at = 170, pos=-0.2, tck=0, labels="Body Surface Index (m^2)")

#CI distribution
plot(1, 1, xlim = c(0, 2), ylim = range(CI), type = 'n', xlab = '', ylab = '', xaxt = 'n')
vioplot(CI, at=1, add=T)
axis(2, at = 380, pos=-0.2, tck=0, labels="Cardiac index (liters/min m^2)")

#AT distribution
plot(1, 1, xlim = c(0, 2), ylim = range(CI), type = 'n', xlab = '', ylab = '', xaxt = 'n')
```

```
115 vioplot(CI, at=1, add=T)
116 axis(2, at = 380, pos=-0.2, tck=0, labels="Cardiac index (liters/
       min m^2)")
117
118 #MCT distribution
119 plot(1, 1, xlim = c(0, 2), ylim = range(MCT), type = 'n', xlab =
       '', ylab = '', xaxt = 'n')
120 vioplot(MCT, at=1, add=T)
121 axis(2, at = 350, pos=-0.2, tck=0, labels="Mean circulation time (
       sec)")
122
123 #UO distribution
124 plot(1, 1, xlim = c(0, 2), ylim = range(UO), type = 'n', xlab = '',
        ylab = '', xaxt = 'n')
125 vioplot(UO, at=1, add=T)
126 axis(2, at = 250, pos=-0.2, tck=0, labels="Urinary Output (ml/hr)")
127
128 #PVI distribution
129 plot(1, 1, xlim = c(0, 2), ylim = range(PVI), type = 'n', xlab =
       '', ylab = '', xaxt = 'n')
130 vioplot(PVI, at=1, add=T)
131 axis(2, at = 600, pos=-0.2, tck=0, labels="Plasma volume index (ml/
       kg)")
132
133 #RCI distribution
134 plot(1, 1, xlim = c(0, 2), ylim = range(RCI), type = 'n', xlab =
       '', ylab = '', xaxt = 'n')
135 vioplot(RCI, at=1, add=T)
136 axis(2, at = 500, pos=-0.2, tck=0, labels="Red Cell Index ml/kg")
137
138 #HG distribution
139 plot(1, 1, xlim = c(0, 2), ylim = range(RCI), type = 'n', xlab =
       '', ylab = '', xaxt = 'n')
140 vioplot(RCI, at=1, add=T)
141 axis(2, at = 500, pos=-0.2, tck=0, labels="Hemoglobin (gm/100ml)")
142
143 #HCT distribution
144 plot(1, 1, xlim = c(0, 2), ylim = range(HCT), type = 'n', xlab =
       '', ylab = '', xaxt = 'n')
145 vioplot(HCT, at=1, add=T)
146 axis(2, at = 375, pos=-0.2, tck=0, labels="Hematocrit (percent)")
147
148
149
150
151 ##########comparing covariates to survival status
152 #continuous covariates are compared using boxplots
```

```
#contingency tables are used for discrete covariates

ggplot(data=physio.data2,aes(x=SURVIVE,y=AGE)) +
geom_boxplot(notch=FALSE,varwidth=T) +
scale_x_discrete('Survival Status',breaks=c(0,1),
labels=c("Survived","Died"))


ggplot(data=physio.data2,aes(x=SURVIVE,y=HT)) +
geom_boxplot(notch=FALSE,varwidth=T) +
scale_x_discrete('Survival Status',breaks=c(0,1),
labels=c("Survived","Died"))

table(SURVIVE,Sex)


table(SURVIVE,SHOCK_TYP)

p3<-ggplot(data=physio.data2,aes(x=SURVIVE,y=SBP)) +
geom_boxplot(notch=FALSE,varwidth=T,fill=c("blue","red")) +
scale_x_discrete('Survival Status',breaks=c(0,1),
labels=c("Survived","Died")) +
labs(y="Systolic Blood Pressure (mmHg)")

p1<-ggplot(data=physio.data2,aes(x=SURVIVE,y=MAP)) +
geom_boxplot(notch=FALSE,varwidth=T,fill=c('blue','red')) +
scale_x_discrete('Survival Status',breaks=c(0,1),
labels=c("Survived","Died")) +
labs(y="Mean Arterial Pressure (mmHg)")

ggplot(data=physio.data2,aes(x=SURVIVE,y=HR)) +
geom_boxplot(notch=FALSE,varwidth=T) +
scale_x_discrete('Survival Status',breaks=c(0,1),
labels=c("Survived","Died"))


p4<-ggplot(data=physio.data2,aes(x=SURVIVE,y=DBP)) +
geom_boxplot(notch=FALSE,varwidth=T,fill=c("blue","red")) +
scale_x_discrete('Survival Status',breaks=c(0,1),
labels=c("Survived","Died")) +
labs(y="Diastolic Blood Pressure (mmHg)")

ggplot(data=physio.data2,aes(x=SURVIVE,y=MCVP)) +
geom_boxplot(notch=FALSE,varwidth=T,fill=c('blue','red')) +
scale_x_discrete('Survival Status',breaks=c(0,1),
labels=c("Survived","Died")) +
labs(y="Mean Arterial Pressure (mmHg)")
```

```
200
201 p2 <- ggplot(data=physio.data2,aes(x=SURVIVE,y=BSI)) +
202 geom_boxplot(notch=FALSE,varwidth=T,fill=c('blue','red')) +
203 scale_x_discrete('Survival Status',breaks=c(0,1),
204 labels=c("Survived","Died")) +
205 labs(y=expression(paste("Mean Central Venous Pressure (cm  ",H[2],"
    O)")))
206
207 ggplot(data=physio.data2,aes(x=SURVIVE,y=CI)) +
208 geom_boxplot(notch=FALSE,varwidth=T) +
209 scale_x_discrete('Survival Status',breaks=c(0,1),
210 labels=c("Survived","Died"))
211
212 ggplot(data=physio.data2,aes(x=SURVIVE,y=AT)) +
213 geom_boxplot(notch=FALSE,varwidth=T) +
214 scale_x_discrete('Survival Status',breaks=c(0,1),
215 labels=c("Survived","Died"))
216
217 ggplot(data=physio.data2,aes(x=SURVIVE,y=MCT)) +
218 geom_boxplot(notch=FALSE,varwidth=T) +
219 scale_x_discrete('Survival Status',breaks=c(0,1),
220 labels=c("Survived","Died"))
221
222 p5<-ggplot(data=physio.data2,aes(x=SURVIVE,y=UO)) +
223 geom_boxplot(notch=FALSE,varwidth=T,fill=c("blue","red")) +
224 scale_x_discrete('Survival Status',breaks=c(0,1),
225 labels=c("Survived","Died")) +
226 labs(y="Urinary Output (ml/hr)")
227
228 ggplot(data=physio.data2,aes(x=SURVIVE,y=PVI)) +
229 geom_boxplot(notch=FALSE,varwidth=T) +
230 scale_x_discrete('Survival Status',breaks=c(0,1),
231 labels=c("Survived","Died"))
232
233 p6<-ggplot(data=physio.data2,aes(x=SURVIVE,y=RCI)) +
234 geom_boxplot(notch=FALSE,varwidth=T,fill=c("blue","red")) +
235 scale_x_discrete('Survival Status',breaks=c(0,1),
236 labels=c("Survived","Died")) +
237 labs(y="Red Cell Index  (ml/kg)")
238
239 ggplot(data=physio.data2,aes(x=SURVIVE,y=HG)) +
240 geom_boxplot(notch=FALSE,varwidth=T) +
241 scale_x_discrete('Survival Status',breaks=c(0,1),
242 labels=c("Survived","Died"))
243
244
245 ggplot(data=physio.data2,aes(x=SURVIVE,y=HCT)) +
```

```
geom_boxplot(notch=FALSE,varwidth=T) +
scale_x_discrete('Survival Status',breaks=c(0,1),
labels=c("Survived","Died"))

table(Sex,SHOCK_TYP,SURVIVE)
#May be worth including interaction term Sex:SHOCK_TYP

grid.arrange(p1,p2,ncol=2)
grid.arrange(p3,p4,ncol=2)
grid.arrange(p5,p6,ncol=2)

########correlation plot


corrplot(cor(physio.data[sapply(physio.data, is.numeric)]))


#######chisq test for associations
chisq.test(SHOCK_TYP,SURVIVE)
#P-value =0.0007616

chisq.test(Sex,SURVIVE)
#P-value is 0.06376

##########t-tests for differences in means

t.test(AGE[SURVIVE==0],AGE[SURVIVE==1])
#P-value is 0.1882

t.test(HT[SURVIVE==0],HT[SURVIVE==1])
#P-values is 0.002579

t.test(SBP[SURVIVE==0],SBP[SURVIVE==1])
#P-value is 0.000116

t.test(MAP[SURVIVE==0],MAP[SURVIVE==1])
#p-value is 9.991e-005

t.test(HR[SURVIVE==0],HR[SURVIVE==1])
#p-value is 0.3006

t.test(DBP[SURVIVE==0],DBP[SURVIVE==1])
#P-value is 0.0003469

t.test(MCVP[SURVIVE==0],MCVP[SURVIVE==1])
#P-value is 0.004267

```

```
t.test(BSI[SURVIVE==0],BSI[SURVIVE==1])
#P-value is 0.02435

t.test(CI[SURVIVE==0],CI[SURVIVE==1])
#P-value is 0.2308

t.test(AT[SURVIVE==0],AT[SURVIVE==1])
#P-value is 0.3418

t.test(MCT[SURVIVE==0],MCT[SURVIVE==1])
#P-value is 0.08888

t.test(UO[SURVIVE==0],UO[SURVIVE==1])
#P-value is 0.0004551

t.test(PVI[SURVIVE==0],PVI[SURVIVE==1])
#P-value is 0.5683

t.test(RCI[SURVIVE==0],RCI[SURVIVE==1])
#P-value is 0.2342

t.test(HG[SURVIVE==0],HG[SURVIVE==1])
#P-value is 0.5729

t.test(HCT[SURVIVE==0],HCT[SURVIVE==1])
#P-value is 0.6216

t.test(HCT[SURVIVE==0],HCT[SURVIVE==1])
#P-value is 0.6216


###########cohen's d

cohen.d(AGE,SURVIVE)
#estimate is -0.2593205

cohen.d(HT,SURVIVE)
#estimate is 0.6589045

cohen.d(SBP,SURVIVE)
#estimate is 0.7908941

cohen.d(MAP,SURVIVE)
#estimate is 4.684463

cohen.d(HR,SURVIVE)
#estimate is -0.1999015
```

```
340
341 cohen.d(DBP,SURVIVE)
342 #estimate is 0.7417424
343
344 cohen.d(MCVP,SURVIVE)
345 #estimate 2.187079
346
347 cohen.d(BSI,SURVIVE)
348 #estimate is 0.4664643
349
350 cohen.d(CI,SURVIVE)
351 #estimate is 0.2275672
352
353 cohen.d(AT,SURVIVE)
354 #estimate is -0.182227
355
356 cohen.d(MCT,SURVIVE)
357 #estimate is -0.349621
358
359 cohen.d(UO,SURVIVE)
360 #estimate is0.5933606
361
362 cohen.d(PVI,SURVIVE)
363 #estimate is 0.1100068
364
365 cohen.d(AT,SURVIVE)
366 #estimate is -0.182227
367
368 cohen.d(MCT,SURVIVE)
369 #estimate is -0.349621
370
371 cohen.d(RCI,SURVIVE)
372 #estimate is .2094894
373
374 cohen.d(HG,SURVIVE)
375 #estimate is 0.1076231
376
377 cohen.d(HCT,SURVIVE)
378 #estimate is 0.09268039
379
380 ##################model selection
381
382 fit1<-glm(SURVIVE~.,data=physio.data2,family='binomial')
383 summary(fit1)
384
385 stepAIC(fit1,direction='both')
386 #stepAIC did not choose SBP and DBP even though they both
```

```
#appeared significant it may be because MAP is very highly
#correlated with SBP and DBP and it is slightly more
#correlated with survival status therefore stepAIC
#only picked MAP
#model chosen by stepAIC  below
#Call:  glm(formula = SURVIVE ~ SHOCK_TYP + MAP + MCVP + BSI + UO +
    RCI,

fit2<-glm(SURVIVE~.+Sex:SHOCK_TYP,data=physio.data2,family='
   binomial')
stepAIC(fit2,direction='both')
summary(fit2)
#interaction term deemed insignificant

fit3<-glm(SURVIVE~SHOCK_TYP+MAP+MCVP+BSI+UO+RCI,data=physio.data2,
   family='binomial')
stepAIC(fit3,direction='both')
summary(fit3)
#decided to get rid of UO and RCI based on results from
#summary(fit3). Kept MAP because there was a very large
#difference in means of MAP between the living and dead patients

###final model
fit4<-glm(SURVIVE~SHOCK_TYP+MAP+MCVP+BSI,data=physio.data2,family='
   binomial')
stepAIC(fit4,direction="both")
summary(fit4)
extractAIC(fit4)

############diagnostics
one.fourth.root=function(x){
x^0.25
}

source("examine.logistic.reg.R")


physio.data2$SURVIVE<-as.numeric(as.character(physio.data2$SURVIVE)
   )
SURVIVE<-as.numeric(as.character(SURVIVE))

g<-2
MAP_interval<-cut(MAP,quantile(MAP,0:g/g),include.lower=TRUE)
MCVP_interval<-cut(MCVP,quantile(MCVP,0:g/g),include.lower=TRUE)
BSI_interval<-cut(BSI,quantile(BSI,0:g/g),include.lower=TRUE)

w <- aggregate(formula = SURVIVE ~ SHOCK_TYP+MAP_interval+
```

```
   MCVP_interval+BSI_interval,
data = physio.data2, FUN = sum)
n<- aggregate(formula= SURVIVE ~SHOCK_TYP+MAP_interval+
   MCVP_interval+BSI_interval,
data=physio.data2,FUN=length)
w.n<-data.frame(w,trials=n$SURVIVE,prop=round(w$SURVIVE/n$SURVIVE
   ,2))
w.n

mod.prelim1<- glm(formula=SURVIVE/trials~SHOCK_TYP+MAP_interval+
   MCVP_interval+BSI_interval,
family=binomial(link=logit),data=w.n,
weights=trials)


save1<-examine.logistic.reg(mod.prelim,identify.points = T,
scale.n=one.fourth.root,scale.cookd=sqrt)

w.n.diag1<- data.frame(w.n,pi.hat=round(save1$pi.hat,2),
std.res=round(save1$stand.resid,2),
cookd=round(save1$cookd,2),
h=round(save1$h,2))

p<-length(mod.prelim1$coef)
ck.out<-abs(w.n.diag1$std.res)>2 |
w.n.diag1$cookd>4/nrow(w.n)|w.n.diag1$h>0.5
extract.EVPs<-w.n.diag1[ck.out,]
extract.EVPs
#a higher cutoff was used for leverage since there were so few data
    points
#and most data points were above the 3p/n cutoff

xtable(extract.EVPs[,c('SHOCK_TYP','MAP_interval','MCVP_interval','
   SURVIVE','BSI_interval','std.res','cookd','h')])

vif(fit4)
xtable(vif(fit4))

####################training and testing
set.seed(1)
n<- nrow(physio.data2)
p<-0.7
c<-0.4
train<- sample(n,p*n)
train_set<-physio.data2[train,]
test_set<-physio.data2[-train,]

```

```
fit_train <- glm(SURVIVE~SHOCK_TYP+MAP+MCVP+BSI,
data=train_set,
family=binomial(link="logit"))
nrow(physio.data2[-train,])
test_probs <- predict.glm(fit_train,newdata=test_set,
type='response')

predicted_class<- as.numeric(test_probs > c)
length(predicted_class)
length(test_set$SURVIVE)
test_confusion_matrix<-table(predicted_class,test_set$SURVIVE,
dnn=c("Predicted","Actual"))
test_confusion_matrix

sensitivity<-test_confusion_matrix[2,2]/(test_confusion_matrix
   [2,2]+test_confusion_matrix[1,2])
sensitivity
specificity<-test_confusion_matrix[1,1]/(test_confusion_matrix
   [1,1]+test_confusion_matrix[2,1])
specificity


####model inferences

betahat<-formatC(signif(fit5$coeff,digits=2),digits=2,format='f',
   flag='#')
OR<-formatC(signif(exp(fit5$coeff),digits=3),digits=3,format='f',
   flag='#')
SE<-formatC(signif(summary(fit5)$coeff[,2],digits=3),digits=3,
   format='f',flag='#')
cibounds<-formatC(signif(exp(confint(fit5)),digits=3),digits=3,
   format='f',flag='#')
pval<-formatC(signif(summary(fit5)$coeff[,4],digits=2),digits=2,
   format='f',flag='#')
x<-cbind(betahat,OR,SE,pval,matrix(paste("(",cibounds[,1],
",",cibounds[,2],")")))
colnames(x)<-cbind("Coefficient","Odds ratio","SE","p-value",
"95% CI on OR")
x
xtable(x)
```

Appendix B:Additional Boxplots of covariates by survival

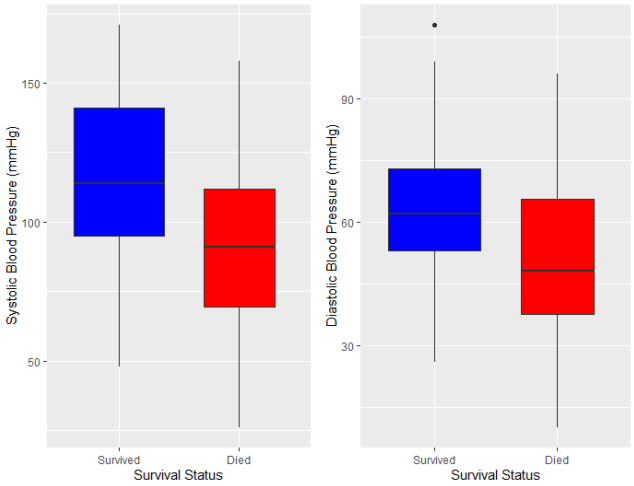

Figure 3: Boxplot of SBP by survival status (left) and Boxplot of DBP by survival status (right) these variables were highly correlated with MAP and were ultimately not used in the final model.

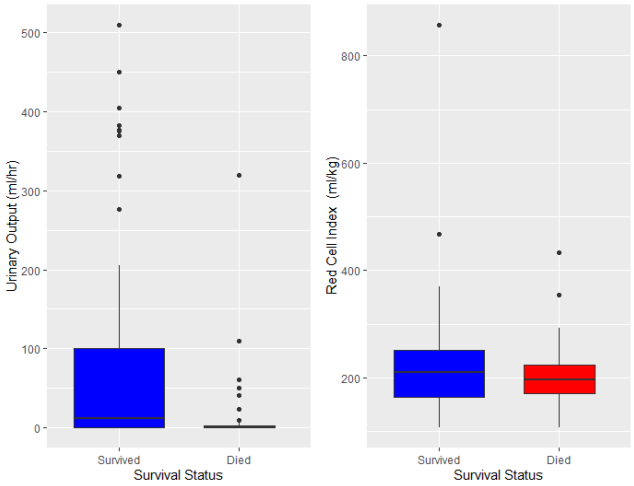

Figure 4: Boxplot of urinary output by survival status (left) and Boxplot of red cell index by survival status (right) removal of these variables increased AIC score by a marginal amount while simplifying the model.

Appendix C: Diagnostic Plots

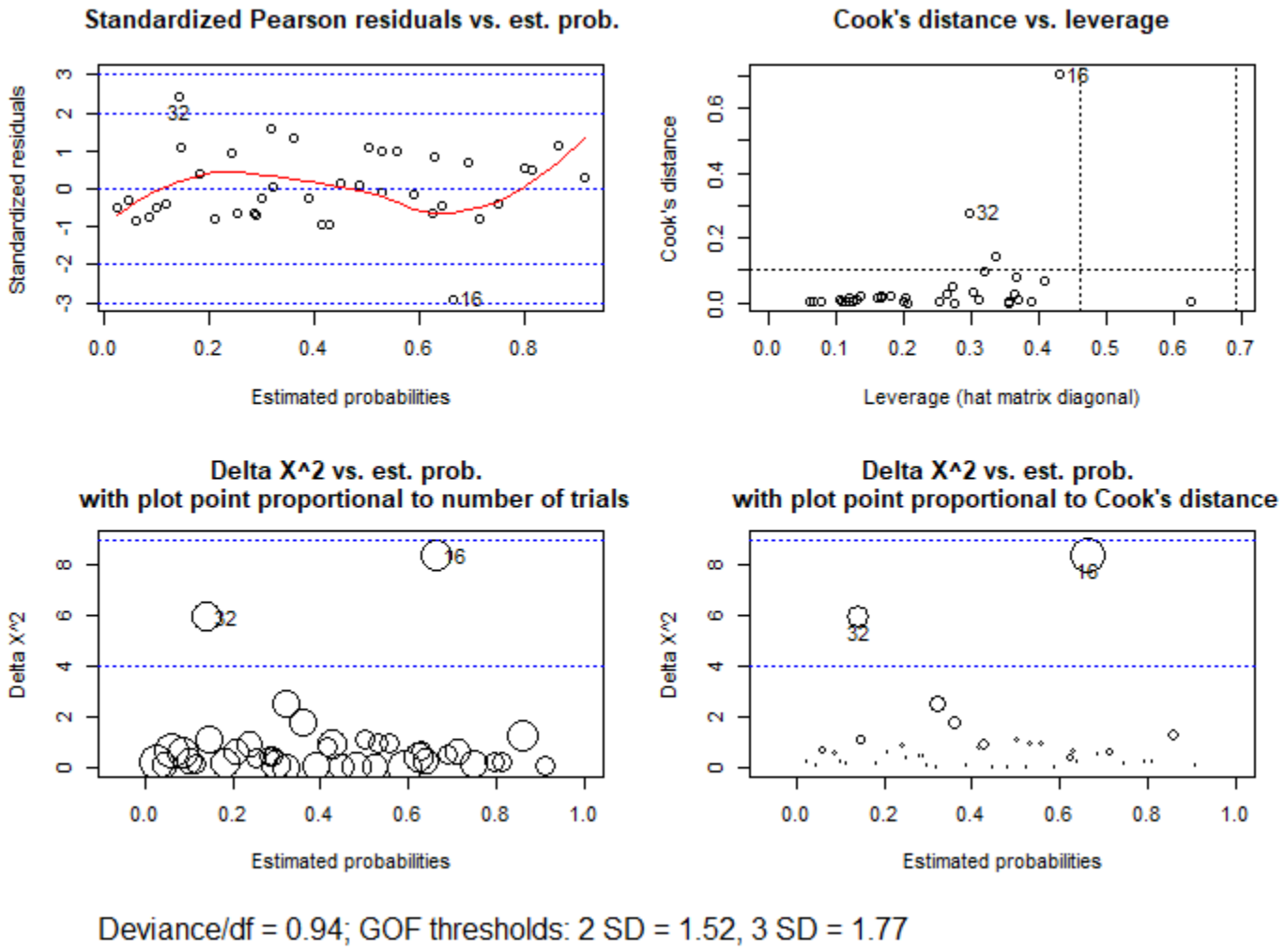

Figure 5: Residual plot (top left) Cook's distance vs leverage plot (top right) and Delta X^2 plots (bottom)