University of Illinois Urbana-Champaign

Department of Statistics

Heart Failure Disease

Group Members:
Jiayi Pan
Botao Wang
Yin T. Ho

A project report submitted for the course of

STAT-448 Advanced Data Analysis

Contents

1 Introduction

Heart failure is a serious medical condition that occurs when the heart is unable to pump enough blood to meet the body's needs. It is a leading cause of hospitalization and death worldwide, and its incidence is increasing with the aging population. To better understand and manage heart failure, researchers have been collecting and analyzing clinical data from patients.

Table 1.1: Data Directory

Features	**Explanations**
Age	Age of the patient (years)
Anaemia	Decrease of red blood cells or hemoglobin
High Blood Pressure	If the patient has hypertension
Creatinine Phosphokinase (CPK)	Level of the CPK enzyme in the blood (mcg/L)
Diabetes	If the patient has diabetes
Ejection Fraction	Percentage of blood leaving the heart at each contraction (percentage)
Platelets	Platelets in the blood (kiloplatelets/mL)
Sex	Woman or Man
Serum Creatinine	Level of serum creatinine in the blood (mg/dL)
Serum sodium	Level of serum sodium in the blood (mEq/L)
smoking	If the patient smokes or not
time	Follow-up period (days)
Death Event	If the patient deceased during the follow-up period

Note: False=0; True=1. Woman=0; Man=1.

The Heart Failure Clinical Records Data Set, available on theUCI Machine Learning Repository [1], is a valuable resource for researchers and clinicians alike. It contains data on 299 patients with heart failure, including demographic information, medical history, laboratory test results, and medications, as well as the patients' survival status. This dataset has been used for various research purposes, such as predicting patient survival, and identifying risk factors for heart failure. Its availability allows for the development of more accurate and personalized approaches to managing heart failure and improving patient outcomes.

2 Preliminarily Analyses

2.1 Response Variable

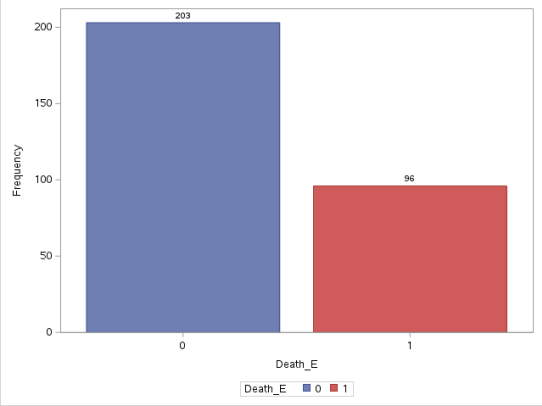


Figure 2.1: Barplot of Death Event

Table 2.1: Summary Statistics

Analysis Variable : Death_E								
N Miss	Range	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	Std Dev	Skewness
0	1.0000	0.0000	0.0000	0.0000	0.3211	1.0000	1.0000	0.4677	0.7703

Table 2.2: Normality Tests

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.588141	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.432741	Pr > D	<0.0100
Cramer-von Mises	W-Sq	11.11927	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	62.78882	Pr > A-Sq	<0.0050

[1] UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records

When people had heart failures, there were 203 patients survived and 96 patients passed away (Figure 2.1). This indicate the proportion of survival rate is relatively high, with a mean of 0.3211, a standard deviation of 0.4677, and a range of 0 to 1(Table 2.1) . Death Event is a binary variable, and all the p-values of Death Event in normality tests (Table 2.2) are smaller than 5% significance level, so it is not normally distributed.

2.2 Categorical Variables

In this study, medical risk factors are categorical variables, namely anaemia, hbp, diabetes, and smoking,

Table 2.3: Frequency Tables between Categorical Variables vs `Death_E`

Table of anaemia by Death_E

Frequency / Expected	Death_E = 0	Death_E = 1	Total
anaemia = 0	120 115.42	50 54.582	170
anaemia = 1	83 87.582	46 41.418	129
Total	203	96	299

(a) Anaemia

Table of hbp by Death_E

Frequency / Expected	Death_E = 0	Death_E = 1	Total
hbp = 0	137 131.71	57 62.288	194
hbp = 1	66 71.288	39 33.712	105
Total	203	96	299

(b) hbp

Table of diab by Death_E

Frequency / Expected	Death_E = 0	Death_E = 1	Total
diab = 0	118 118.13	56 55.866	174
diab = 1	85 84.866	40 40.134	125
Total	203	96	299

(c) Diabetes

Table of sex by Death_E

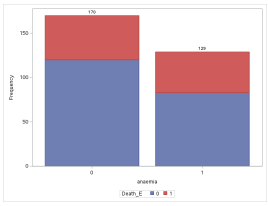
Frequency / Expected	Death_E = 0	Death_E = 1	Total
sex = 0	71 71.288	34 33.712	105
sex = 1	132 131.71	62 62.288	194
Total	203	96	299

(d) Sex

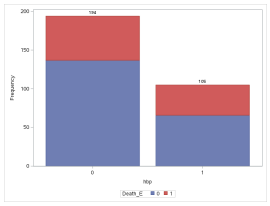
Table of smk by Death_E

Frequency / Expected	Death_E = 0	Death_E = 1	Total
smk = 0	137 137.82	66 65.177	203
smk = 1	66 65.177	30 30.823	96
Total	203	96	299

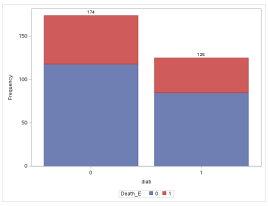
(e) Smoking



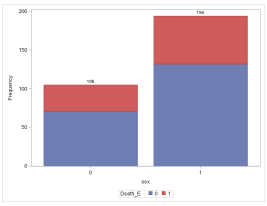
(a) Anaemia



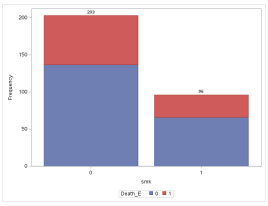
(b) hbp



(c) Diabetes



(d) Sex



(e) Smoking

Figure 2.2: Barplots of Categorical Variables

Table 2.4: Independence Tests between Categorical Variables and `Death_E`

Statistic	DF	Value	Prob
Chi-Square	1	1.3131	0.2518
Likelihood Ratio Chi-Square	1	1.3086	0.2527
Continuity Adj. Chi-Square	1	1.0422	0.3073
Mantel-Haenszel Chi-Square	1	1.3087	0.2526
Phi Coefficient		0.0663	
Contingency Coefficient		0.0661	
Cramer's V		0.0663	

(a) Anaemia

Statistic	DF	Value	Prob
Chi-Square	1	1.8827	0.1700
Likelihood Ratio Chi-Square	1	1.8630	0.1723
Continuity Adj. Chi-Square	1	1.5435	0.2141
Mantel-Haenszel Chi-Square	1	1.8764	0.1707
Phi Coefficient		0.0794	
Contingency Coefficient		0.0791	
Cramer's V		0.0794	

(b) hbp

Statistic	DF	Value	Prob
Chi-Square	1	0.0011	0.9732
Likelihood Ratio Chi-Square	1	0.0011	0.9732
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0011	0.9732
Phi Coefficient		-0.0019	
Contingency Coefficient		0.0019	
Cramer's V		-0.0019	

(c) Diabetes

Statistic	DF	Value	Prob
Chi-Square	1	0.0056	0.9405
Likelihood Ratio Chi-Square	1	0.0056	0.9405
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0056	0.9406
Phi Coefficient		-0.0043	
Contingency Coefficient		0.0043	
Cramer's V		-0.0043	

(d) Sex

Statistic	DF	Value	Prob
Chi-Square	1	0.0476	0.8272
Likelihood Ratio Chi-Square	1	0.0478	0.8270
Continuity Adj. Chi-Square	1	0.0073	0.9318
Mantel-Haenszel Chi-Square	1	0.0475	0.8275
Phi Coefficient		-0.0126	
Contingency Coefficient		0.0126	
Cramer's V		-0.0126	

(e) Smoking

These bar plots (Figure 2.2) and frequency table (Table 2.3) show relationships between medical risk factors, sex, and patient mortality, corresponding to specific Death Event within the population being studied. It found that a large proportion of the patients did not have one or more of the following medical conditions: anaemia, high blood pressure, diabetes, and smoking. However, majority of the patients tends to have higher survival rate, no matter if female or male patients had these medical conditions. Meanwhile, the total number of females is smaller than males, but both had low risk in mortality respectively. All the categorical variables in the independence tests (Table 2.4) are statistically insignificant because their p-values are greater than 5% significance level. This means each categorical variable has little association with death event.

2.3 Continuous Variables

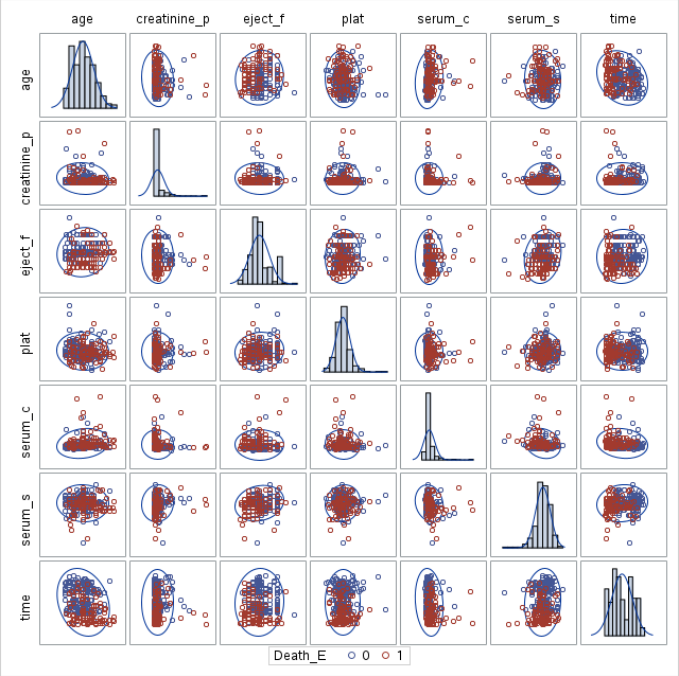


Figure 2.3: Correlation Matrix

Table 2.5: Summary Statistics

Variable	N Miss	Range	Minimum	Lower Quartile	Mean	Upper Quartile	Maximum	Std Dev	Skewness
age	0	55.0000	40.0000	51.0000	60.8339	70.0000	95.0000	11.8948	0.4231
creatinine_p	0	7838.0000	23.0000	115.0000	581.8395	582.0000	7861.0000	970.2879	4.4631
eject_f	0	66.0000	14.0000	30.0000	38.0836	45.0000	80.0000	11.8348	0.5554
plat	0	824900.0000	25100.0000	212000.0000	263358.0293	304000.0000	850000.0000	97804.2369	1.4623
serum_c	0	8.9000	0.5000	0.9000	1.3939	1.4000	9.4000	1.0345	4.4560
serum_s	0	35.0000	113.0000	134.0000	136.6254	140.0000	148.0000	4.4125	-1.0481
time	0	281.0000	4.0000	73.0000	130.2609	205.0000	285.0000	77.6142	0.1278

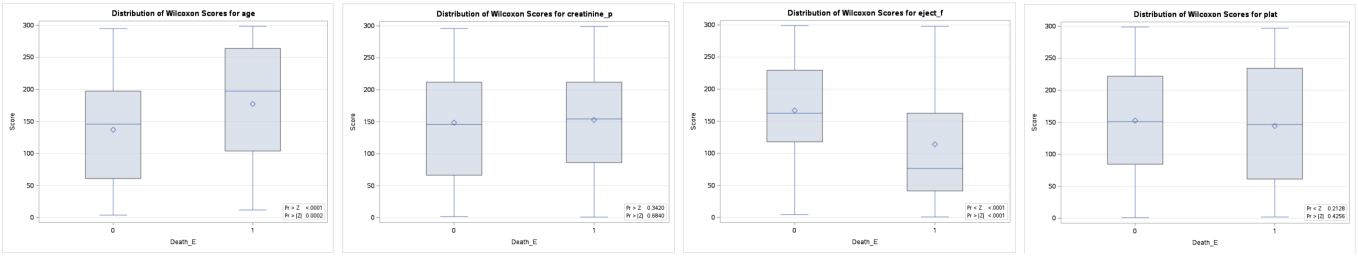
Table 2.6: Pearson Correlation

	age	creatinine_p	eject_f	plat	serum_c	serum_s	time
age	1.00000	-0.08158 0.1594	0.06010 0.3003	-0.05235 0.3670	0.15919 0.0058	-0.04597 0.4284	-0.22407 <.0001
creatinine_p	-0.08158 0.1594	1.00000	-0.04408 0.4476	0.02446 0.6735	-0.01641 0.7775	0.05955 0.3047	-0.00935 0.8722
eject_f	0.06010 0.3003	-0.04408 0.4476	1.00000	0.07218 0.2133	-0.01130 0.8457	0.17590 0.0023	0.04173 0.4722
plat	-0.05235 0.3670	0.02446 0.6735	0.07218 0.2133	1.00000	-0.04120 0.4779	0.06212 0.2843	0.01051 0.8563
serum_c	0.15919 0.0058	-0.01641 0.7775	-0.01130 0.8457	-0.04120 0.4779	1.00000	-0.18910 0.0010	-0.14932 0.0097
serum_s	-0.04597 0.4284	0.05955 0.3047	0.17590 0.0023	0.06212 0.2843	-0.18910 0.0010	1.00000	0.08764 0.1305
time	-0.22407 <.0001	-0.00935 0.8722	0.04173 0.4722	0.01051 0.8563	-0.14932 0.0097	0.08764 0.1305	1.00000

This summary statistics table (Tabel 2.5) presents key descriptive statistics for seven continuous variables that were measured in this study. The variables include age, time, ejection fraction, platelets, creatinine and serum levels. For each variable, the table (Tabel 2.5) presents the minimum and maximum values, the mean, the standard deviation, lower and upper quartiles, and skewness. Though the histograms (Figure 2.3) with density of age, ejection fraction, platelets, and time looks like normally distributed, their absolute values of skewness are greater 0.1 (Tabel 2.5). The normality tests (Table B.1) also reflect all the continuous variables are not normally distributed because their p-values are smaller than 5% significance level. The correlation matrix (Figure 2.3) show age vs time, eject fraction vs serum sodium, serum creatinine vs creatinine sodium, and serum creatinine vs time are weakly associated to each other in a group. Based on the Pearson correlation table (Table 2.6), these groups' p-values are smaller than 5% significance level, and their absolute values of Pearson correlation coefficients are between 0 and 0.3. Therefore, these groups' continuous variables are weakly correlated.

2.3.1 Location Tests

Each continuous variable in different death-survival groups is not normal distributed (Table B.2 and B.3) and has unequal variance (Table A.1 and A.1). So, the pooled t test is not appropriate in this situation, which is opposite to the Wilcoxon rank sum test. Hence, the Wilcoxon rank sum test is used to detecting location for each predictor.



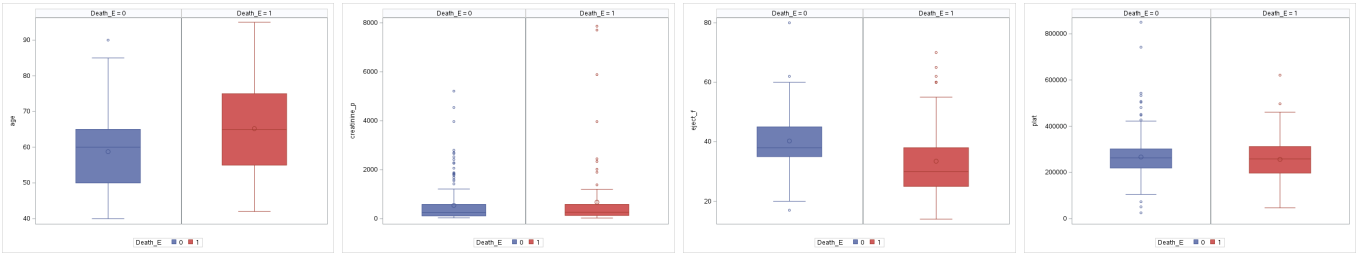
(a) Age **(b)** `creatinine_p` **(c)** `eject_f` **(d)** plat

(e) `serum_c` **(f)** `serum_s` **(g)** time

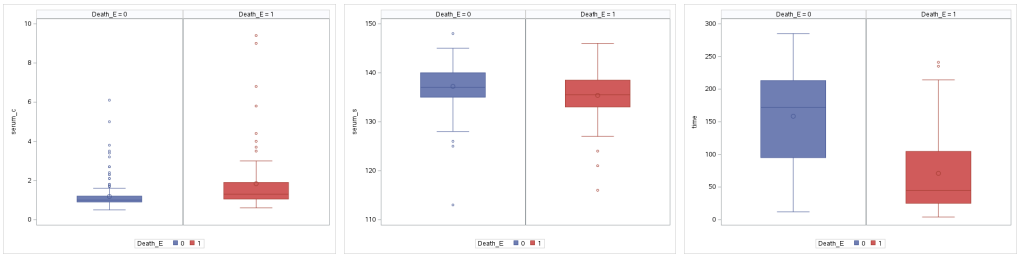
Figure 2.4: Wilcoxon Scores (Rank Sums) Tests for Continuous Variables

The box plots (Figure 2.4) of creatinine phosphokinase and platelets in different death-survival groups are overlapping, and only their p-values are greater than 5% significance level in the Wilcoxon two sample test (Table C.2). This implies that the median of age, eject fraction, serum creatinine, serum sodium, and time have statistically significant differences between dead and survival groups.

2.3.2 Associations with Death Event



(a) Age **(b)** `creatinine_p` **(c)** `eject_f` **(d)** plat



(e) `serum_c` **(f)** `serum_s` **(g)** time

Figure 2.5: Boxplots of Continuous Variables vs `Death_E`

The box plots (Figure 2.5) of age, eject fraction, serum creatinine, and time have some variation between dead and survival groups, so they may have moderate associations with Death Event. On the contrary, creatinine phosphokinase and platelets might have a weak association with Death Event.

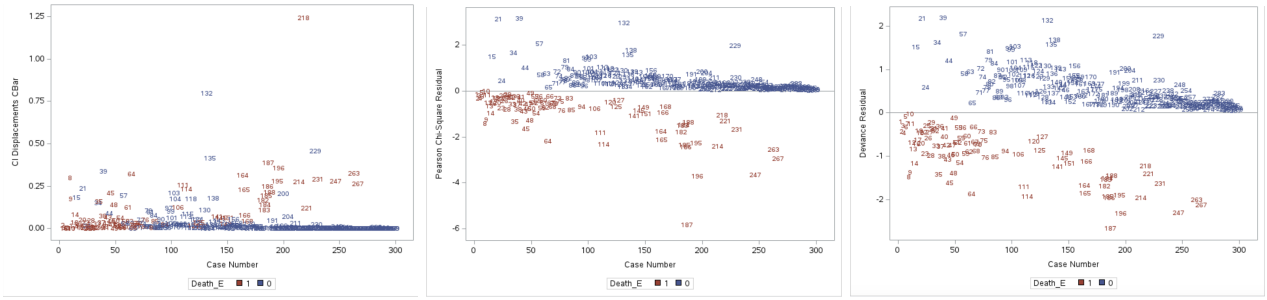
3 Logistics Regression

Logistic regression is a type of generalized linear model that is commonly used for modeling response variables that have binary or multi-class values. Its main purpose is to predict the probability of the dependent variable and establish the relationship between factors that influence the response variable. As a result, logistic regression is frequently the model of choice for solving binary classification problems.

Full Model:

$$\log\left(\frac{P(\text{Death Event})}{1-P(\text{Death Event})}\right) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Anaemia} + \beta_3 \text{CreatininePhosphokinase} + \beta_4 \text{Diabetes} + \beta_5 \text{Eject Fraction} + \beta_6 \text{SerumCreatinine} + \beta_7 \text{SerumSodium} + \beta_8 \text{Smoking} + \beta_9 \text{Time} + \beta_{10} \text{HighBloodPressure} + \beta_{11} \text{Platelets} + \beta_{12} \text{Sex} + \varepsilon$$

3.1 Residual Diagnostics



(a) CBar **(b)** Pearson Residuals **(c)** Deviance Residuals

Figure 3.1: Influence Diagnostics

After the full model has been constructed, this model is well fitted since there is no significant trend in residuals (Figure 3.1b and 3.1c). However, No. 218 observation is an influential point in the raw data because its CBar is greater than 1(Figure 3.1a), which would be be dropped in Section 3.2.

3.2 Stepwise Selection

Table 3.1: Summary of Stepwise Selection

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	375.069	234.368
SC	378.767	252.853
-2 Log L	373.069	224.368

(a)

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	148.7018	4	<.0001
Score	120.3076	4	<.0001
Wald	71.9751	4	<.0001

(b)

Summary of Stepwise Selection							
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	time		1	1	84.8630		<.0001
2	eject_f		1	2	25.2050		<.0001
3	age		1	3	12.2623		0.0005
4	serum_c		1	4	9.3502		0.0022

(c)

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
age	1	9.3081	0.0023
eject_f	1	24.0988	<.0001
serum_c	1	7.9445	0.0048
time	1	51.5764	<.0001

(d)

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.8334	1.0563	0.6225	0.4301
age	1	-0.0464	0.0152	9.3081	0.0023
eject_f	1	0.0798	0.0162	24.0988	<.0001
serum_c	1	-0.5630	0.1998	7.9445	0.0048
time	1	0.0210	0.00292	51.5764	<.0001

(e)

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	0.955	0.927	0.984
eject_f	1.083	1.049	1.118
serum_c	0.569	0.385	0.842
time	1.021	1.015	1.027

(f)

As the results of stepwise selection, the reduced model remains time, eject fraction, age, and serum creatinine (Table 3.1e). Comparing with the Intercept only model, the reduced model is better because it has smaller criterions(AIC, SC, and -2 Log L) (Table 3.1e) and extremely small p-values in the Chi-square tests (Table 3.1a). This indicates that the reduced model is more adequate than the intercept-only model.

These results (Table 3.1c) are partially consistent with the conclusions from preliminary analysis (Section 2) on the association between death-survival and categorical as well as continuous variables. The predictor, Serum Sodium, is excluded because it has high correlation with Serum Creatinine (Table 2.6) which has been dropped automatically. All the remains are statistically significant with less than 5% p-values (Table 3.1d and Table 3.1e), so this reduced model is also a final model in the stepwise selection. Additionally, the 95% wald confidence intervals of the remains do not include 1, so each remain has statistically differences between death-survival groups (Table 3.1f). Therefore, the reduced model is written as

$$\log\left(\frac{P(\text{Death Event})}{1-P(\text{Death Event})}\right) = -0.8334 - 0.0464\cdot\text{Age} + 0.0798\cdot\text{EjectFraction} - 0.5630\cdot\text{SerumCreatinine} + 0.0210\cdot\text{Time}$$

(Table 3.1e and Table 3.1f).

3.3 Model Performance

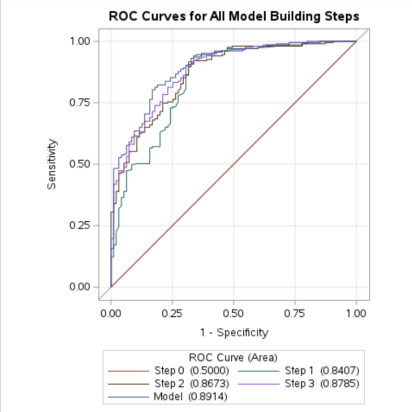


Figure 3.2: Model includes Time

Table 3.2: Confusion Matrix includes Time

Frequency	Table of Death_E by _INTO_		
	INTO(Formatted Value of the Predicted Response)		
Death_E	0	1	Total
0	183	20	203
1	30	65	95
Total	213	85	298

Table 3.3: Confusion Matrix excludes Time

Frequency	Table of Death_E by _INTO_		
	INTO(Formatted Value of the Predicted Response)		
Death_E	0	1	Total
0	183	20	203
1	49	47	96
Total	232	67	299

Table 3.4: New Parameter Estimates

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.3531	0.8396	7.8551	0.0051
age	1	-0.0517	0.0123	17.6549	<.0001
eject_f	1	0.0700	0.0142	24.1859	<.0001
serum_c	1	-0.6659	0.1592	17.5051	<.0001

Comparing AUC (Area Under Curve) values in ROC (Receiver Operating Characteristic) curves (Figure 3.2) is one of the methods to evaluate different models' performances during the stepwise process. It is observed (Figure 3.2) that the model's predictive ability improves as more predictors are added, as evidenced by the increasing AUC values from Step 1 (0.8407) to the final model (0.8914). The AUC value of 0.8914 for the final model is relatively large, indicating that the binary classifier can effectively distinguish between death-survival groups. Therefore, the performance of the final model can be regarded as good to a certain degree.

The confusion matrices were constructed to compare the observed frequencies of the death event (presence = 0 or 1) with the predicted frequencies generated by the final model using a default prediction threshold of 50%. Based on this table (Table 3.2), the overall fraction of correct predictions for the survival-death data can be calculated by

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FN} = \frac{183+65}{183+20+30+65} \approx 0.832215$$

.

Despite achieving a high level of accuracy in classifying the response variable, the final model is exhibiting signs of overfitting. The model was constructed using stepwise selection, which identified four predictors - eject fraction, age, serum creatinine, and "time" - out of twelve explanatory variables. However, including "time" in the final model raises interpretational issues since it suggests that patients are less likely to experience a negative outcome as the follow-up period increases, which is not a causal relationship. This is because "time" refers to the length of the follow-up period, and the reason why more patients with a negative outcome correspond to shorter "time" in the data is simply because they experienced the event earlier during the follow-up period, whereas the reason why more patients with a positive outcome correspond to longer "time" is just because they remained alive until the end of the follow-up period. Consequently, it is advisable to exclude "time" from the set of predictors before performing the logistic regression. Following this concept and the earlier described model building process, it is feasible to construct a new frequency table (Table 3.3).

Optimal Model:

$$\log\left(\frac{P(\text{Death Event})}{1-P(\text{Death Event})}\right) = 2.3531 - 0.0517 \cdot \text{Age} + 0.0700 \cdot \text{EjectFraction} - 0.6659 \cdot \text{SerumCreatinine}$$

The new frequency table (Table 3.3), constructed by removing the variable "time" from the logistic regression model, shows a slightly different distribution of patients compared to the original table (Table 3.2). By comparing the two contingency tables, we can see that the new sensitivity remains the same at approximately 0.9015 [2], which indicates that the model is still able to correctly identify the majority of Death Events. However, the new specificity has decreased to approximately 0.4896 [3] (Table 3.3), which is lower than the previous specificity of approximately 0.68424 [4](Table 3.2). This decrease in specificity may be attributed to the removal of the variable "time," which was significantly correlated with Death Event and may have contributed to better performance on the training data in the original model. However, the inclusion of this variable may have caused overfitting, which could have led to poorer classification performance.

$$\text{Accuracy}_{New} = \frac{TP+TN}{TP+FN+TN+FN} = \frac{183+47}{183+20+47+65} \approx 0.76923$$

Overall, while the new frequency table (Table 3.3) may result in slightly worse specificity, removing the variable "time" could potentially resolve the issue of overfitting in the final model in stepwise selection (Section 3.2)and lead to better classification performance.

4 Decision Tree

The decision tree algorithm is a widely used machine learning approach for both regression and classification problems. It is composed of branches and nodes, where each node represents a feature, and each branch represents a decision based on the feature value. The terminal nodes of the tree contain the final predictions or outcomes. The main objective of the decision tree algorithm is to divide the data recursively, based on the values of the input features, so that the resulting subsets are as pure as possible. This purity implies that the subsets contain only data points belonging to the same class or with similar target values. This recursive partitioning process is the core of the decision tree algorithm.

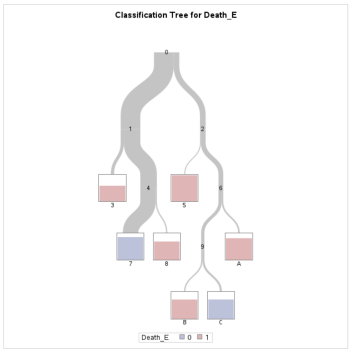


Figure 4.1: Overview Tree

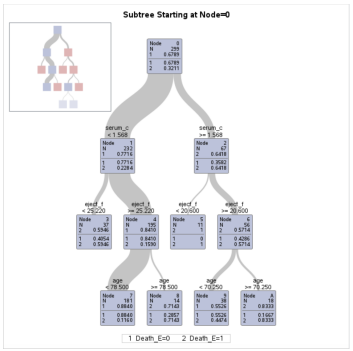


Figure 4.2: Final Tree

Table 4.1: Model Information

Split Criterion Used	Entropy
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	4
Number of Leaves Before Pruning	55
Number of Leaves After Pruning	7
Model Event Level	0

Table 4.2: Confusion Matrix

Actual	Predicted 0	Predicted 1	Error Rate
0	177	26	0.1281
1	27	69	0.2813

It has been realized that "time" is not a significant variable, although it is statistically significant in the reduced model (Section 3.3). Accordingly, "time" has been omitted in a full model while constructing a decision tree model. The decision tree was constructed using the entropy split criterion for both splitting and pruning. The final model has two branches and a depth of four, with the number of terminal nodes reduced from 55 to 7 after pruning (Table 4.1). This approach addresses the problems of overfitting and interpretation difficulties, making the model more reliable and user-friendly. The tree (Figure 4.2) leads to the similar result as the optimal model does (Section 3.3) and predicts death events in patients based on three most important predictors: serum creatinine, ejection fraction, and age. Though these findings (Figure 4.2) underscore the importance of serum creatinine levels, ejection fraction, and age in identifying patients who are at risk of mortality.

[2] $Sensitivity_{Table3.3} = \frac{TP}{TP+FN} = \frac{183}{183+20} \approx 0.9015$
[3] $Specificity_{Table3.3} = \frac{TN}{FP+TN} = \frac{47}{149+47} \approx 0.4896$
[4] $Specificity_{Table3.2} = \frac{TN}{FP+FN} = \frac{65}{30+65} \approx 0.68424$

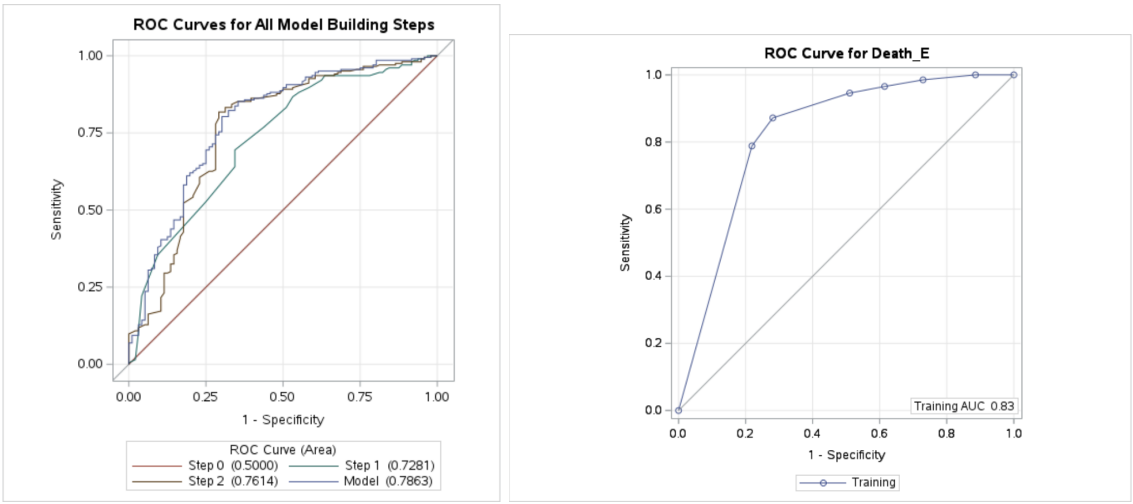
- The level of serum creatinine is the most important predictor of mortality. Patients with serum creatinine levels of 1.568 or higher are at greater risk of mortality.
- The ejection fraction is the second important predictor. Patients with an ejection fraction of less than 20.6% are at higher risk of mortality.
- Age is a significant factor in predicting mortality, but it is not as strong a predictor as serum creatinine or ejection fraction. Patients over the age of 78.5 with an ejection fraction greater than or equal to 25.22% are at higher risk of mortality.
- Patients with an ejection fraction greater than or equal to 25.22% and serum creatinine levels less than 1.568 are still at risk of mortality, but this risk is not as high as those with higher serum creatinine levels.

In summary, the proportion of accurate predictions for the survival and death outcomes can be calculated by

$$\text{Accuarcy} = \frac{TP + TN}{TP + FN + TN + FN} = \frac{177 + 69}{177 + 26 + 27 + 69} \approx 0.822742$$

(Table 4.2).

5 Conclusion



(a) Optimal Logistics Regression **(b)** Decision Tree

Figure 5.1: Model Comparison

The logistic regression and decision tree analyses revealed that age, ejection fraction, and serum creatinine levels are significant factors that impact patient mortality or survival in heart failure cases. The initial logistic regression model demonstrated high accuracy of 83.22%(Table 3.2). However, after addressing issues of causality interpretation and overfitting by removing the variable "time," the optimal model's accuracy slightly decreased to 76.92%(Table 3.3). The final decision model, using tree-based analysis, identified crucial variables for predicting patient outcomes and achieved an accuracy of 82.27% (Figure 5.1b and Table 4.2).

The findings of this study provide valuable insights for healthcare providers to make informed decisions to improve patient outcomes and provide better care. The decision tree model identifies different risk factors for death events based on serum creatinine, eject fraction, and age. The optimal logistic model equation suggests that higher age and serum creatinine levels are associated with a higher probability of death events, while higher eject fraction is associated with a lower probability of death events.

However, the constructed models have some limitations, including the small sample size of 299 patients and the lack of data on other potential factors that could impact heart failure patients' mortality, such as comorbidities, lifestyle factors, and medication use. Additionally, the optimal model's predictive power may vary when applied to a larger, more diverse patient population. Future studies with larger sample sizes and additional variables could improve the accuracy and generalizability of the models.

In conclusion, this study presents a reliable model for predicting heart failure patients' mortality risk based on age, eject fraction, and serum creatinine levels. While further research is necessary to validate the model's effectiveness in larger populations, the study's findings offer valuable insights to healthcare providers to improve patient outcomes and optimize care. The models' limitations highlight the need for additional variables and larger sample sizes to improve model accuracy and generalizability.

6 Appendices

A Descriptive Statistics

Table A.1: Basic Statistics for Continuous Variables when `Death_E`=0

Location		Variability	
Mean	58.76191	Std Deviation	10.63789
Median	60.00000	Variance	113.16471
Mode	60.00000	Range	50.00000
		Interquartile Range	15.00000

(a) Age

Location		Variability	
Mean	540.0542	Std Deviation	753.79957
Median	245.0000	Variance	568214
Mode	582.0000	Range	5179
		Interquartile Range	473.00000

(b) `creatinine_p`

Location		Variability	
Mean	40.26601	Std Deviation	10.85996
Median	38.00000	Variance	117.93879
Mode	35.00000	Range	63.00000
		Interquartile Range	10.00000

(c) `eject_f`

Location		Variability	
Mean	266657.5	Std Deviation	97531
Median	263000.0	Variance	9512335419
Mode	263358.0	Range	824900
		Interquartile Range	83000

(d) plat

Location		Variability	
Mean	1.184877	Std Deviation	0.65408
Median	1.000000	Variance	0.42782
Mode	1.000000	Range	5.60000
		Interquartile Range	0.30000

(e) `serum_c`

Location		Variability	
Mean	137.2167	Std Deviation	3.98292
Median	137.0000	Variance	15.86368
Mode	137.0000	Range	35.00000
		Interquartile Range	5.00000

(f) `serum_s`

Location		Variability	
Mean	158.3399	Std Deviation	67.74287
Median	172.0000	Variance	4589
Mode	187.0000	Range	273.00000
		Interquartile Range	118.00000

(g) time

Table A.2: Basic Statistics for Continuous Variables when `Death_E`=1

Location		Variability	
Mean	65.21528	Std Deviation	13.21456
Median	65.00000	Variance	174.62448
Mode	60.00000	Range	53.00000
		Interquartile Range	20.00000

(a) Age

Location		Variability	
Mean	670.1979	Std Deviation	1317
Median	259.0000	Variance	1733385
Mode	582.0000	Range	7838
		Interquartile Range	453.50000

(b) `creatinine_p`

Location		Variability	
Mean	33.46875	Std Deviation	12.52530
Median	30.00000	Variance	156.88322
Mode	25.00000	Range	56.00000
		Interquartile Range	13.00000

(c) `eject_f`

Location		Variability	
Mean	256381.0	Std Deviation	98526
Median	258500.0	Variance	9707310182
Mode	263358.0	Range	574000
		Interquartile Range	115000

(d) plat

Location		Variability	
Mean	1.835833	Std Deviation	1.46856
Median	1.300000	Variance	2.15667
Mode	1.000000	Range	8.80000
		Interquartile Range	0.85000

(e) `serum_c`

Location		Variability	
Mean	135.3750	Std Deviation	5.00158
Median	135.5000	Variance	25.01579
Mode	134.0000	Range	30.00000
		Interquartile Range	5.50000

(f) `serum_s`

Location		Variability	
Mean	158.3399	Std Deviation	67.74287
Median	172.0000	Variance	4589
Mode	187.0000	Range	273.00000
		Interquartile Range	118.00000

(g) time

B Normality Tests

Table B.1: Normality Tests for Continuous Variables

Test	Statistic		p Value	
Shapiro-Wilk	W	0.97547	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.069751	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.235874	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.642448	Pr > A-Sq	<0.0050

(a) Age

Test	Statistic		p Value	
Shapiro-Wilk	W	0.514263	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.286765	Pr > D	<0.0100
Cramer-von Mises	W-Sq	8.096309	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	41.90613	Pr > A-Sq	<0.0050

(b) `creatinine_p`

Test	Statistic		p Value	
Shapiro-Wilk	W	0.947316	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.168123	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.944623	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	5.802017	Pr > A-Sq	<0.0050

(c) `eject_f`

Test	Statistic		p Value	
Shapiro-Wilk	W	0.911509	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.116068	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.924776	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	4.989043	Pr > A-Sq	<0.0050

(d) plat

Test	Statistic		p Value	
Shapiro-Wilk	W	0.551466	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.265251	Pr > D	<0.0100
Cramer-von Mises	W-Sq	6.935724	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	36.45086	Pr > A-Sq	<0.0050

(e) `serum_c`

Test	Statistic		p Value	
Shapiro-Wilk	W	0.939028	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.11254	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.524928	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.093756	Pr > A-Sq	<0.0050

(f) `serum_s`

Test	Statistic		p Value	
Shapiro-Wilk	W	0.946783	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.104807	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.829905	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	4.970228	Pr > A-Sq	<0.0050

(g) time

Table B.2: Normality Tests for Continuous Variables when `Death_E`=0

Test	Statistic		p Value	
Shapiro-Wilk	W	0.979641	Pr < W	0.0048
Kolmogorov-Smirnov	D	0.066764	Pr > D	0.0252
Cramer-von Mises	W-Sq	0.142093	Pr > W-Sq	0.0313
Anderson-Darling	A-Sq	0.962072	Pr > A-Sq	0.0164

(a) Age

Test	Statistic		p Value	
Shapiro-Wilk	W	0.627714	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.2516	Pr > D	<0.0100
Cramer-von Mises	W-Sq	4.189999	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	22.49463	Pr > A-Sq	<0.0050

(b) `creatinine_p`

Test	Statistic		p Value	
Shapiro-Wilk	W	0.91993	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.209278	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.102481	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	6.442492	Pr > A-Sq	<0.0050

(c) `eject_f`

Test	Statistic		p Value	
Shapiro-Wilk	W	0.87304	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.135289	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.899987	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	4.968999	Pr > A-Sq	<0.0050

(d) plat

Test	Statistic		p Value	
Shapiro-Wilk	W	0.588627	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.262657	Pr > D	<0.0100
Cramer-von Mises	W-Sq	4.073281	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	21.97445	Pr > A-Sq	<0.0050

(e) `serum_c`

Test	Statistic		p Value	
Shapiro-Wilk	W	0.924775	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.128764	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.412171	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.366785	Pr > A-Sq	<0.0050

(f) `serum_s`

Test	Statistic		p Value	
Shapiro-Wilk	W	0.942217	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.132897	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.873446	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	4.805655	Pr > A-Sq	<0.0050

(g) time

Table B.3: Normality Tests for Continuous Variables when `Death_E`=1

Test	Statistic		p Value	
Shapiro-Wilk	W	0.968888	Pr < W	0.0221
Kolmogorov-Smirnov	D	0.111787	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.114663	Pr > W-Sq	0.0744
Anderson-Darling	A-Sq	0.760307	Pr > A-Sq	0.0470

(a) Age

Test	Statistic		p Value	
Shapiro-Wilk	W	0.439243	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.358224	Pr > D	<0.0100
Cramer-von Mises	W-Sq	3.561376	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	18.1682	Pr > A-Sq	<0.0050

(b) `creatinine_p`

Test	Statistic		p Value	
Shapiro-Wilk	W	0.926549	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.146355	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.345665	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.237556	Pr > A-Sq	<0.0050

(c) `eject_f`

Test	Statistic		p Value	
Shapiro-Wilk	W	0.971356	Pr < W	0.0336
Kolmogorov-Smirnov	D	0.108582	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.137976	Pr > W-Sq	0.0356
Anderson-Darling	A-Sq	0.724192	Pr > A-Sq	0.0591

(d) plat

Test	Statistic		p Value	
Shapiro-Wilk	W	0.608422	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.253408	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.959628	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	10.80958	Pr > A-Sq	<0.0050

(e) `serum_c`

Test	Statistic		p Value	
Shapiro-Wilk	W	0.958213	Pr < W	0.0038
Kolmogorov-Smirnov	D	0.131275	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.162565	Pr > W-Sq	0.0173
Anderson-Darling	A-Sq	0.946987	Pr > A-Sq	0.0174

(f) `serum_s`

Test	Statistic		p Value	
Shapiro-Wilk	W	0.862312	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.171337	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.7836	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	4.61807	Pr > A-Sq	<0.0050

(g) time

C Location Tests

C.1 Wilcoxon Scores (Rank Sums) Tests

Table C.1: Summary Statistics for Continuous Variables Classified by `Death_E`

Death_E	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	203	27827.0	30450.0	696.603384	137.078818
1	96	17023.0	14400.0	696.603384	177.322917

Average scores were used for ties.

(a) Age

Death_E	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	203	30166.0	30450.0	696.634859	148.600985
1	96	14684.0	14400.0	696.634859	152.958333

Average scores were used for ties.

(b) `creatinine_p`

Death_E	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	203	33882.50	30450.0	693.137669	166.908867
1	96	10967.50	14400.0	693.137669	114.244792

Average scores were used for ties.

(c) `eject_f`

Death_E	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	203	31006.50	30450.0	697.778070	152.741379
1	96	13843.50	14400.0	697.778070	144.203125

Average scores were used for ties.

(d) plat

Death_E	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	203	26004.0	30450.0	694.897434	128.098522
1	96	18846.0	14400.0	694.897434	196.312500

Average scores were used for ties.

(e) `serum_c`

Death_E	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	203	32967.50	30450.0	694.991458	162.401478
1	96	11882.50	14400.0	694.991458	123.776042

Average scores were used for ties.

(f) `serum_s`

Death_E	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	203	36994.50	30450.0	697.952090	182.238916
1	96	7855.50	14400.0	697.952090	81.828125

Average scores were used for ties.

(g) time

Table C.2: Two Sample Tests for Continuous Variables Classified by `Death_E`

Statistic	Z	Pr > Z	Pr > |Z|	t Approximation Pr > Z	t Approximation Pr > |Z|
17023.00	3.7647	<.0001	0.0002	0.0001	0.0002

Z includes a continuity correction of 0.5.

(a) Age

Statistic	Z	Pr > Z	Pr > |Z|	t Approximation Pr > Z	t Approximation Pr > |Z|
14684.00	0.4070	0.3420	0.6840	0.3422	0.6843

Z includes a continuity correction of 0.5.

(b) `creatinine_p`

Statistic	Z	Pr < Z	Pr > |Z|	t Approximation Pr < Z	t Approximation Pr > |Z|
10967.50	-4.9514	<.0001	<.0001	<.0001	<.0001

Z includes a continuity correction of 0.5.

(c) `eject_f`

Statistic	Z	Pr < Z	Pr > |Z|	t Approximation Pr < Z	t Approximation Pr > |Z|
13843.50	-0.7968	0.2128	0.4256	0.2131	0.4262

Z includes a continuity correction of 0.5.

(d) plat

Statistic	Z	Pr > Z	Pr > |Z|	t Approximation Pr > Z	t Approximation Pr > |Z|
18846.00	6.3973	<.0001	<.0001	<.0001	<.0001

Z includes a continuity correction of 0.5.

(e) `serum_c`

Statistic	Z	Pr < Z	Pr > |Z|	t Approximation Pr < Z	t Approximation Pr > |Z|
11882.50	-3.6216	0.0001	0.0003	0.0002	0.0003

Z includes a continuity correction of 0.5.

(f) `serum_s`

Statistic	Z	Pr < Z	Pr > |Z|	t Approximation Pr < Z	t Approximation Pr > |Z|
7855.500	-9.3760	<.0001	<.0001	<.0001	<.0001

Z includes a continuity correction of 0.5.

(g) time