The Correlation Between Echocardiography Results and Cardiovascular Patients' Survival Months

Anjali Chauhan, Erin Zhu, Fan Cheng

1 Abstract

Cardiovascular diseases (CVDs) are the number 1 cause of death globally. In the United Statesalone, approximately 1.5 million heart attacks and strokes occur every year in the United States.More than 800,000 people in the United States die from cardiovascular disease each year—that's 1in every 3 deaths, and about 160,000 of them occur in people under age 65 [1]. Echocardiography Is one of the most prominent methods used for CVD diagnosis. In this study, we aim to examine whether echocardiography results are good predictors for patients' total survival months, and potentially identify the key echocardio factors that are indicative of the patients' survivals. The data are obtained from a U.S. clinical trial studied by Steven Salzberg at Harvard University in 1989 [2]. All 132 patients suffered from heart attacks at some point before the study was conducted. Echocardiography examinations were performed on all patients. Multivariate regression analysis and exhaustive, forward and backward cross validation were performed to study the potential correlations between survival months and echocardiography results. However, we found that the variables were unable to generate a good predictive linear regression model, with the highest r-squared model being approximately 30 percent. By defining the existing limitations of the analysis, the study confirms our future research objective of using cox regression and decision trees to further study the correlation between echocardiography results and patients' survival.

2 Introduction

Heart attack, or myocardial infarction (MI), occurs when the flow of blood to the heart is blocked. WHO reported that cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year [3]. CVDs create a significant impact both socially and economically. According to the American Heart Association, the direct and indirect costs of cardiovascular diseases and stroke totaled more than 316.6 billion, which includes health expenditures and lost productivity [4]. Echocardiography is one of the most prominent methods used for CVD diagnosis as it provides crucial information including the variables listed in the Methodology below. In this research, we aim to use the echocardiography explanatory variables listed above to predict the patients' total survival months. This analysis will help identify the key echocardio factors that affect the patients' survivals, and examine whether echocardiography results are good predictors for patients' total survival months.

3 Methodology

3.a Data Summaries

The data set was collected in 1989, it contains information of 132 patients in a U.S. clinical trial studied by Steven Salzberg at Harvard University. All 132 patients suffered from heart attacks at some point before the study was conducted. Echocardiographic examinations were performed on all patients. The variables studied include:

- Survival Month : The number of months a patient survived (has survived, if patient is still alive). Because all the patients had their heart attacks at different times, it is possible that some patients have survived less than one year but they are still alive. Check the second variable to confirm this. Such patients cannot be used for the prediction task mentioned above.The unit of this variable is months.

- Still-alive : Dummy variable. **0 = dead at end of survival period, 1 = still alive.**

- Age at Heart Attack : Age in years when the patient's first heart attack occurred. The unit of this variable is years.

- Pericardial-Effusion (PE) : Dummy Variable. Pericardial effusion is fluid around the heart. **0=no fluid, 1=fluid.** Pericardial effusion is the buildup of extra fluid in the space around the heart. If too much fluid builds up, it can put pressure on the heart. This can prevent it from pumping normally. PE is examined using echocardiography.

- Fractional-Shortening (FS) : A measure of contractility around the heart. Fractional shortening (FS) is calculated by measuring the percentage change in left ventricular diameter during systole, examined using echocardiography. The unit of FS variable is observed as percentage values (%).

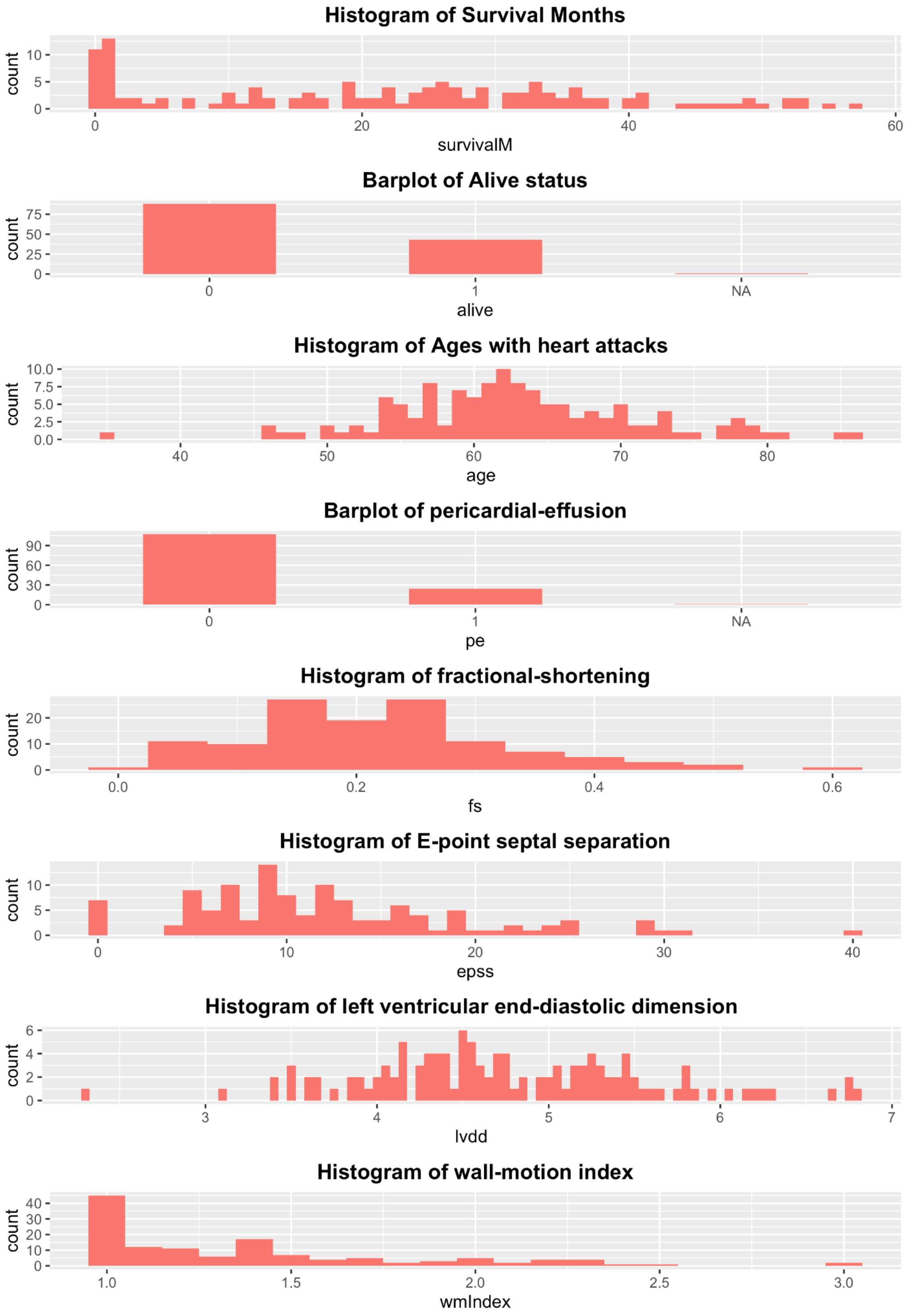
- EPSS : E-point septal separation, another measure of contractility. Larger Numbers are increasingly abnormal. EPSS can be examined using echocardiography by measuring distance in space separating the anterior MV leaflet from the septal wall. The unit of EPSS is millimeter (mm).

- LVDD : left ventricular end-diastolic dimension, examined using echocardiography. This is a measure of the size of the heart at the end-diastole. Large hearts tend to be sick hearts.The LVDD variable is observed as percentage values (%).

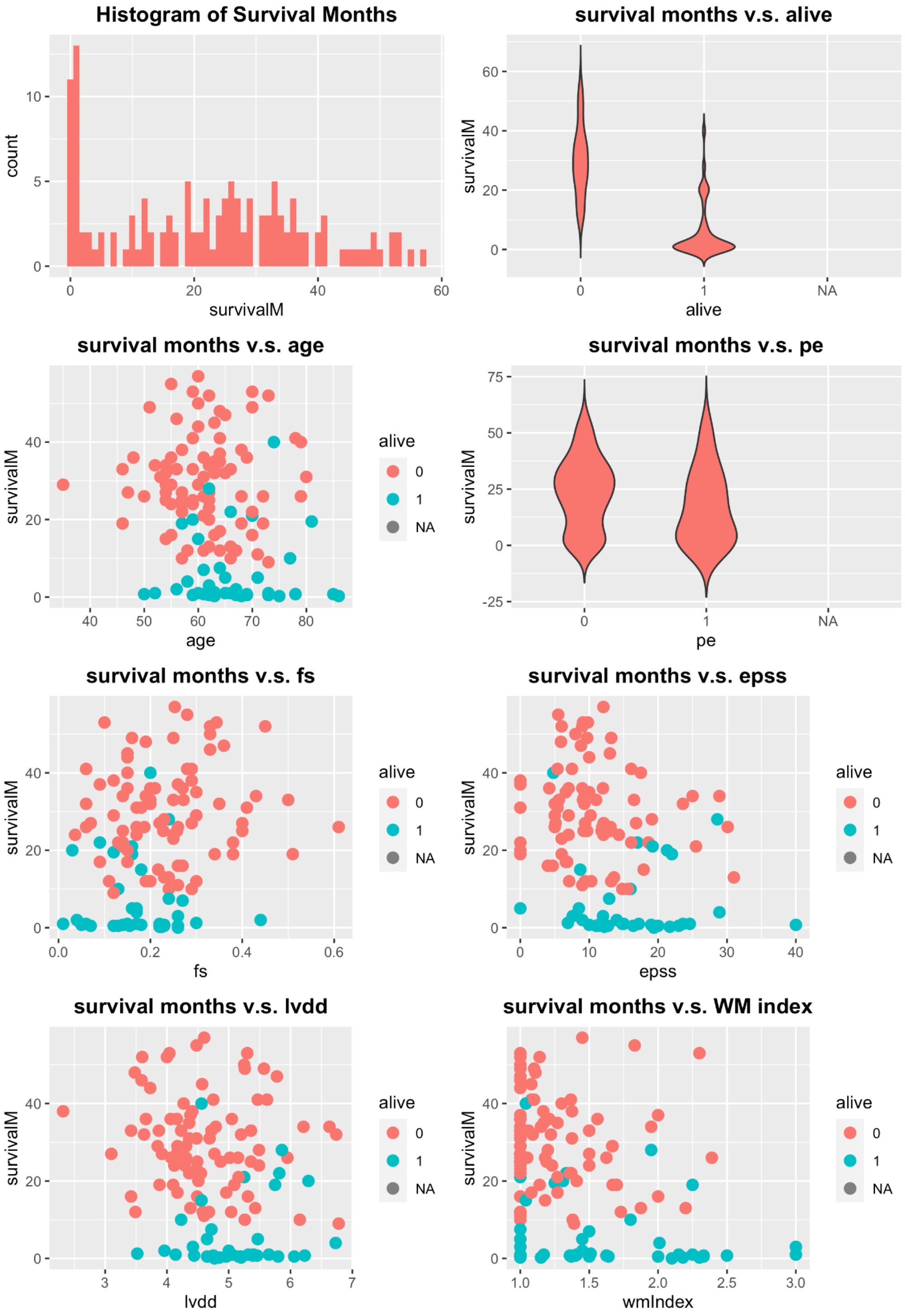
- Wall-Motion-Index (wmInd) : equals wall-motion-score divided by number of segments seen. Using a standard transthoracic echocardiography sequence, each myocardial segment is assigned a score from 1 to 4. Usually 12-13 segments are seen in an echocardiogram.

3.b Visualization

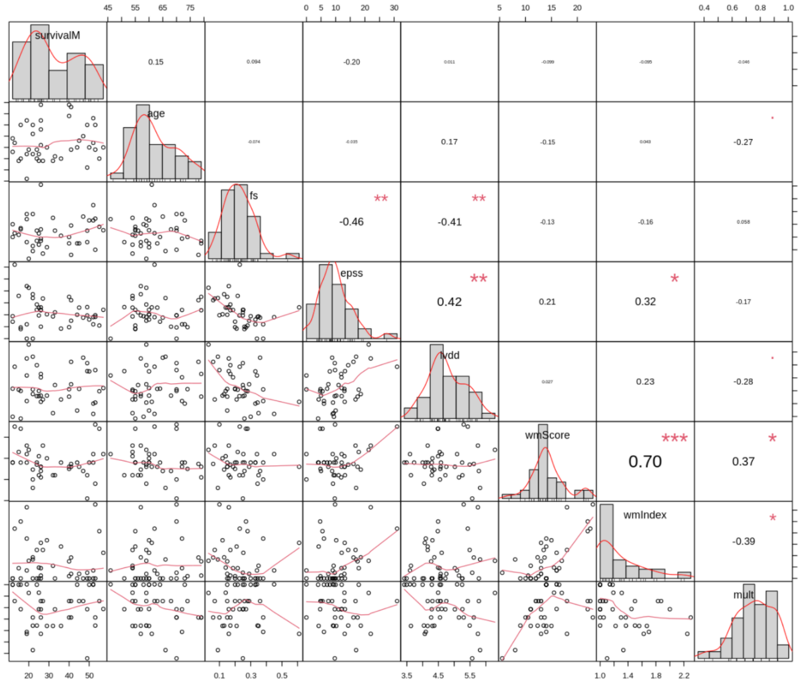
As introduced above, this dataset has one continuous response variable: Survival Month, five numeric explanatory variables: Age, Fractional-Shortening (FS), E-point septal separation(EPSS), left ventricular end-diastolic dimension(lvdd), Wall-Motion-Index (wmInd), and two dummy variables: Alive and Pericardial-Effusion (PE). All the numeric variables except wmInd are likely to be fitted by a "bell-shaped" curve(with some skewness). An unusual high frequency is noticed at low survival months, which is linked to the patients who joined the study recently and they were usually alive. The number of dead and alive patients are 88 and 43 respectively, suggesting it’s an unbalanced dataset.

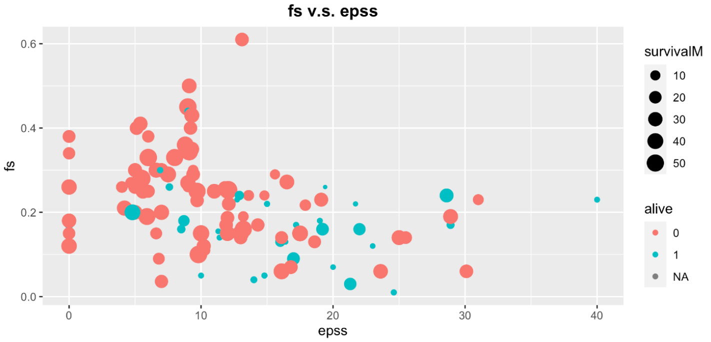


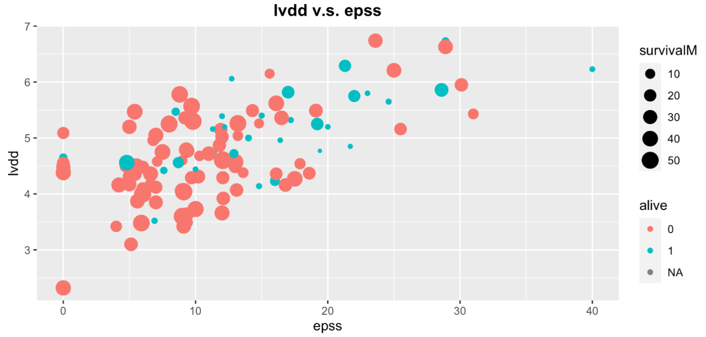
Then the relationships between a single explanatory variable and Survival month are examined. Survival month is larger on average when the patients are not alive. This is an usual case in clinical trials when the expected event(death) hasn’t been observed for newly joined testees before the trial ends. For another category variable, PE, the distribution of survival month varies a little in different groups of PE. For all the continuous variables except wmInd, barely no effects on survival month is observed. In patients recorded as dead, survival months decrease as wmInd increases.

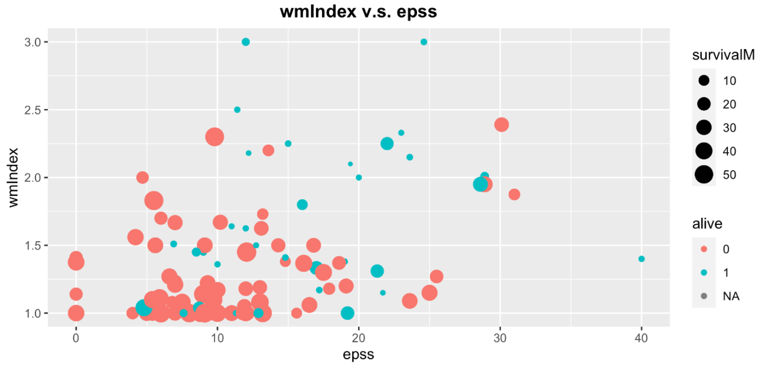


A correlation plot is generated to identify the correlations between variables: **epss & lvdd, fs & epss, and wmIndex*epss.** Further visualizations were conducted to verify correlations between these variables.

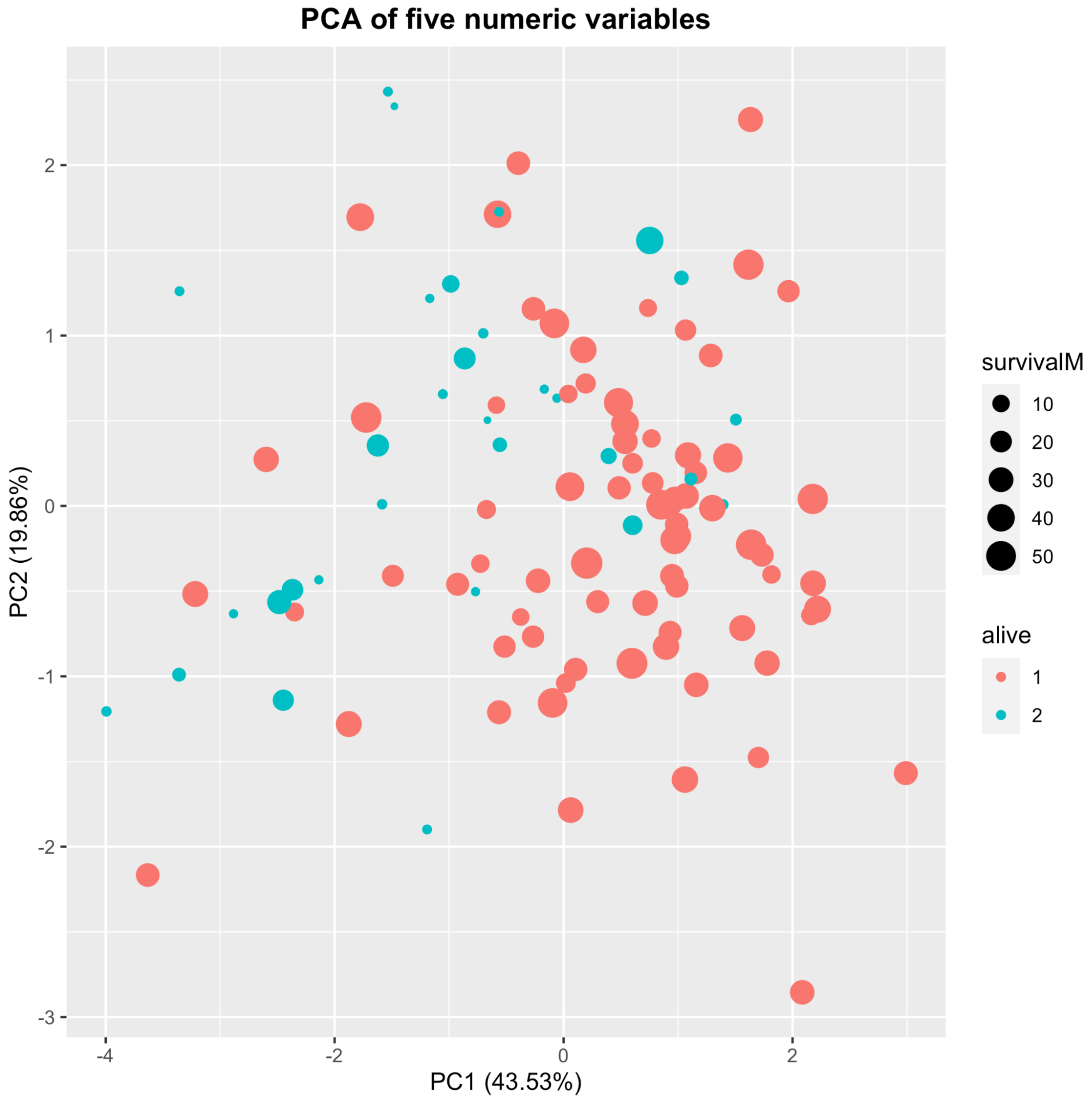








PCA is used to generalize an overview of the relationship and structure of all variables after being studied in pairs. From the analysis we found that Principle Component 1(PC1) and Principle Component 2(PC2) can only explain 43.53% and 19.86% variance in data. Thus not an ideal cluster of patients with similar survival month is shown on the plot. Survival month is reflected by the size of the points. Though an obvious separate is between the alive or not, points representing different survival months are mixed together. This also calls for some non-linear modelling techniques as we'll discuss later.

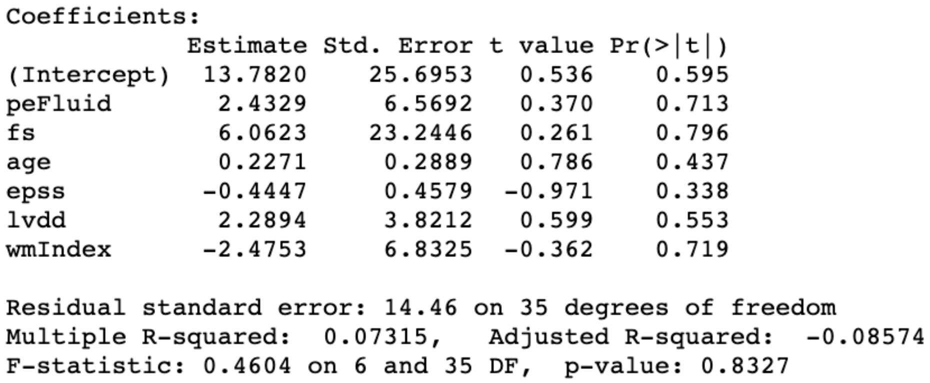


4 Data Analysis

4.a Fitting a Model

In this study, we are using echocardiography explanatory variables listed above to predict the patients' total survival months using multivariate linear regression models . Therefore, the variable ***Still-alive*** from the original dataset is excluded from our list of explanatory variables. First of all, we generated a model fitting all the variables, without considering any potential interactions:

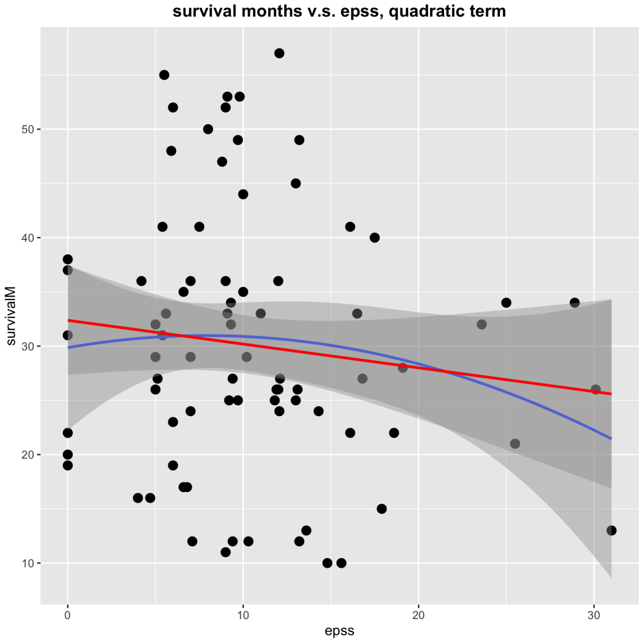
$$\textit{Survival Months}=\beta 0 + \beta 1*pe+\beta 2*fs+\beta 3*age+\beta 4*epss+\beta 5*lvdd+\beta 6*wmInd$$



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.7820    25.6953   0.536    0.595
peFluid       2.4329     6.5692   0.370    0.713
fs            6.0623    23.2446   0.261    0.796
age           0.2271     0.2889   0.786    0.437
epss         -0.4447     0.4579  -0.971    0.338
lvdd          2.2894     3.8212   0.599    0.553
wmIndex      -2.4753     6.8325  -0.362    0.719

Residual standard error: 14.46 on 35 degrees of freedom
Multiple R-squared:  0.07315,	Adjusted R-squared:  -0.08574 
F-statistic: 0.4604 on 6 and 35 DF,  p-value: 0.8327
```

The results are far from satisfaction, as all the coefficient estimates have P-Values higher than 0.05, and the adjusted r^2 is negative. Based on our correlation plot, we found comparatively stronger correlations between these pairs: **epss & lvdd, fs & epss, and wmIndex*epss.**

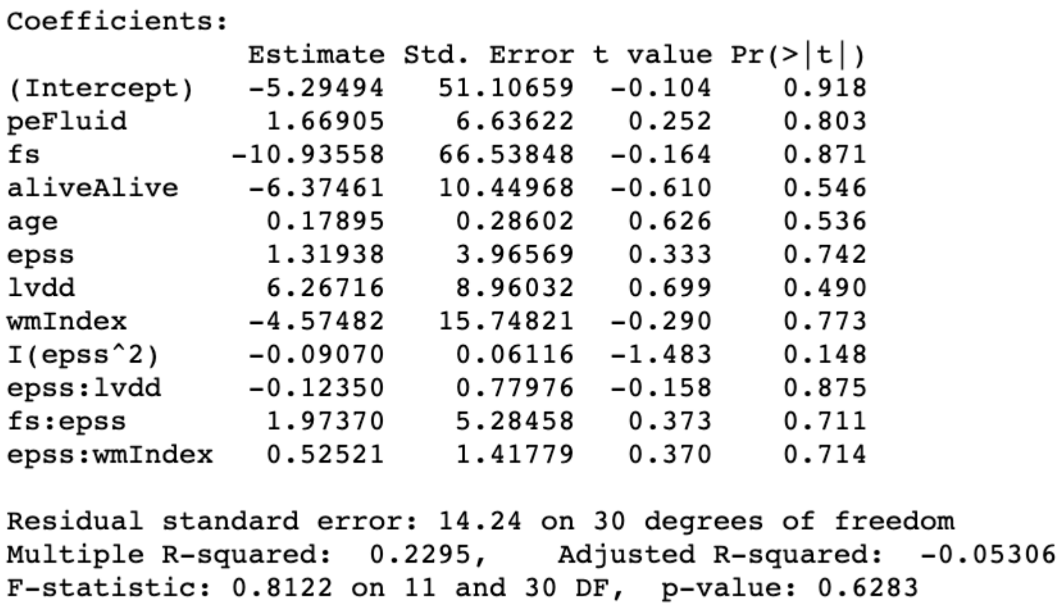
Therefore, we included the interaction terms between these variables as explanatory variables in our second model. Besides, based on our visualizations we observed that epss displays some quadratic pattern, so we also included I(epss^2) as one of the explanatory variables.


survival months v.s. epss, quadratic term

$$Surval\ Months=\beta0\ +$$

$$\beta1*pe+\beta2*fs+\beta3*age+\beta4*epss+\beta5*lvdd+\beta6*epss*lvdd+\beta7*wmIndex+\beta8*epss*lvdd+\beta9*epss*fs+\beta10*$$

$$\beta11*wmIndex*epss+\beta12*I(epss\verb|^|2)$$



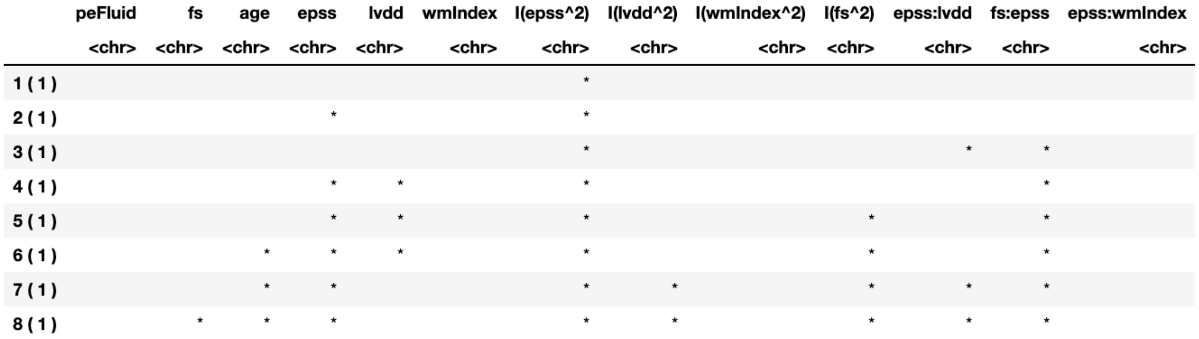
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5.29494   51.10659  -0.104    0.918
peFluid        1.66905    6.63622   0.252    0.803
fs           -10.93558   66.53848  -0.164    0.871
aliveAlive    -6.37461   10.44968  -0.610    0.546
age            0.17895    0.28602   0.626    0.536
epss           1.31938    3.96569   0.333    0.742
lvdd           6.26716    8.96032   0.699    0.490
wmIndex       -4.57482   15.74821  -0.290    0.773
I(epss^2)     -0.09070    0.06116  -1.483    0.148
epss:lvdd     -0.12350    0.77976  -0.158    0.875
fs:epss        1.97370    5.28458   0.373    0.711
epss:wmIndex   0.52521    1.41779   0.370    0.714

Residual standard error: 14.24 on 30 degrees of freedom
Multiple R-squared:  0.2295,    Adjusted R-squared:  -0.05306
F-statistic: 0.8122 on 11 and 30 DF,  p-value: 0.6283
```

Again, the results are far from satisfaction, as all the coefficient estimates have P-Values higher than 0.05, and the adjusted r^2 is still negative.

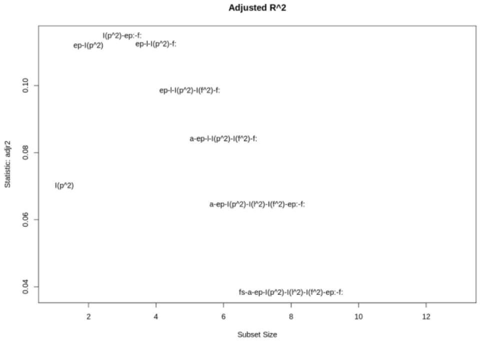
4.b Model Selection

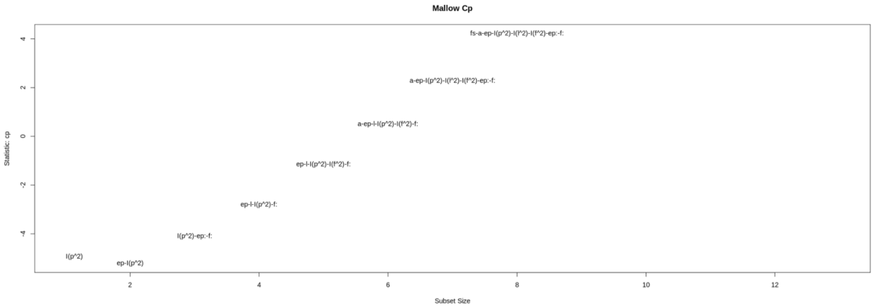
Based on the unfavorable results from the previous model fitting process, we included all possible quadratic terms in the model selection process. Here we used *regsubsets()* to choose the best combination of features.

*survivalM~pe+fs+age+epss+lvdd+epss*lvdd+wmIndex+epss*fs+wmIndex*epss+I(epss^2)+I(lvdd^2)+I(wmIndex^)+I(fs^2)*



	peFluid	fs	age	epss	lvdd	wmIndex	I(epss^2)	I(lvdd^2)	I(wmIndex^2)	I(fs^2)	epss:lvdd	fs:epss	epss:wmIndex
	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
1 (1)							*						
2 (1)				*			*						
3 (1)							*				*	*	
4 (1)				*	*		*					*	
5 (1)				*	*		*			*		*	
6 (1)			*	*	*		*			*		*	
7 (1)			*	*			*	*		*	*	*	
8 (1)		*	*	*			*	*		*	*	*	





Based on this selection result, we identified the models that have the highest adjusted r^2 the model 3, the model with the most ideal Cp is model 6, and the model with the smallest BIC value is model 1.

4.c Cross Validation

Finally, we used cross-validation to test the model's ability to predict new data. For the three models that we identified in the previous steps, the predictions errors are:

Model 3	Model 6	Model 1
242.196	327.42	178.662

Model 1 produces the smallest prediction error.

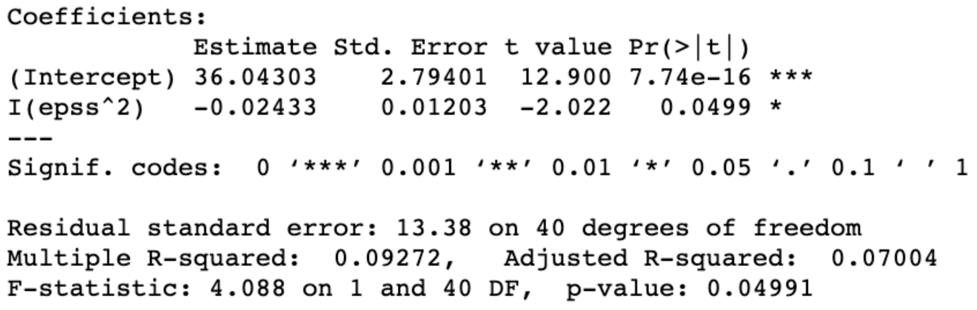
4.c.2. Leave-one-out Cross Validation

Model 3	Model 6	Mode 1
334.778	545.06	182.185

Similarly, Model 1 produces the smallest prediction error.

Eventually, we fitted the linear regression model for model 1:

$$Surval\ Months=\beta 0 + \beta 1*(epss^2)$$



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.04303    2.79401  12.900 7.74e-16 ***
I(epss^2)   -0.02433    0.01203  -2.022   0.0499 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.38 on 40 degrees of freedom
Multiple R-squared:  0.09272,   Adjusted R-squared:  0.07004
F-statistic: 4.088 on 1 and 40 DF,  p-value: 0.04991
```

We can see that although the coefficient estimates are significant, the adjusted R^2 is only 0.07, which indicates that the model is not reliable at all.

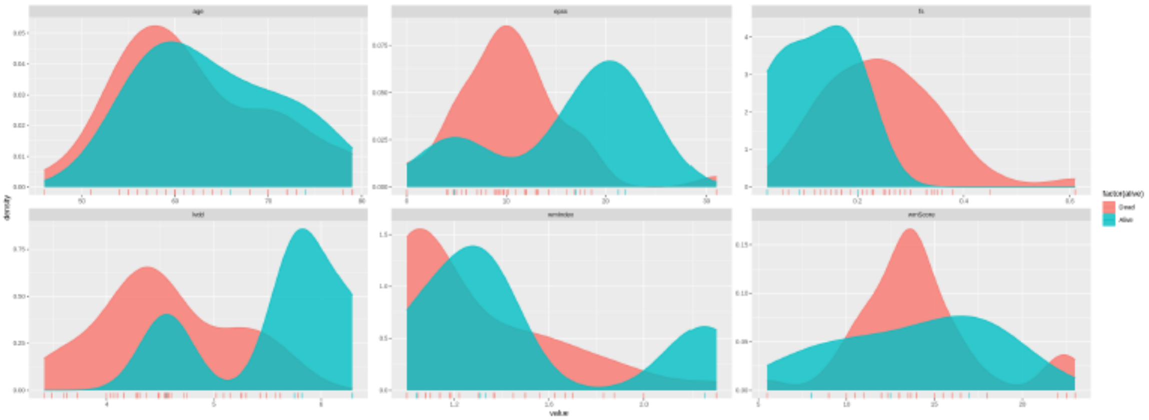
5 Conclusion and Discussion

5.a Findings

In conclusion, we found that we are unable to generate a proper multivariate linear regression model to identify the correlation between echocardiography results and the patients' survival months. Based on our model selection result, the best model selected were only able to generate an adjusted R-squared lower than 10 percent.

5.b Limitations

By reviewing our analysis process and result, we concluded the main limitations that potentially caused the failure of this study. First, the data set that we choose only has less than 200 records, with many missing values that we had to omit or replace with mean values, while we have 6 variables to test in this study. The small sample size might have caused the inaccuracy of the model generated. Additionally, based on the further EDA, it is highly possible that the relationship is not linear between the variables we are interested in, and data we have is imbalanced.



Secondly, there might be a variety of factors that can affect a patient' survival months, including gender, lifestyle (smoking, drinking, sleeping habits), other underlying health conditions and so on. The echocardiography might only make a very limited contribution to the patient' survival months.

Moreover, linear regression models might not be suitable for this study. After extensive research,we found that for similar studies, cox regression and decision trees are applied in further studies,which can generate models with much higher accuracy. Details of these models are listed in the *Future Outlook* section.

5.3. Future Outlook

In our future studies, we will first aim to collect a larger, more up-to-date data set. We will then apply cox regression and decision trees, and random forest to identify the potential correlation between echocardiography results and the patients' survival months. We still believe that such study can potentially help identify key echocardiography factors that are indicative of the patients' survival, thus better precautions can be taken in advance. Below are the models that we plan to look into in our future studies:

1. **Survival Analysis: Cox Regression**

   Censoring data, Kaplan-Meier Method, Log Rank Test and Cox Proportional Hazards Models

   Refer:

   https://www.datacamp.com/community/tutorials/survival-analysis-R

   https://www.kaggle.com/yukikitayama/survival-analysis

   This might be a suitable model for our data before it simultaneously evaluates the effect of various factors on the survival (response variable).

2. **Decision Trees**

Upon further research we see that the decision tree models can achieve upto 80% accuracy. These models can capture classes significantly. However, the imbalanced data can hinder its performance metric AUC so using Kappa is a good metric to use.

Refer: https://www.kaggle.com/loganalive/echocardiogram-dataset-uci

3. **Random Forest**

An even better model than Decision Trees is a Random Forest model. Random Forest is just a collection of several decision trees and is much more robust, limiting overfitting. These are also very effective at "estimating missing data and maintaining accuracy" when huge proportions of the data are missing. It deals efficiently with the imbalanced data by balancing errors.

Refer:

https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991

https://www.datacamp.com/community/tutorials/decision-trees-R

References:

1. Cdc. (2021, February 26). Million hearts® costs & consequences. Retrieved April 13, 2021, from https://millionhearts.hhs.gov/learn-prevent/cost-consequences.html
2. UCI machine Learning repository: ECHOCARDIOGRAM data set. (n.d.). Retrieved April 13, 2021, from https://archive.ics.uci.edu/ml/datasets/Echocardiogram
3. Cardiovascular diseases. (n.d.). Retrieved April 13, 2021, from https://www.who.int/health-topics/cardiovascular-diseases/
4. 2021 heart disease and Stroke Statistics UPDATE fact sheet ... (n.d.). Retrieved April 13, 2021, from https://www.heart.org/-/media/phd-files-2/science-news/2/2021-heart-and-stroke-stat-update/2021_heart_disease_and_stroke_statistics_update_fact_sheet_at_a_glance.pdf?la=en