Assignment 3

Will Rauen

2024-03-13

Setup

For this investigation, we will be using a multitude of packages to answer a research question related to logistic regression.

Data

```
echo_data <- read.csv("/Users/williamrauen/Desktop/DS 3100/Assignments/Assignment 3/Data/echocardiogram

dim(echo_data)
```

```
## [1] 132  13
```

The data used in this dataset comes from the UC Irvine Machine Learning Repository and can be accessed with the link:

https://archive.ics.uci.edu/dataset/38/echocardiogram

This specific dataset takes data from 123 patients who had recently had a heart attack with a follow up a year later to see if patients were still alive. Looking at the dimension of the data, each of the 123 rows represents a different patient involved in the study while each column represents either a predictor variable, or classification. The predictor variables that are continuous have to do with the patients age at heart attack, months they have survived, and vital stats related to the patients heart. There are also some binary variables such as whether or not there is pericardial effusion fluid around the heart. The variable of interest and classification for this investigation is whether or not the patient survived a year after their heart attack, and is represented as a binary variable in the data set. There are also variables which were said to be ignored by the creators of the data set and those variables are consequently renamed to indicate that. Lastly, there was missingness in the data, so data wrangling will be required.

Varaibles of Interest

The dependent variable of interest as aforementioned is whether or not the patient survived at least a year after their heart attack. This is recorded in the data set as a binary variable in the data where a '0' indicates the patient died after 1 year and a '1' indicates the patient survived at least 1 year. The dependent variables of interest for the investigation are either continuous variables such as age of the patient, measures of contracility, size of the heart, and an index score of the health of heart walls, or a binary variable like the pericardial fluid variable mentioned before.

Research Question

In this assignment, I aim to find what features best determine whether or not a patient who suffered a heart attack will be alive after a one-year survival period.

Data Wrangling

Firstly, I renamed each of the variable columns, including ones I did not plan on using to make the data set clearer.

```
#Renaming columns

echo_data <- echo_data %>%
  rename(months_survival = X11) %>%
  rename(still_alive = X0) %>% #(binary)
  rename(age_at_heart_attack = X71) %>%
  rename(pericardial_effusion = X0.1) %>% #(binary)
  rename(fractional_shortening = X0.260) %>%
  rename(epss = X9) %>%
  rename(lvdd = X4.600) %>%
  rename(wall_motion_score = X14) %>%
  rename(wall_motion_index = X1) %>%
  rename(ignore_var = X1.1) %>%
  rename(ignore_var2 = X1.2) %>%
  rename(alive_at_1 = X0.2)
```

The next thing I made sure of was that there was no missing data points in the data set. Missing values in the data were represented with question marks ('?') so I wrangled the data set to get rid of these rows. There was also one row filled with NA so I also filtered that row out of the data set.

```
echo_data <- echo_data[rowSums(echo_data == '?') == 0, ]

echo_data <- echo_data %>% #Gets rid of NA values
  na.omit()
```

The next step is to subset the data so that I only have the columns of data that pertain to the investigation. We get rid of columns such as months survival and still alive since those would heavily influence our analysis as those automatically indicate survival and other columns that the data set strictly don't have a use for like the names column. One thing of note is that the column wall_motion_index was said to be used instead of wall_motion_score by the creators of the data set so I will be following suit.

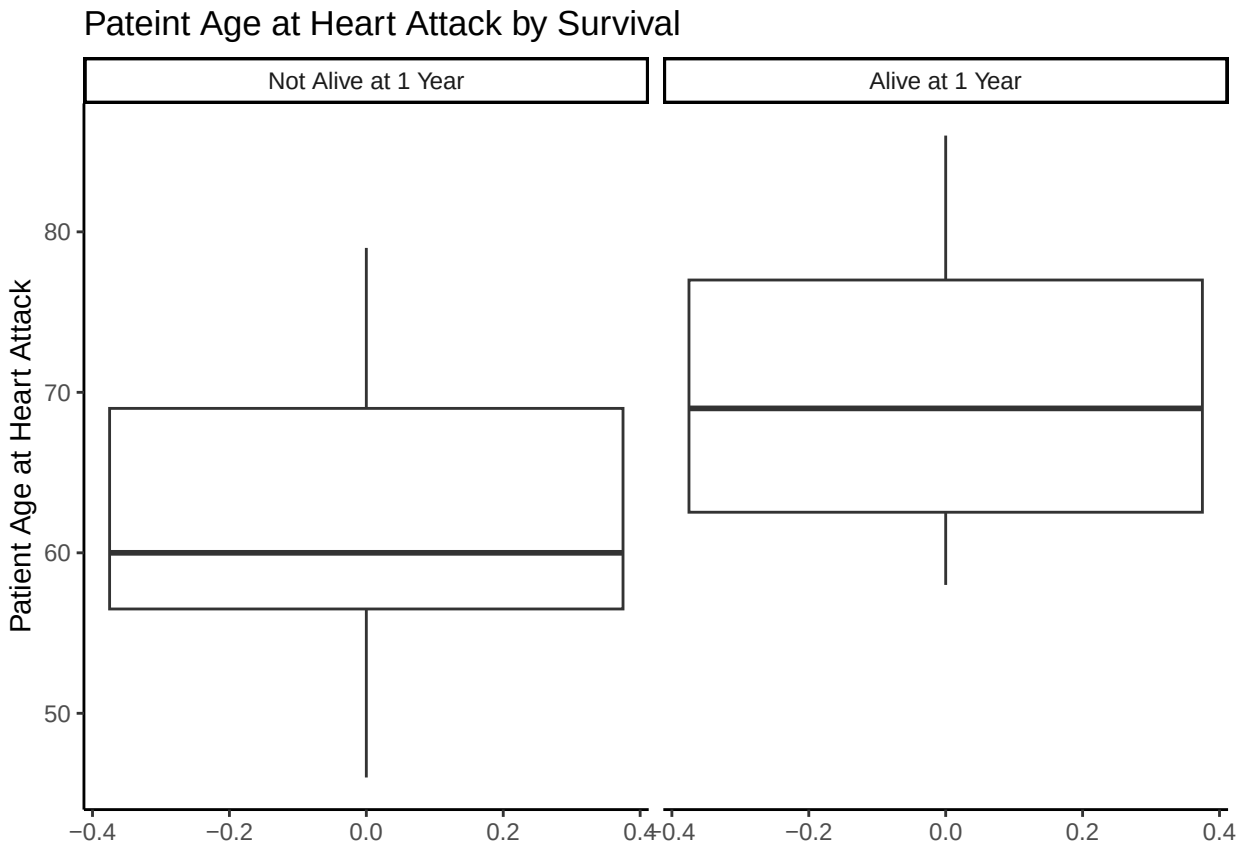
```
#Subsets data by each column
echo_new <- echo_data[c(3,4,5,6,7,9,13)]
```

From here, I realized each of the continuous variables were being treated as categorical values, so I converted each of the data points in the columns into to correct data type. I also made sure that the binary varaibles were converted to factor variables.

```
#Changing each variable into a numeric variable

echo_new <- echo_new %>%
  mutate(age_at_heart_attack = as.numeric(age_at_heart_attack)) %>%
  mutate(fractional_shortening = as.numeric(fractional_shortening)) %>%
  mutate(epss = as.numeric(epss)) %>%
  mutate(lvdd = as.numeric(lvdd)) %>%
  mutate(wall_motion_index = as.numeric(wall_motion_index)) %>%
  mutate(alive_at_1 = as.factor(alive_at_1)) %>%
  mutate(pericardial_effusion = as.factor(pericardial_effusion))
# make sure classification variable is type factor
```

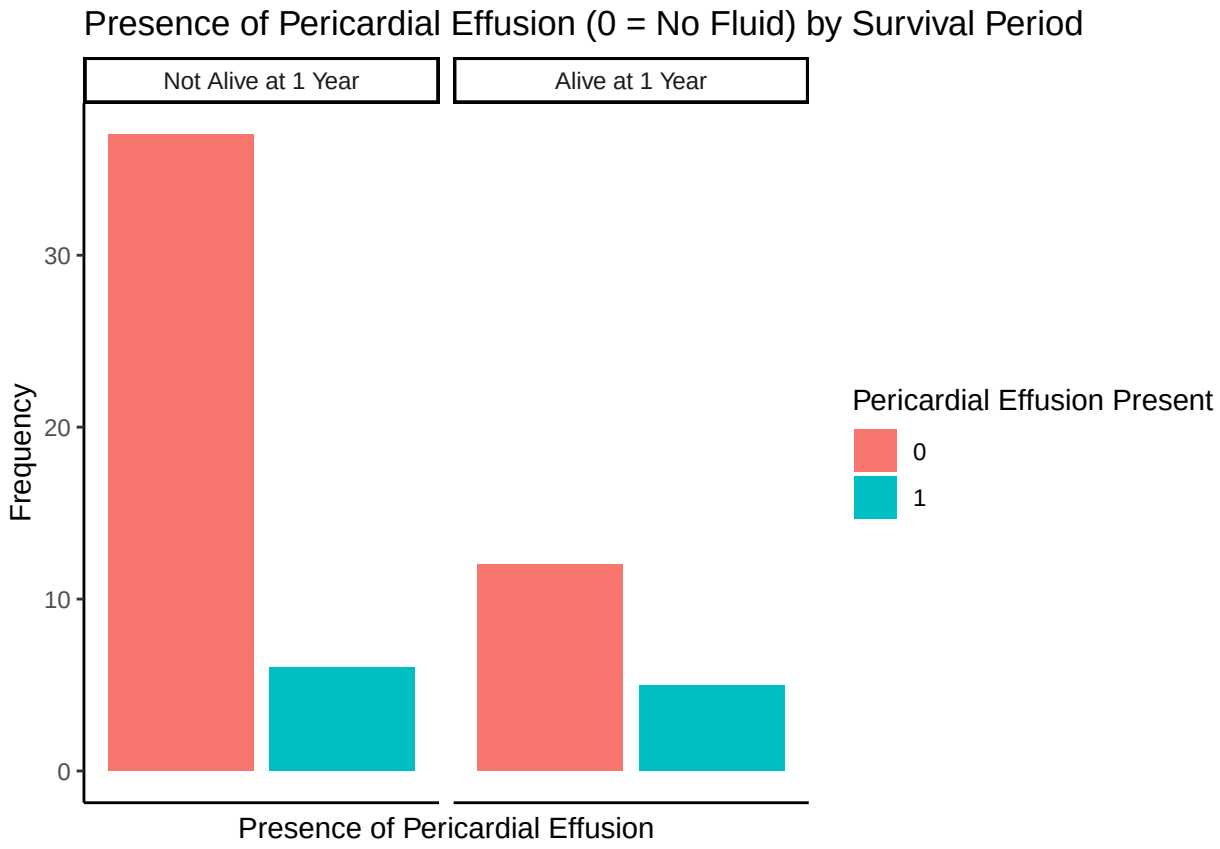
Lastly, lets visualize some of these variables by whether or not the patient made it through the one year survival period, and see if there are any readily apparent variables of interest for our dependent variable.

```
#Labels for data
custom_labels <- labeller(alive_at_1 = c("0" = "Not Alive at 1 Year",
                                         "1" = "Alive at 1 Year"))

#Graphing relationships between Independent and Dependent Variables

echo_new %>%
  ggplot(aes(x = age_at_heart_attack)) +
  geom_boxplot() +
  facet_wrap(~alive_at_1, labeller = custom_labels) +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) +
  coord_flip() +
  theme_classic() +
  labs(title = 'Pateint Age at Heart Attack by Survival',
       x = 'Patient Age at Heart Attack')
```

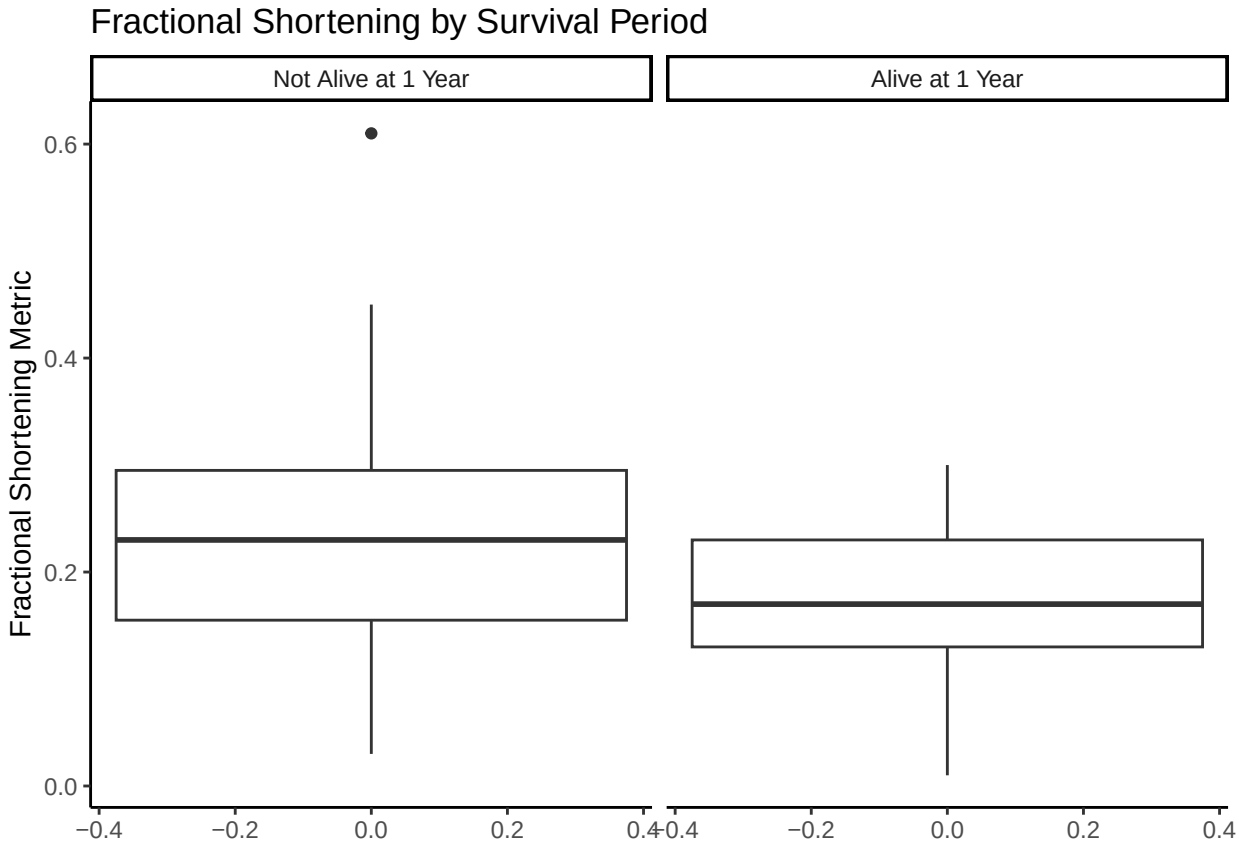



```
echo_new %>%
  ggplot(aes(x = pericardial_effusion, fill = pericardial_effusion)) +
  geom_bar() +
  facet_wrap(~alive_at_1, labeller = custom_labels) +
  theme_classic() +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank()) +
  labs(title = 'Presence of Pericardial Effusion (0 = No Fluid) by Survival Period',
       x = 'Presence of Pericardial Effusion',
       y = 'Frequency',
```

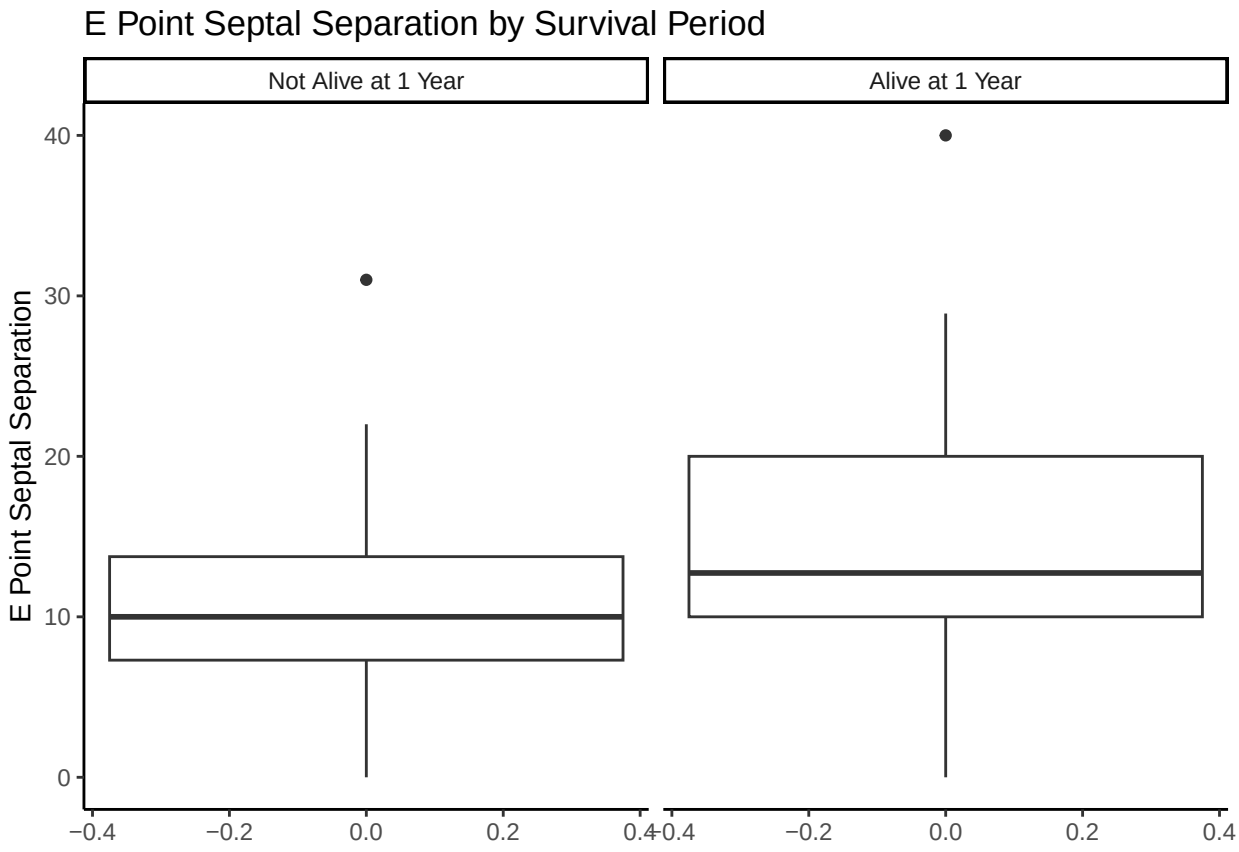
```
       fill = 'Pericardial Effusion Present')
```



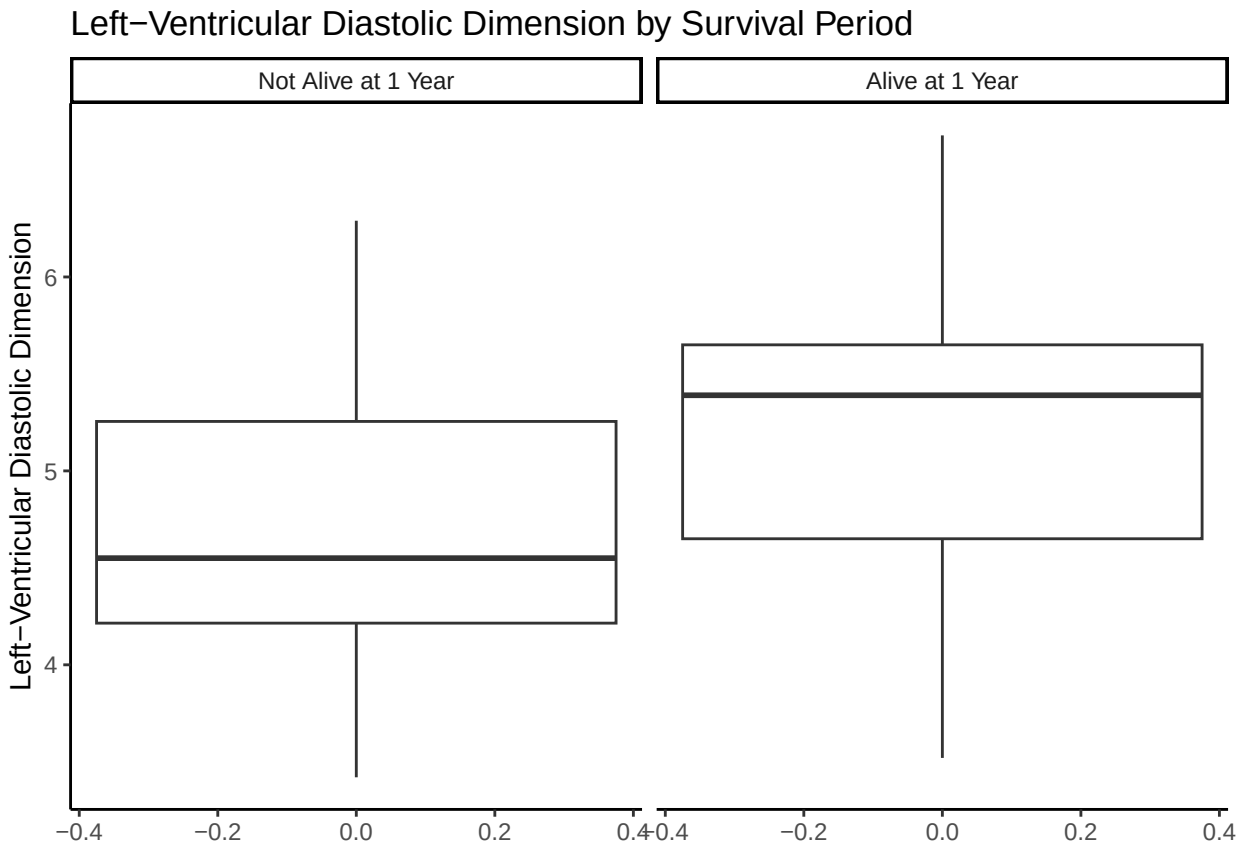

```
echo_new %>%
  ggplot(aes(x = fractional_shortening)) +
  geom_boxplot() +
  facet_wrap(~alive_at_1, labeller = custom_labels) +
  coord_flip() +
  theme_classic() +
  labs(title = 'Fractional Shortening by Survival Period',
       x = 'Fractional Shortening Metric')
```



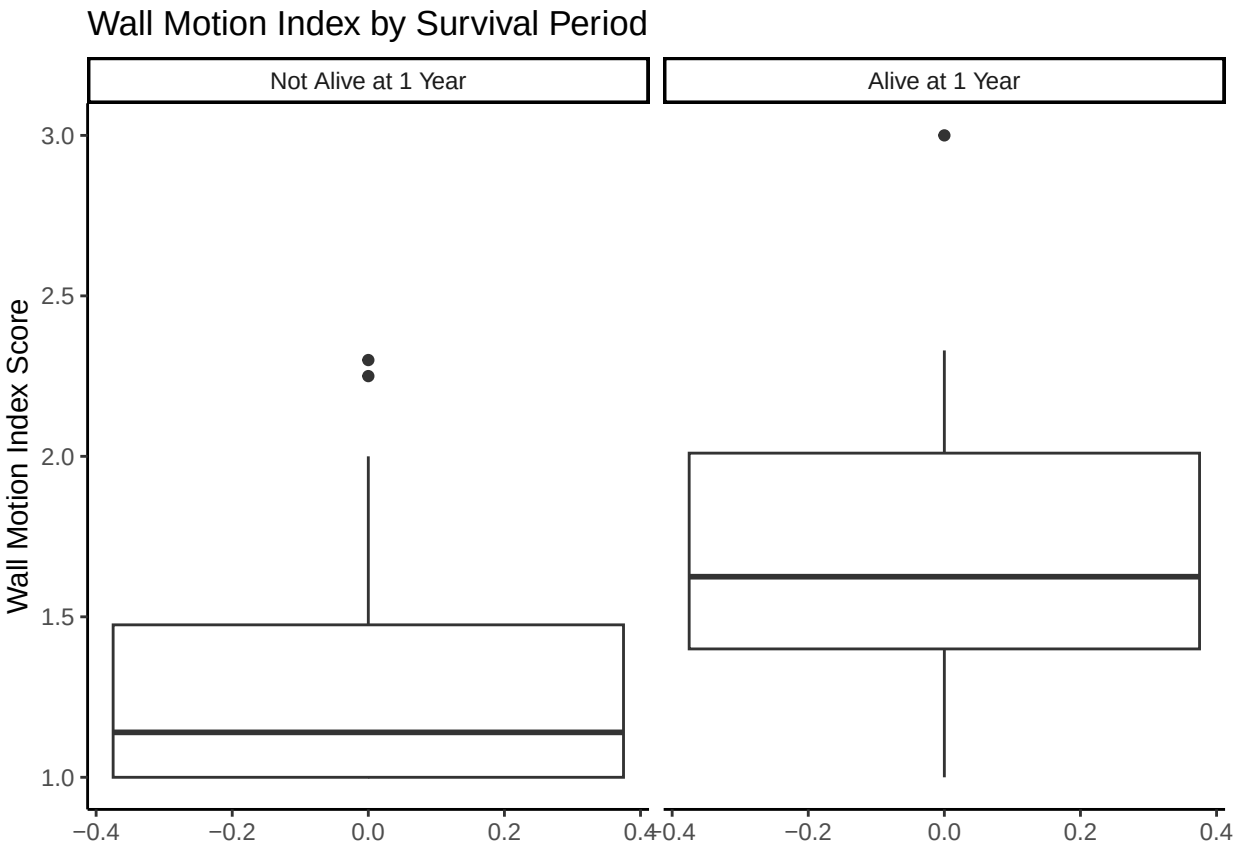

```
echo_new %>%
  ggplot(aes(x = epss)) +
  geom_boxplot() +
  facet_wrap(~alive_at_1, labeller = custom_labels) +
  coord_flip() +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) +
  theme_classic() +
  labs(title = 'E Point Septal Separation by Survival Period',
       x = 'E Point Septal Separation')
```




```
echo_new %>%
  ggplot(aes(x = lvdd)) +
  geom_boxplot() +
  facet_wrap(~alive_at_1, labeller = custom_labels) +
  coord_flip() +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) +
  theme_classic() +
  labs(title = 'Left-Ventricular Diastolic Dimension by Survival Period',
       x = 'Left-Ventricular Diastolic Dimension')
```




```
echo_new %>%
  ggplot(aes(x = wall_motion_index)) +
  geom_boxplot() +
  facet_wrap(~alive_at_1, labeller = custom_labels) +
  coord_flip() +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) +
  theme_classic() +
  labs(title = 'Wall Motion Index by Survival Period',
       x = 'Wall Motion Index Score')
```



After visualizing all the dependent variables, there does not immediately seem to be any blatant variables that heavily influence whether a patient is classified as living through the one-year survival period. Data analysis will hopefully narrow down these features.

Data Analysis

Feature Selection

The first part of our data analysis will have us run the logistic regression involving all the dependent variables mentioned above.

```
# Logistic regression involving all predicting variables

glm_echo <- glm(alive_at_1 ~ age_at_heart_attack + pericardial_effusion +
                  fractional_shortening + epss + lvdd + wall_motion_index,
                data = echo_new,
                family = binomial(link='logit'))

summary(glm_echo)
```

```
##
## Call:
## glm(formula = alive_at_1 ~ age_at_heart_attack + pericardial_effusion +
##     fractional_shortening + epss + lvdd + wall_motion_index,
##     family = binomial(link = "logit"), data = echo_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.7257  -0.5817  -0.3141   0.4756   2.4899
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -13.58225    4.84483  -2.803  0.00506 **
## age_at_heart_attack     0.11869    0.04723   2.513  0.01198 *
## pericardial_effusion1   1.00695    0.89160   1.129  0.25874
## fractional_shortening  -2.26566    4.07416  -0.556  0.57814
## epss                    0.01339    0.06006   0.223  0.82363
## lvdd                    0.37178    0.60522   0.614  0.53902
## wall_motion_index       2.01414    0.94771   2.125  0.03356 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 71.529  on 59  degrees of freedom
## Residual deviance: 47.917  on 53  degrees of freedom
## AIC: 61.917
##
## Number of Fisher Scoring iterations: 5
```

The initial logistic regression reveals that there only appear to be two significant features of the data that predict a patient living through the survival period: the age of the patient when they have the heart attack and the patient's heart wall health (heart motion index score). This is because these two features are below the alpha level of 0.05.

Another aspect of the prediction variables we can quickly check is if any of them are collinear. Using the car package we can measure variable multicollinearity by finding each of the variable's Variation Inflation Factor (VIF). Any value above 10 means we should be wary of collinearity.

```
#Testing for Multicollinearity using Variation Inflation Factor
vif(glm_echo)
```

```
##   age_at_heart_attack  pericardial_effusion fractional_shortening
##              1.079886              1.085084              1.161220
##                  epss                  lvdd     wall_motion_index
##              1.809475              1.608013              1.292498
```

None of the VIF scores indicate that there are variables that are collinear with one another as there are no VIF values above 10. Thus, we cannot immediately remove any features at the moment.

Returning to the initial logistic regression, we found that there were two significant features. Let's assume for now that these are the significant features we should use for our model.

```
# New Logistic regression model
glm_echo_new <- glm(alive_at_1 ~ age_at_heart_attack + wall_motion_index,
                    data = echo_new,
                    family = binomial(link='logit'))
summary(glm_echo_new)
```

```
##
## Call:
## glm(formula = alive_at_1 ~ age_at_heart_attack + wall_motion_index,
##     family = binomial(link = "logit"), data = echo_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.9043  -0.6518  -0.3274   0.5249   2.3487
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -12.94001    3.66576  -3.530 0.000416 ***
## age_at_heart_attack   0.12389    0.04566   2.713 0.006659 **
## wall_motion_index     2.56657    0.85539   3.000 0.002696 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 71.529  on 59  degrees of freedom
## Residual deviance: 50.422  on 57  degrees of freedom
## AIC: 56.422
##
## Number of Fisher Scoring iterations: 5
```

```
vif(glm_echo_new)
```

```
## age_at_heart_attack   wall_motion_index
##            1.075986            1.075986
```

This model reveals that the two variables are even more significant predictive features of our model, and once again, the VIF of the variables in the new model reveal that they are not collinear. As a result, it is safe to move forward with this model for our logistic regression.

Odds Interpretations and Average Marginal Effects

As a reminder, the dichotomous variable for the survival period are as follows:

0 = Did not live to 1 year survival period 1 = Lived through the 1 year survival period

```
#Odds Ratio
echo_coef <- coef(glm_echo_new)

echo_coef
```

```
##         (Intercept) age_at_heart_attack   wall_motion_index
##          -12.940006            0.123885            2.566574
```

```
# Log Odds Ratio
exp(echo_coef)
```

```
##         (Intercept) age_at_heart_attack   wall_motion_index
##        2.400085e-06        1.131886e+00        1.302114e+01
```

The log odds ratio coefficients are 0.12 for the age_at_heart_attack variable and 2.57 for the wall_motion_index. This can be interpreted as for every unit increase in the patients age, the log-odds that patient does not make it through the survival period increases by 0.12 units. Similarly, the other feature can be interpreted as for every unit increase in a patient's heart wall motion index score, the log odds that patient makes it through the survival period increases by 2.57. This is all assuming we hold the other variable constant in each of these scenarios.

The odds ratios similarly tells us that for every 1 unit increase in age of the patient, the odds of making it through the 1 year survival period increases by a factor of 1.13 holding constant other features in the model. Consequently, for every 1 unit increase in heart wall health of the patient, the odds of being starter increases by a factor of 13.

Using the margins and Desctools packages, we can also find the McFadden R squared value along with the average marginal effects.

```
# Finds average marginal effects
marg_echo <- margins(glm_echo_new)
summary(marg_echo)
```

```
##               factor    AME     SE      z      p  lower  upper
##  age_at_heart_attack 0.0167 0.0047 3.5176 0.0004 0.0074 0.0259
##    wall_motion_index 0.3452 0.0803 4.2965 0.0000 0.1877 0.5026
```

```
#Finds McFadden's' R Squared Score
PseudoR2(glm_echo_new)
```

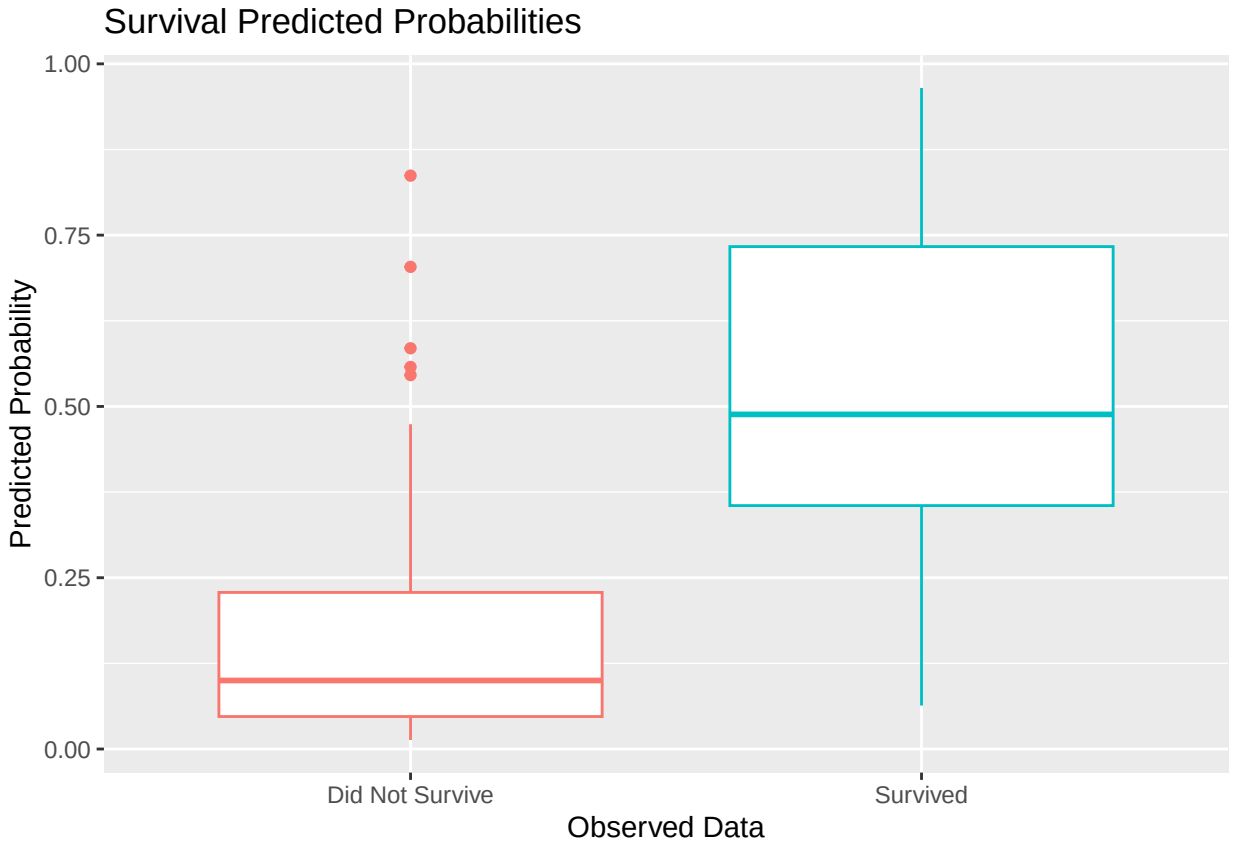
```
##  McFadden
## 0.2950827
```

Once again, the AME's reaffirm what we found before with the odds and log odds ratios. These are essentially the closest analogous values for a slope we can have. The values can be interpreted as for every 1 unit increase in the age of the patient, the probability of making it through the one year survival period increase by 1.67% and for every one unit increase in heart-wall motion score, the probability of making it through the one year survival period increases by 34.52%. The McFadden R-square found is 0.295 which is between the range of 0.2 and 0.4 which indicates this is a good model fit.

Model Prediction

To test the model, we need to test the fitted values from the logistic regression model to our data. Once we do so, we can visualize these predited values using a boxplot.

```
# Create data frame for prediction comparison with original data and new fitted values
echo_pred <- as.data.frame(cbind(echo_new$alive_at_1,
                                 glm_echo_new$fitted.values))

# Renaming column names in the data frame
echo_pred <- echo_pred %>%
  rename(original = V1, predprob = V2) %>%
  mutate(original = ifelse(original == '1', 0, 1)) #correction in dataframe

# Labeling the original data by whether or not they survived
echo_pred$original <- factor(echo_pred$original,
                         labels = c('Did Not Survive', 'Survived'))

# Using 0.5 as decision boundary to create new column for predicted class
echo_pred$predicted_factor <- ifelse(echo_pred$predprob > .5, 1, 0)

# Labeling predicted values by their factors
echo_pred$predicted_factor <- factor(echo_pred$predicted_factor,
                                     labels = c('Did Not Survive', 'Survived'))

#Boxplot to understand range of predicted probabilities
ggplot(echo_pred, aes(x=original, y=predprob, group=original, col=original)) +
  geom_boxplot() +
  labs(title = 'Survival Predicted Probabilities',
       x = 'Observed Data',
       y='Predicted Probability') +
  theme(legend.position = "none")
```



Since we've found our predicted probabilities and correctly labelled them, we can successfully create a mosaic plot that accurately summarizes our data.

```
# Creates confusion matrix for our model prediction
confusionMatrix(echo_pred$predicted_factor, echo_pred$original,
                positive = 'Survived')
```

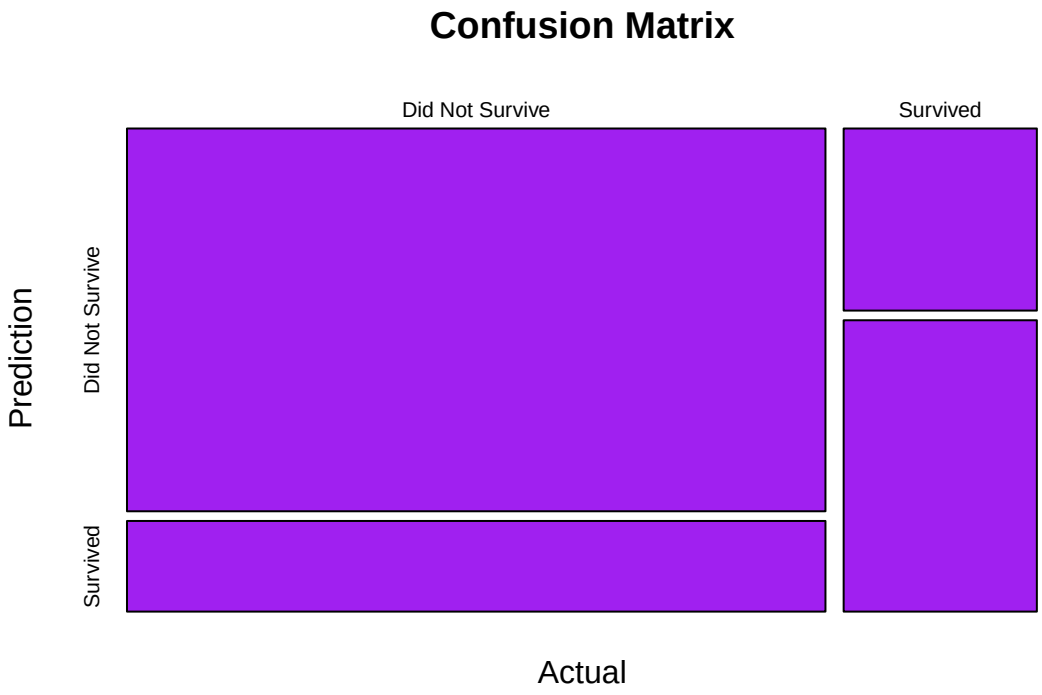
```
## Confusion Matrix and Statistics
##
##                  Reference
## Prediction        Did Not Survive Survived
##   Did Not Survive              38        9
##   Survived                      5        8
##
##                Accuracy : 0.7667
##                  95% CI : (0.6396, 0.8662)
##     No Information Rate : 0.7167
##     P-Value [Acc > NIR] : 0.2402
##
##                   Kappa : 0.3814
##
##  Mcnemar's Test P-Value : 0.4227
##
##             Sensitivity : 0.4706
##             Specificity : 0.8837
##          Pos Pred Value : 0.6154
##          Neg Pred Value : 0.8085
##              Prevalence : 0.2833
```

```
##          Detection Rate : 0.1333
##    Detection Prevalence : 0.2167
##       Balanced Accuracy : 0.6772
##
##        'Positive' Class : Survived
##
```

```
# Mosaic Table
echo_mos_tab <- table(echo_pred$predicted_factor, echo_pred$original)
echo_mos_tab
```

```
##
##                   Did Not Survive Survived
##   Did Not Survive              38        9
##   Survived                      5        8
```

```
# Mosaic Plot creation
mosaicplot(echo_mos_tab, main = 'Confusion Matrix', xlab = 'Actual', ylab = 'Prediction',
           color = 'purple')
```



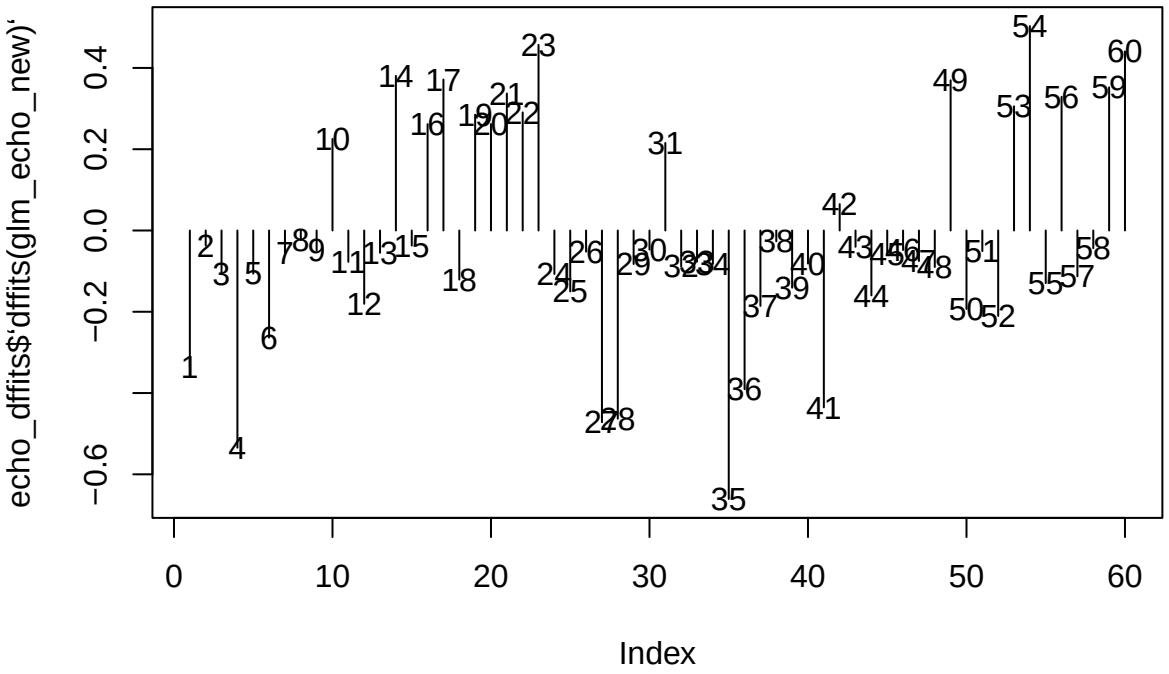
The mosaic table reveals that our model has mediocre sensitivity with a value of 0.47 and high specificity with a value of 0.88. This means our model has the drawback of predicting false negatives with the benefit of rarely predicting false positives. In the context of this investigation, this means our model is more likely to predict someone will not make it through the 1 year survival period when they actually do compared to a patient who is predicted to make it through the 1 year survival period but actually do not.

Outlier Detection

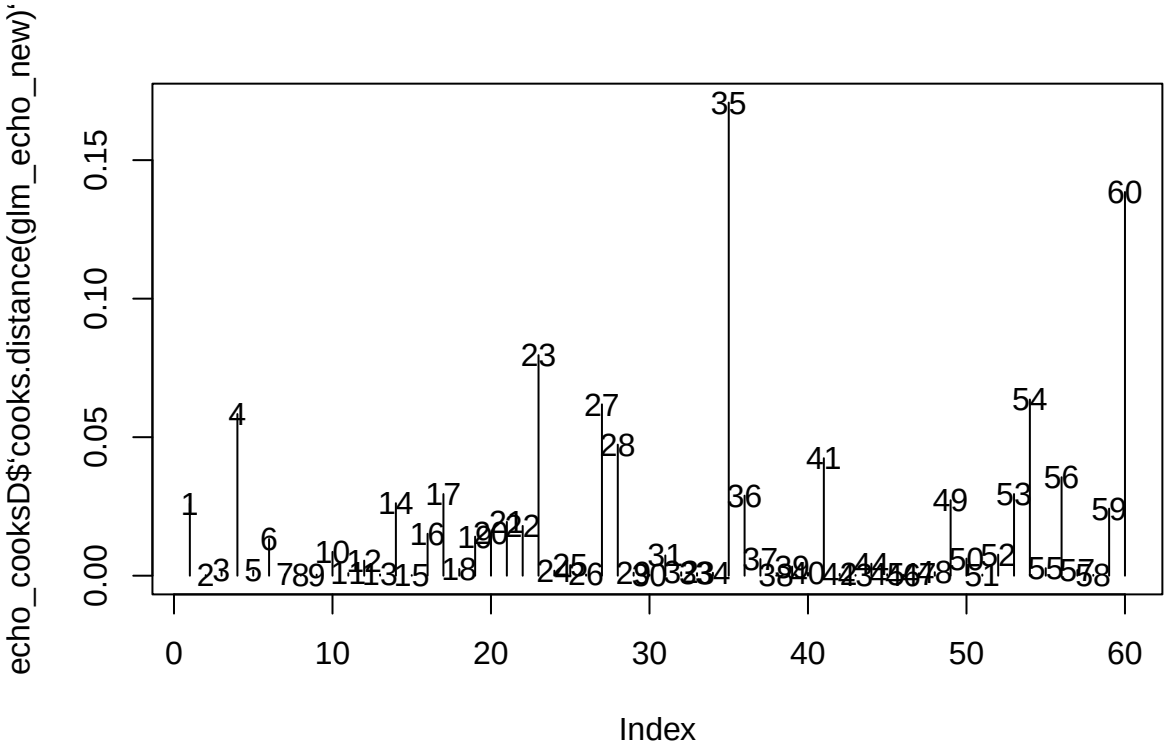
A risk that this model has is that it can be heavily influenced by outliers considering the small data sample of 60. Using DFBETAS, Cook's Distance, and DFFITS, we can try to identify possible outliars in the data set.

```
#DFFITS
echo_dffits <- as.data.frame(dffits(glm_echo_new))
plot(echo_dffits$`dffits(glm_echo_new)`, type='h')
```

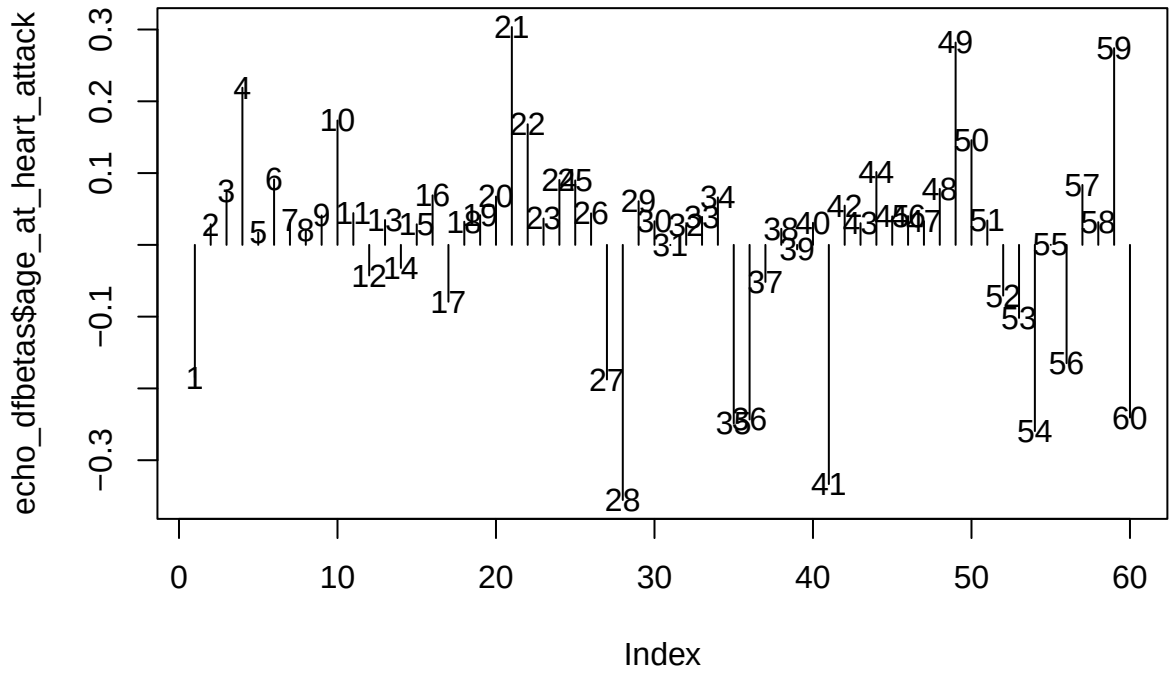
```
text(echo_dffits$`dffits(glm_echo_new)`,
     row.names(echo_dffits$`dffits(glm_echo_new)`))
```



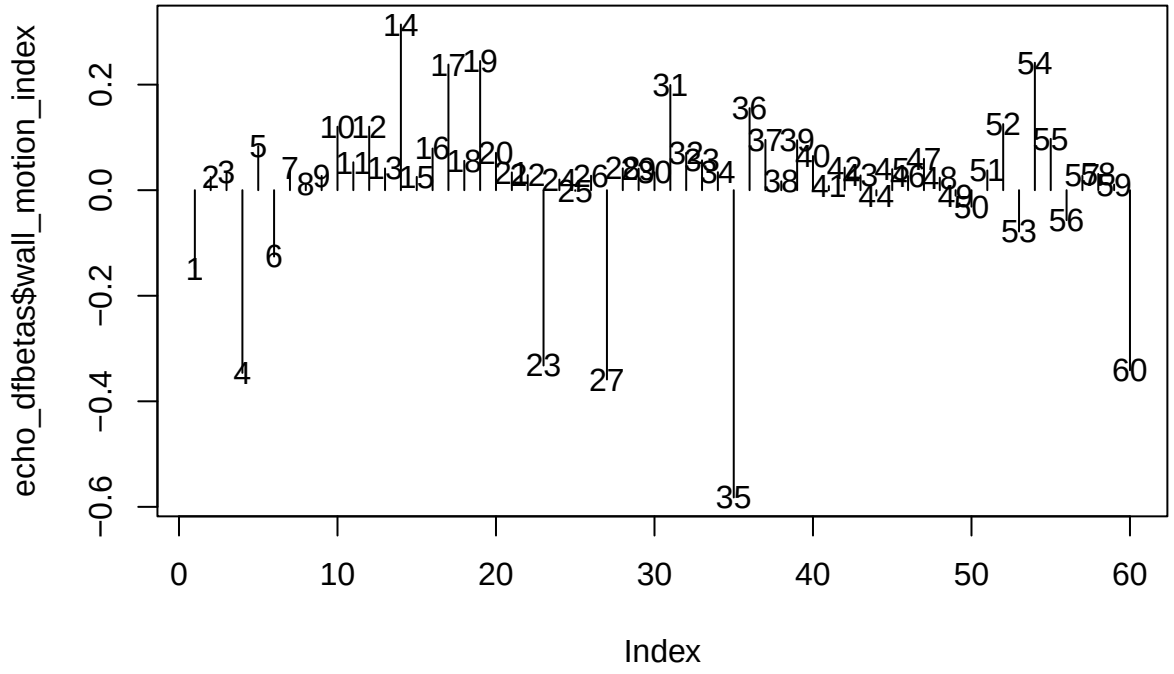

```
#COOKS DISTANCE
echo_cooksD <- as.data.frame(cooks.distance(glm_echo_new))
plot(echo_cooksD$`cooks.distance(glm_echo_new)`, type='h')
text(echo_cooksD$`cooks.distance(glm_echo_new)`,
     row.names(echo_cooksD$`cooks.distance(glm_echo_new)`))
```



```
#DFBETAS
echo_dfbetas <- as.data.frame(dfbetas(glm_echo_new))

plot(echo_dfbetas$`age_at_heart_attack`, type='h')
text(echo_dfbetas$`age_at_heart_attack`,
     row.names(echo_dfbetas$`age_at_heart_attack`))
```




```
plot(echo_dfbetas$`wall_motion_index`, type = 'h')
text(echo_dfbetas$`wall_motion_index`,
     row.names(echo_dfbetas$`wall_motion_index`))
```



A common outlier across these tests is row 35. Thus, it is important to see how our results would change by removing this row.

Re-Running the Model

When rerunning the model, the first thing we need to do is get rid of the 35th row of data. After that, we can follow the same steps as before in the investigation.

```
#Getting rid of the 35th row in the data
echo_data2 <- echo_new[-35, ]
```

```
# Logistic regression involving all predicting variables

glm_echo2 <- glm(alive_at_1 ~ age_at_heart_attack + wall_motion_index,
                 data = echo_data2,
                 family = binomial(link='logit'))

summary(glm_echo2)
```

```
##
## Call:
## glm(formula = alive_at_1 ~ age_at_heart_attack + wall_motion_index,
##     family = binomial(link = "logit"), data = echo_data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7593  -0.5869  -0.2760   0.4459   2.4587
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -14.7061     4.0873  -3.598 0.000321 ***
## age_at_heart_attack   0.1380     0.0490   2.816 0.004860 **
## wall_motion_index     3.1770     0.9986   3.181 0.001465 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.856  on 58  degrees of freedom
## Residual deviance: 46.291  on 56  degrees of freedom
## AIC: 52.291
##
## Number of Fisher Scoring iterations: 5
```

The model has similar log odds values for the predicting features and are similarly statistically significant. This would give us analogous interpretations for the odds and AME interpretations.

```
# Predictions for rerun model

echo_pred2 <- as.data.frame(cbind(echo_data2$alive_at_1,
                                  glm_echo2$fitted.values))

# Renaming column names in the data frame
echo_pred2 <- echo_pred2 %>%
  rename(original = V1, predprob = V2) %>%
  mutate(original = ifelse(original == '1', 0, 1)) #correction in dataframe
```

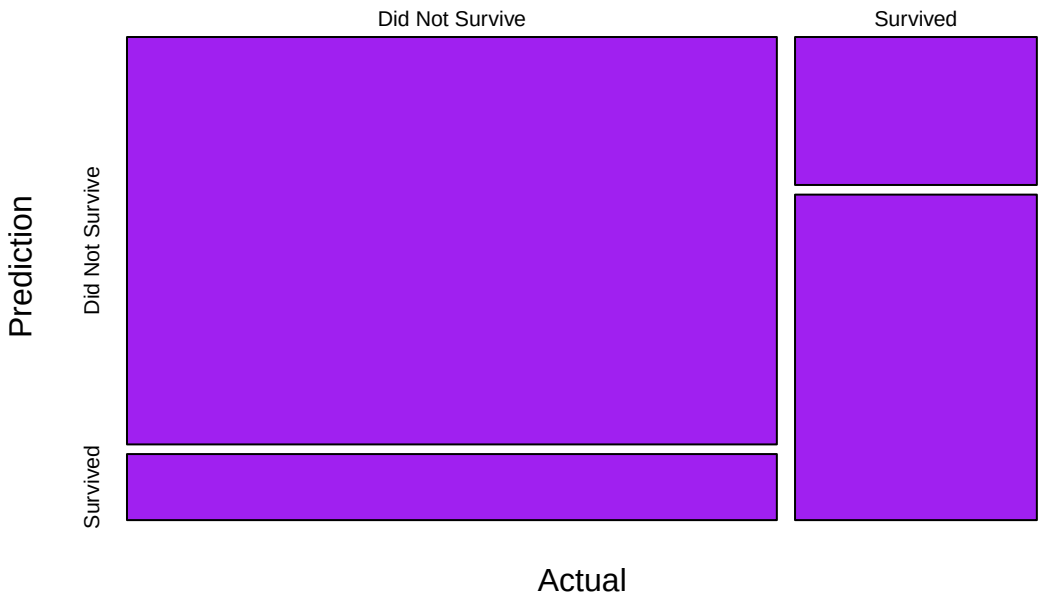
```r
# Labeling the original data by whether or not they survived
echo_pred2$original <- factor(echo_pred2$original,
                        labels = c('Did Not Survive', 'Survived'))

# Using 0.5 as decision boundary to create new column for predicted class
echo_pred2$predicted_factor <- ifelse(echo_pred2$predprob > .5, 1, 0)

# Labeling predicted values by their factors
echo_pred2$predicted_factor <- factor(echo_pred2$predicted_factor,
                                      labels = c('Did Not Survive', 'Survived'))

confusionMatrix(echo_pred2$predicted_factor, echo_pred2$original,
                positive = 'Survived')
```

```
## Confusion Matrix and Statistics
##
##                  Reference
## Prediction        Did Not Survive Survived
##   Did Not Survive              37        6
##   Survived                      5       11
##
##                Accuracy : 0.8136
##                  95% CI : (0.6909, 0.9031)
##     No Information Rate : 0.7119
##     P-Value [Acc > NIR] : 0.05267
##
##                   Kappa : 0.5374
##
##  Mcnemar's Test P-Value : 1.00000
##
##             Sensitivity : 0.6471
##             Specificity : 0.8810
##          Pos Pred Value : 0.6875
##          Neg Pred Value : 0.8605
##              Prevalence : 0.2881
##          Detection Rate : 0.1864
##    Detection Prevalence : 0.2712
##       Balanced Accuracy : 0.7640
##
##        'Positive' Class : Survived
##
```

```r
# Mosaic Table
echo_mos_tab2 <- table(echo_pred2$predicted_factor, echo_pred2$original)
echo_mos_tab2
```

```
##
##                   Did Not Survive Survived
##   Did Not Survive              37        6
##   Survived                      5       11
```

```r
# Mosaic Plot creation
mosaicplot(echo_mos_tab2, main = 'Confusion Matrix', xlab = 'Actual', ylab = 'Prediction',
           color = 'purple')
```



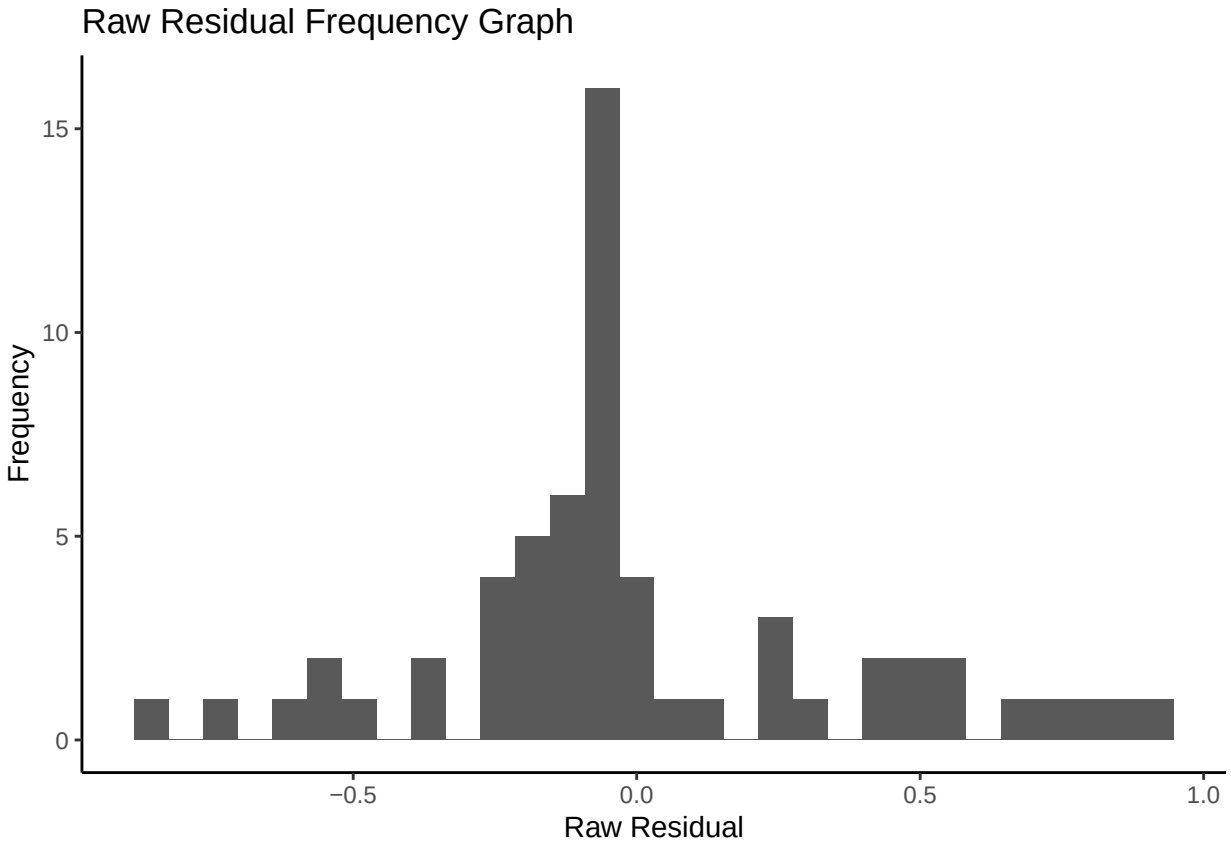
The major difference between this model (removal of outlier) and the last one is that it increased the sensitivity of the model by roughly 20%. However, Mcnemar's P-value is extremely close to 1 which would indicate that this model is not statistically significant at all compared to the previous model where it was around 0.42. Thus, there are drawbacks of both models, and it is difficult to determine if one is better than the other.

Residual Analysis & Independence Discussion

In logistic regression, there is no one formal way to determine or test independence of the data without a vector of residuals. However, we have to assume an independence of data under the assumption of logistic regression. In order to do so, we look at outliers, residuals, and measures of collinearity to gauge the strength of the model and the data, but we can also think about our data critically. Given that our sample is of different patients, we can assume that they are likely independent of one another (unless there are confounding variables e.g.:perhaps some of the patients are related to each other) with the limitation that we don;t have great sample size.

Nonetheless, it is also important to take a look at the residuals of the data set. The simplest type of residuals are called 'response residuals' or otherwise known as 'raw residuals' which is essentially the difference between the observation of the data and its predicted probability. For this part of the investigation, we will use the original data from before we took out the outlier.

```
# Calculating Residuals and graphing Histogram Plot

echo_pred %>%
  mutate(residual = ifelse(original == 'Did Not Survive', #ifelse for Residual
                           0 - predprob,
                           1 - predprob)) %>%
  ggplot(aes(x = residual)) +
  geom_histogram(bins = 30) +
  theme_classic() +
  labs(title = 'Raw Residual Frequency Graph',
       x = 'Raw Residual',
       y = 'Frequency')
```



The residual plot reveals that the raw residuals are somewhat normally distributed which is an indicator of independent residuals. This plot has some spaces and gaps as it tapers to larger magnitude of values but this can be explained by the fact that our data is not a very large sample. Ultimately, using what we know about the residuals, measures of multicollinearity, outliers, and context of the data set, I think it is safe to say that our data adheres to general assumptions of independence for logistic regression.

Discussion

This investigation used a data set involving patients who had heart attacks and aimed to determine the significant features in determining whether these patients would live to the 1 year survival period. Once the data was cleaned, we were left with 60 data points of patients with 6 possible prediction variables we could use for the logistic regressions. After running the initial logistic regression with all possible prediction variables, we found that there were two significant variables of the data set: age of the patient and the patient's heart wall index score. Isolating the two features in another logistic regression sequence also showed the two variables remained significant, and after confirming they were not collinear using their VIF values, I determined these were likely the best variables for predicting whether a patient made it through the one year survival period.

The logistic regression output revealed that there were positive relations between the age of the patient and their heart wall health and the likelihood they lived through the 1 year survival period. Upon looking at these results, it made sense for the heart wall variable to be correlated with greater likelihood of surviving, but the age variable seems to be out of place. It is important to know that the age variable and a very minimal impact on survival rate compared to the heart wall variable yet was statistically significant. This is likely a direct result of not having enough data points since even when the outlier was removed for a later diagnostic the results remained similar. Later in the analysis, the logistic model revealed a mediocre sensitivity value but high specificity value which in the context of the investigation is likely a good thing as it indicates a patient that is predicted to not make it through 1 year is more likely to beat the diagnosis, than a patient

who is predicted to survive 1 year and not make it.

As aforementioned, outlier diagnostics were conducted and I determined one patient was an outlier. The new model ran without the outlier revealed similar results to the original logistic regression model but had more accurate predictions (see mosaic plot). However this model was not as statistically significant according to the Mcnemar's P value so I decided to finish the investigation using the original logistic regression model. The last thing discussed in the investigation found that the model likely upheld assumptions of independence of logistic regression by looking at residuals, outliers, and VIF scores.

Ultimately, the investigation concluded that the most significant features in predicting whether or not a patient makes it through a 1 year survival period after suffering a heart attack are their age and heart wall health. However, interpretations of the logistic model used found some conflicting results as higher age marginally correlated with a higher likelihood of survival. This was likely due to the small sample size of the data so future investigations may want to look at the same predicting features but with a greater sample size.

Link to Data:

https://archive.ics.uci.edu/dataset/38/echocardiogram