Identifying Predictors of Survival Due to Heart Failure

Math 449 Final Project

Nicholas Kane

May 2025

1 Introduction

Cardiovascular disease is one of the leading causes of death worldwide. A large subset of these cases is heart failure. Heart failure occurs when the heart can no longer pump blood efficiently enough to keep up with the needs of the body. In this project, factors related to heart failure are studied with patient survival as response variable. Through logistic regression and model fitting it was found that the most significant factors were age, serum creatinine, ejection fraction, and time between follow up visits.

2 Data Set

The data set used in this project is of heart failure clinical records collected from the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan) during April-December 2015. The data set contains a row for each patient and measures 13 variables. The variables are age (integer), anemia (binary), creatinine phosphokinase level (integer), diabetes (binary), ejection fraction (integer), high blood pressure (binary), platelets (continuous), serum creatinine (continuous), serum sodium (integer), sex (binary), smoking (binary), and time between follow up (integer), and death (binary). Death will be the binary response variable, with 1 being death and 0 being survival. There were no missing values in the data set, so all 12 predictors will be used in the initial fit.

3 Model Selection

3.1 Predictor Selection

When attempting to fit all predictors, we obtain the coefficients in Table 1. From the summary, we can see that age, serum creatinine, ejection fraction, and time are the only significant predictors. We can confirm this my performing backwards step wise selection. This model has all four predictors that were significant in the first model, however it also includes serum sodium. We can use the drop1 function and see that serum sodium can be dropped without a significant effect on the model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	10.1849	5.6566	1.80	0.0718
age	0.0474	0.0158	3.00	0.0027
anaemia	-0.0075	0.3605	-0.02	0.9835
creatinine_phosphokinase	0.0002	0.0002	1.25	0.2117
diabetes	0.1451	0.3512	0.41	0.6794
ejection_fraction	-0.0767	0.0163	-4.69	0.0000
high_blood_pressure	-0.1027	0.3587	-0.29	0.7747
platelets	-0.0000	0.0000	-0.64	0.5254
serum_creatinine	0.6661	0.1815	3.67	0.0002
serum_sodium	-0.0670	0.0397	-1.69	0.0919
sex	-0.5337	0.4139	-1.29	0.1973
smoking	-0.0135	0.4126	-0.03	0.9739
time	-0.0210	0.0030	-6.98	0.0000

Table 1: Coefficients for the full model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.6045	1.0361	0.58	0.5596
age	0.0433	0.0149	2.91	0.0036
serum_creatinine	0.7198	0.1746	4.12	0.0000
ejection_fraction	-0.0748	0.0156	-4.81	0.0000
time	-0.0206	0.0029	-7.15	0.0000

Table 2: Final Model Coefficients

When fitting the model with age, serum creatinine, ejection fraction, and time, we can see that using the drop1 function that all four predictors are significant this prediction equation can be found below (A=age, SC = serum creatinine, EF = ejection fraction, T = time). These coefficients can be found in Table 2.

$$P(\hat{Y}=1) = \frac{e^{0.6045+0.0433A+0.7198SC-0.0748EF-0.0206T}}{1+e^{0.6045+0.0433A+0.7198SC-0.0748EF-0.0206T}}$$

Our model can now be compared to the full model to test if the models are significantly different, this is shown in Table 3. When we compare to the full, we get a p value of 0.56, and thus reject the null and conclude that our model is not significantly different. We can also compare to the null model which has similar results in Table 4. Since the p value is ≈ 0 we determine our model is better than the null.

We can also show that the model is a good, but not great fit since the majority of residuals are less than |3|. These residuals can be seen in Figure 1.

Resid. Df	Resid. Dev	Df	Deviance	P-val
294	226.3	NA	NA	NA
286	219.55	8	6.747	0.56

Table 3: LRT comparing final model to full model

Resid. Df	Resid. Dev	Df	Deviance	P-val
298	375.35	NA	NA	NA
294	226.3	4	149.05	2.2e-16

Table 4: LRT comparing final model to null model

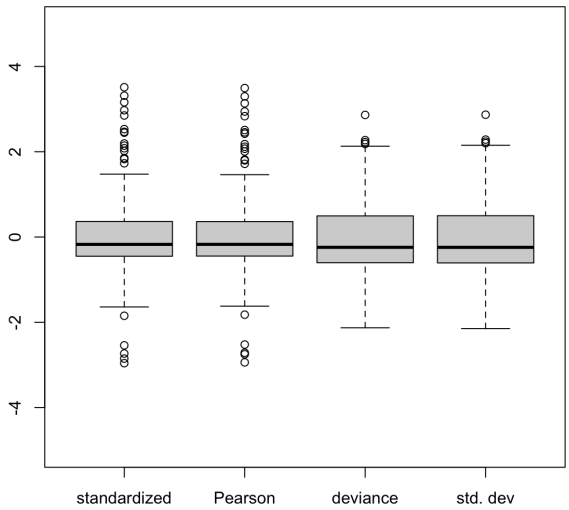

Figure 1: Residuals for the final model fit

3.2 Inference

We can infer the multiplicative effect of our predictors by finding the 95% confidence interval for β and e^{β}. These results are summarized in Table 5. From the table we can observe that serum creatinine has the largest multiplicative effect, multiplying the odds by 2.052. Next is age, which, although > 1, has a more muted multiplicative effect of 1.044, though an increase in age will still increase the odds of a death. Both time and ejection fraction have a negative effect, as increasing the time decreases odds by 0.979, and increasing the ejection fraction decreases odds by 0.928. The latter makes sense as the higher the ejection fraction, the more efficient the heart, and thus the smaller chance of heart failure.

	β	Lower β	Upper β	e^{β}	Lower e^{β}	Upper e^{β}
age	0.043	0.015	0.073	1.044	1.01	1.07
sodium creatinine	0.719	0.387	1.096	2.052	1.473	2.992
ejection fraction	-0.075	-1.06	-0.04	0.928	0.346	0.96
time	-0.021	-0.027	-0.015	0.979	0.973	0.985

Table 5: Coefficient and multiplicative effect for 95% confidence level

3.2.1 Model Visualization

We can visualize the effect of each predictor by plotting their effect. These logistic curves are shown in Figure 2, and from them we can observe the same inferences as made previously. An increase in time and ejection fraction decrease the probability of a death event, and an increase in serum creatinine and age both increase the probability of a death event. Also included in the plots are the margins of error in gray around each curve.

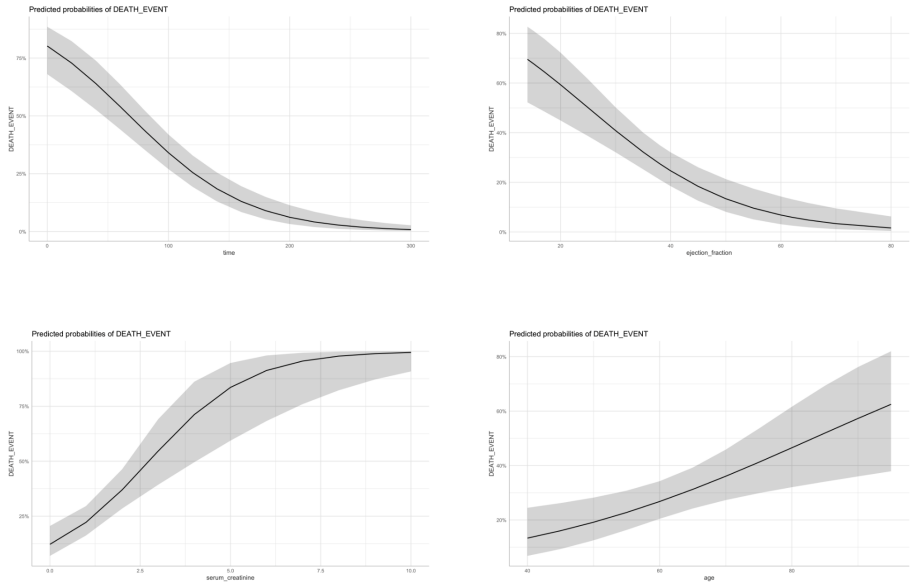

Figure 2: Visualizations of Predictor Effects

3.3 Cut-Off Selection

When $\pi_0 = 0.5$, the model has a sensitivity of 0.902, specificity of 0.688, and accuracy of 0.833. We also can compare these values for a few different cutoff points (0.25, 0.3, 0.4, 0.6, 0.75, and 0.85), and the results can be found in Table 6. We can observe that the optimal cutoff point is somewhere between 0.3 and 0.4, and from the ROC curve we confirm this, and find an optimal cutoff point of 0.312. Although this does not maximize accuracy, it maximizes the tradeoff between sensitivity and specificity (these results can be observed in Table 7 which compares the logit and probit models). The ROC curve also gives us an AUC of 0.891, which further proves we have a good fit for the data.

Cutoff (π_0)	sensitivity	specificity	accuracy
0.25	0.739	0.843	0.773
0.3	0.798	0.833	0.809
0.4	0.872	0.75	0.833
0.5	0.902	0.688	0.833
0.6	0.936	0.635	0.840
0.75	0.975	0.375	0.783
0.85	0.903	0.292	0.759

Table 6: Comparing various values of cutoff point π_0

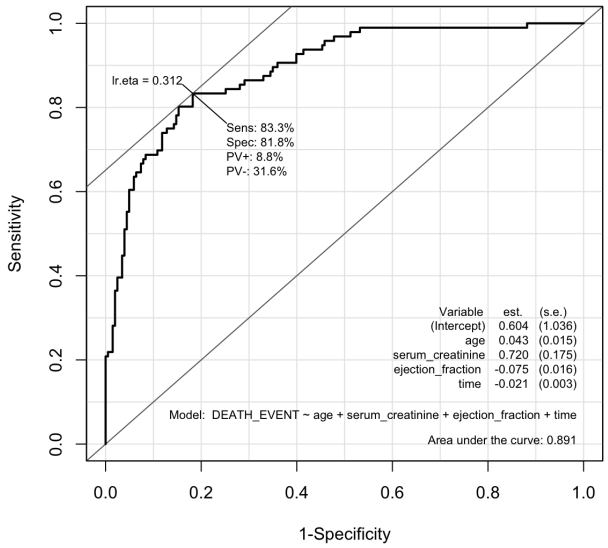

Figure 3: Receiver Operating Characteristic for Final Logit Model

3.4 Cross Validation

Leave-one-out and 10-fold cross-validation are both performed to estimate the model accuracy on the full dataset with a cutoff value of $\pi_0 = 0.5$. Using LOOCV, we obtain an accuracy of 87%, using 10-fold cross validation, we obtain an accuracy of 70.3%. The lower accuracy from 10-fold CV can be explained by an imbalance of the data in the set. Since about 1/3 of the cases are deaths and 2/3 are survivals, when splitting up into groups of 10, we may encounter more discrepancies in the model.

3.5 Link Selection

In addition to the logistic model, we can also test probit and identity models. The probit model has the same value for specificity, but has lower sensitivity and accuracy values, so we conclude that the logit model is better.

Model	sensitivity	specificity	accuracy	AIC
logit	0.813	0.833	0.819	236.3
probit	0.793	0.833	0.806	236.4

Table 7: Comparing the logit and probit models

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.389	0.594	0.657	0.511
age	0.0247	0.0083	2.958	0.0031
serum_creatinine	0.4065	0.0994	4.089	0.0000
ejection_fraction	-0.0439	0.0087	-5.081	0.0000
time	-0.0116	0.0015	-7.676	0.0000

Table 8: Probit Model Coefficients

4 Conclusion

In this project, predictors of survival after heart failure were analyzed using logistic regression. It was found that significant predictors were age, serum creatinine, ejection fraction, and time between visits, with serum creatinine having the most significant effect. This model classifies a death event reasonably well, with an accuracy of 81.0% when internally analyzed, and 70.3% when compared to the validation set. A possible improvement on this project would be to better balance the data by obtaining multiple models by sampling from the survivors. This would most likely increase our cutoff point, and give a higher accuracy and AUC.