College of Arts and Sciences
Dr. Bing Zhang Department of Statistics

University of Kentucky

College of Arts and Sciences
Dr. Bing Zhang Department of Statistics

Dr. Bing Zhang Department of Statistics

Improving the prediction of the Heart Failure Patients' Survival using Machine Learning Approach

Author:
Pronob Barman

July 30, 2021

1 30-Day Mortality Data Analysis

Introduction

We are interested in analyzing the dataset of 30 days mortality. This dataset contains sample size $n = 294$, which excludes the censored data. We totally have 13 variables, but we manually create another variable Y as our response variable for the data analysis purpose. Below are the variables that used in the D and G (2020):

- Y: binary, response variable. If `time` ≤ 30 and `DEATH EVENT` $= 1$, then Y=1. If `time` > 30, then Y=0.
- X_1: serum creatinine, level of serum creatinine in the blood (mg/dL).
- X_2: ejection fraction, percentage of blood leaving the heart at each contraction (percentage).

We are going to use the logistic regression to perform the data analysis.

Data Analysis

We first perform some of the explanatory data analysis. We plot the scatter plot of X_1 *and* X_2 in Figure 1.

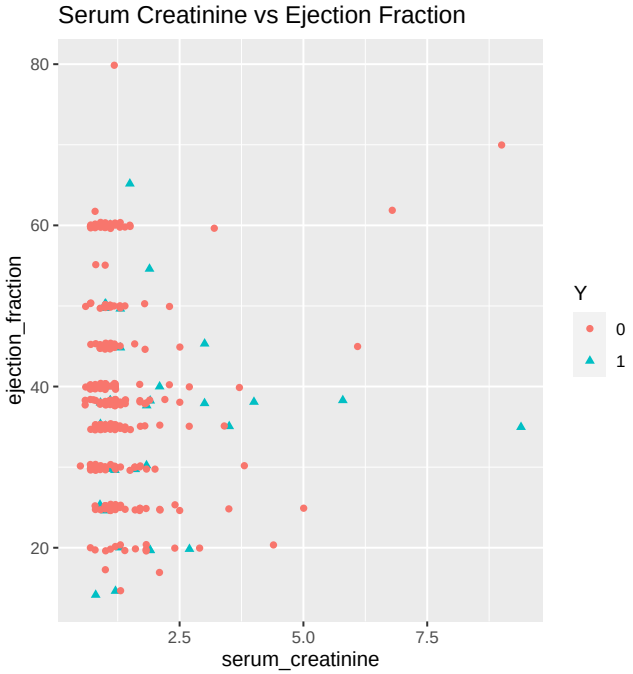

Figure 1: Serum Creatinine vs Ejection Fraction

Next, we provide some the descriptive statistics for both X_1 *and* X_2 and their distribution as well. The distributions are provided in Figure 2. We can find that for X_1, the distribution is really right skewed while X_2 somewhat look like a normal although not perfectly. The summary table is shown in Table 1.

We fit the logistic regression model next and provide the summary output below. The estimates and p-values can be found in Table 2. The mathematical form of hypotheses testing are performed in Equation 1:

$$\begin{aligned} H_0 &: \beta_i = 0 \\ H_1 &: \beta_i \neq 0, \quad i = 1, 2. \end{aligned} \tag{1}$$

In the clinical context, we are testing whether the effect of serum creatinine is contributing in our model. Same argument applies for the other covariates. Furthermore, we can calculate the 95% confidence intervals for the estimates. The method that we used is the profiling confidence interval, which is shown in Table 2.

	X_1 (Y=1)	X_1 (Y=0)	X_2 (Y=1)	X_2 (Y=0)
Min	.8	.5	14	15
Q1	1	.9	30	30
Median	1.3	1.1	38	38
Mean	1.945	1.321	36.37	38.37
Q3	1.9	1.3	42.5	45
Max	9.4	9	65	80

Table 1: Descriptive Statistics

	Estimate	Std. Error	z value	p-value	95% CI Lower limit	95% CI Upper limit
(Intercept)	-1.94410	0.62543	-3.108	0.00188	-3.2019	-0.7330
X_1	0.37984	0.13010	2.920	0.0035	0.1235	.6469
X_2	-0.01741	0.01599	-1.088	0.27642	-.0498	.0132

Table 2: Test result(Model 1)

Now, we already built a fitted logistic model and have parameter estimates, the estimated odds ratio can be calculated. Instead of one unit changed, we are interested in a ten unit changed in X_1, when X_2 is fixed, which is about 44.63177. Also, the 95% confidence interval can be obtained to be (3.4388, 600.441). Again, this is the profile confidence interval.

We next, then, calculate the $P(Y = 1|X_1 = x_1, X_2 = x_2)$ for the nine combinations of x_1 *and* x_2 based on their first, second, and third quartile. Table 3 shows the 3×3 table of numerical score(or, estimated probability) for each combination of serum creatinine and ejection fraction.

		X_1		
		Q1	Q2	Q3
X_2	Q1	.1067	.1142	.1262
	Q2	.0942	.1009	.1117
	Q3	.0843	.0903	.1001

Table 3: Probability estimates table

Based on the score we calculate in above, we label subjects "high risk" if the score exceed a certain value k. Predicting "high risk" patients as 30-days death event($\hat{Y} = 1$), we produce Table 4 of sensitivity, specificity, positive predictive value, and negative predictive value corresponds to each k(=0.25, 0.50).

To see if the classification was appropriate, we plot Figure 3 with contours of the fitted model and the data points. The contour plot shows that the model predicts a higher probability of $Y = 1$ when X_1(serum creatinine) is larger and X_2(ejection fraction) is smaller, yet we see $Y = 1$ event is scattered all over the plot.

Then, we conduct Hosmer-Lemeshow goodness of fit test to see if we need to add more variables to this model. According to the result, the χ^2 was 7.5743 and p-value was 0.4761, which means the observed and expected proportions for each group are not significantly different, and adding more variables is not necessary.

After fitting a model with X_1 and X_2, we build another model that incorporates an interaction term(X_1X_2). Serum creatinine(X_1) is still significant at 99% confidence level, however, the interaction term was not significant. The result is presented in Table 5.

As we did for the first model, again, we can estimate the odds ratios and confidence intervals corresponding ten unit changes in X_1, when X_2 is at its $25^{th}, 50^{th}, 75^{th}$ percentiles. The result is presented in Table 6. We also have included performance Table 7 and the contour plot for model 2 in Figure 4.

For the last part of task 1, we run the Hosmer-Lemeshow goodness of fit testing to see if we need to add more variables in this model. The result shows adding more terms is not necessary(χ^2=7.5474, p-value = 0.4789).

	k=25%	k=50%
Sensitivity	0.08571429	0.02857143
Specificity	0.97297297	0.99613900
Positive Predictive Value	0.30000000	0.50000000
Negative Predictive Value	0.88732394	0.88356164

Table 4: Performance of Model 1(Sensitivity, Specificity, PPV, and NPV for threshold k=0.25 and 0.50)

	Estimate	Std. Error	z value	p-value	95%CI Lower limit	95%CI Upper limit
(Intercept)	-2.830511	0.992926	-2.851	0.00436	-4.86087872	-0.938724722
X1	0.837079	0.423331	1.977	0.04800	0.05216506	1.772978319
X2	0.003101	0.023597	0.131	0.89546	-0.04348949	0.050360571
X1:X2	-0.010073	0.009048	-1.113	0.26560	-0.03241195	0.006292441

Table 5: Test result(Model 2)

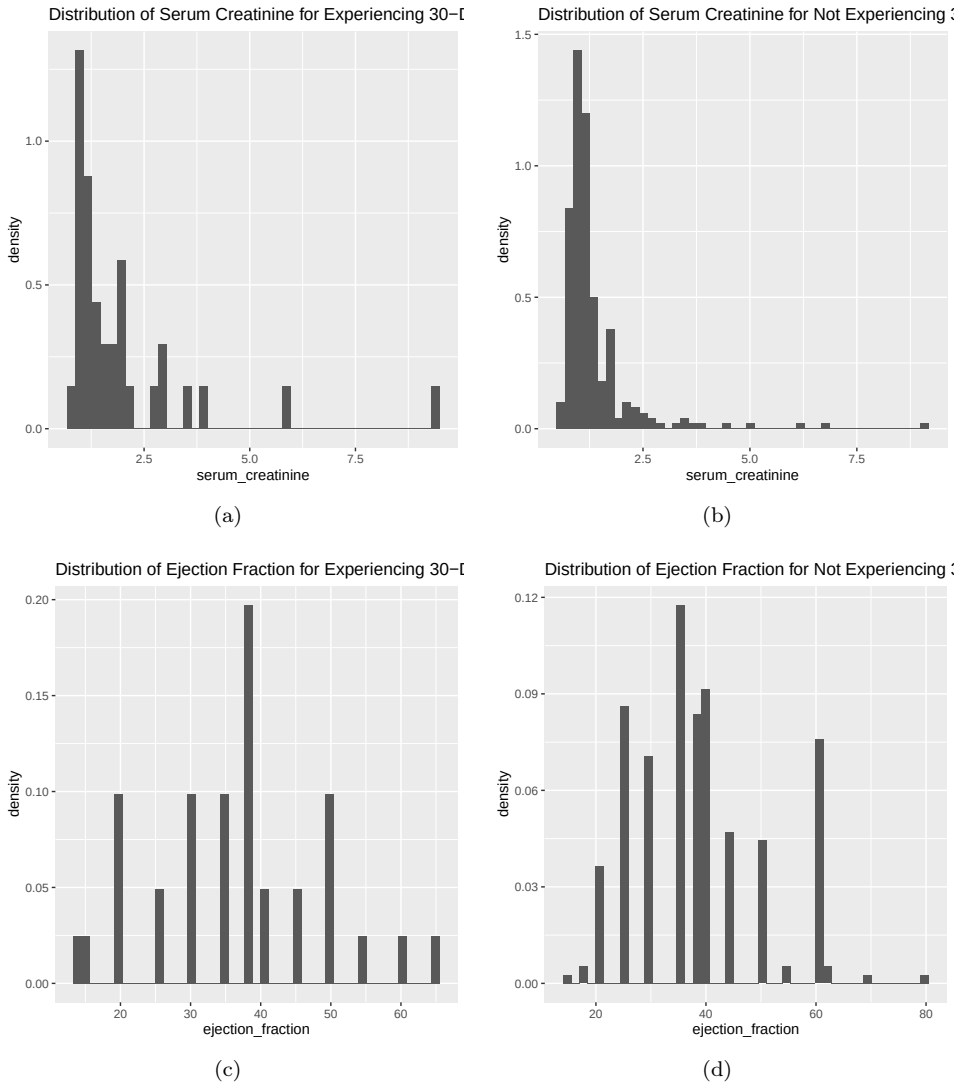

Figure 2: The distributions of (a): serum creatinine experiencing 30 days mortality. (b): serum creatinine not experiencing 30 days mortality. (c): ejection fraction experiencing 30 days mortality. (d): ejection fraction not experiencing 30 days mortality.

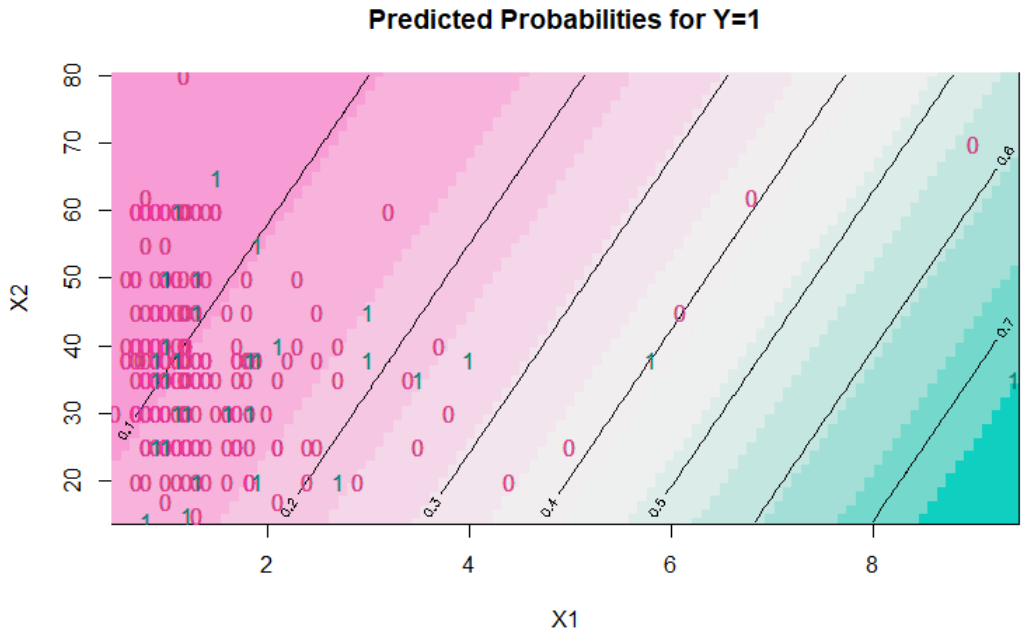

Figure 3: Contour plot of Model 1

Odds ratio	$X_2 = Q_1$	$X_2 = Q_2$	$X_2 = Q_3$
$\frac{\frac{\pi(x_1+10\|x_2)}{1-\pi(x_1+10\|x_2)}}{\frac{\pi(x_1\|x_2)}{1-\pi(x_1\|x_2)}}$	210.39343	93.98763	46.43595

Table 6: Model 2: Odds ratio for ten unit change in X_1

	k=25%	k=50%
Sensitivity	0.1428571	0.02857143
Specificity	0.9729730	0.99227799
Positive Predictive Value	0.4166667	0.33333333
Negative Predictive Value	0.8936170	0.88316151

Table 7: Performance of Model 2(Sensitivity, Specificity, PPV, and NPV for threshold k=0.25 and 0.50)

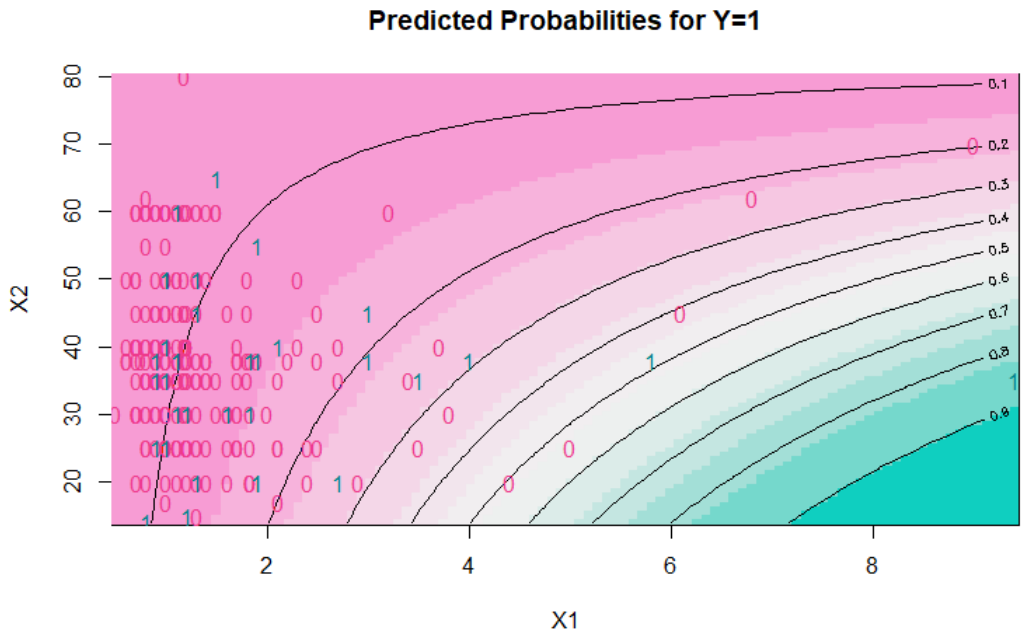

Figure 4: Contour plot of Model 2

2 30-Day Mortality Data Analysis Continued

Introduction

In the second task, we are no longer limited to using serum creatinine and ejection fraction as predictors of 30-day mortality. All the predictors can be included in the model except `time` and `DEATH EVENT`. We are going to find the best model based on different model selection criteria.

Data Analysis

We first want to use the forward selection. The procedure continues until we have 5 predictors in our model as asked by the question. Table 8 show the whole process of AIC based forward selection where
Z_1 =`Age`
Z_2 =`Serum Sodium`
Z_3 =`Serum Creatinine`
Z_4 =`Anaemia`
Z_5 =`Platelets`.

Model	AIC
$M_1 : logit[\pi(z)] = -6.314 + 0.067Z_1$	198.8
$M_2 : logit[\pi(z)] = 7.49 + 0.066Z_1 - 0.101Z_2$	194.1
$M_3 : logit[\pi(z)] = 5.257 + 0.064Z_1 - 0.087Z_2 + 0.277Z_3$	192.6
$M_4 : logit[\pi(z)] = 5.737 + 0.0631Z_1 - 0.092Z_2 + 0.255Z_3 + 0.58Z_4$	192.4
$M_5 : logit[\pi(z)] = 6.277 + 0.066Z_1 - 0.092Z_2 + 0.252Z_3 + 0.608Z_4 - 0.000003Z_5$	192.7

Table 8: Forward Selection Model

There are two main reasons why forward selection is better than backward elimination. Firstly, backward elimination is not a wise choice because it is possible to eliminate the most influential variable first if there is high level of collinearity between the candidate explanatory variables. However, in forward selection process, it is guaranteed that the model includes the most influential variables since we include the variables from the most influential to the least influential. That's why forward selection is better than backward elimination here.

Secondly, backward elimination is only available when the starting model(the largest model) is appropriate. Here, we have only 35 observations with Y equals 1. Therefore, adding twelve variables to one model is not a good strategy. Forward selection, on the other hand, starts with the smallest model so it's more appropriate.

Table 9 shows the AIC and BIC for 5 different models that fitted in the Table 8. From it we see that model M_2 and model M_4 has the lowest BIC and AIC values, respectively, which means the could potentially be our best model candidates.

Model	AIC	BIC
M_1	198.8	206.2
M_2	194.1	205.2
M_3	192.6	207.4
M_4	192.4	210.8
M_5	192.7	214.8

Table 9: AIC and BIC for 5 models

Our selected models are shown in Equation 2 and 3, which are M_2 and M_4 in our setting.

$$logit(\pi(z)) = 7.49 + 0.066Z_1 - 0.101Z_2 \tag{2}$$

$$logit(\pi(z)) = 5.737 + 0.063Z_1 - 0.092Z_2 + 0.255Z_3 + 0.58Z_4 \quad (3)$$

95% confidence intervals table for model M_2, Equation 2

	2.5 %	97.5 %
(Intercept)	-3.06816336	18.09720863
Age	0.03600109	0.09828357
Serum Sodium	-0.17943089	-0.02506284

95% confidence intervals table for model M_4, Equation 3

	2.5 %	97.5 %
(Intercept)	-5.14876990	16.61029036
Age	0.03218827	0.09580282
Serum Sodium	-0.17182686	-0.01376738
Serum Creatinine	-0.03902818	0.54286179
Anaemia	-0.18477302	1.36160158

p-values for tests of null hypotheses on individual model parameters For model M_2, Equation 2,

Age	Serum Sodium
0.000028	0.0089

Since the p-values are less than 0.05 level of significance. So, we can say that age and serum sodium have a significant effect on mortality.

For model M_4, Equation 3

Age	Serum Sodium	Serum Creatinine	Anaemia
0.000092	0.0206	0.0796	0.138

Since the first two p-values for `Age` and `Serum Sodium` are less than 0.05 level of significance. So, we can say that age and serum sodium have a significant effect on mortality but `Serum Creatinine` and `anaemia` have no significant effect on mortality.

We, then, calculate the sensitivity, specificity, positive predictive value, and negative predictive value for thresholds of .25 and .5, which are summarized in Table 10 and Table 11

	k=25%	k=50%
Sensitivity	0.3428571	0.08571429
Specificity	0.9382239	0.99227799
Positive Predictive Value	0.4285714	0.6
Negative Predictive Value	0.9135338	0.88927336

Table 10: Performance of M_2(Sensitivity, Specificity, PPV, and NPV for threshold k=0.25 and 0.50)

	k=25%	k=50%
Sensitivity	0.3714286	0.1142857
Specificity	0.9266409	0.9922780
Positive Predictive Value	0.4062500	0.6666667
Negative Predictive Value	0.9160305	0.8923611

Table 11: Performance of M_4(Sensitivity, Specificity, PPV, and NPV for threshold k=0.25 and 0.50)

Next thing we do is to perform the Hosmer-Lemeshow goodness of fit test.

For model M_2,

```
X-squared = 9.376, df = 8, p-value = 0.3116
```

Since, the p-value=0.3116 is greater than 0.10 level of significance. So, we can say that the model is a good fit.

For model M_4,

```
X-squared = 13.628, df = 8, p-value = 0.09198
```

Since, the p-value=0.09 is less than 0.10 level of significance. So, we can say that the model is a poor fit.

We also can calculate the AIC and BIC for the model we fitted in the first task. Here, X_1 =Serum Creatinine and X_2 =Ejection Fraction.

Without Interaction,

```
AIC   211.6992
BIC   222.7499
```

With Interaction,

```
AIC   212.2851
BIC   227.0194
```

DFBETAS are statistics that indicate the effect that deleting each observation has on the estimates for the regression coefficients. We are not going to list all the dfbeta values here. Instead, graphs can be a very effective way to view it, which is shown in Figure 5. We also summarize some very influential points in Table 12.

	Task 1	
Variable	Model 1	Model 2
Serum Creatinine(X_1)	10 ($X_1 = 9.4$, $X_2 = 35$,$Y = 1$, $dfbeta = 0.0571$) 224($X_1 = 5$, $X_2 = 25$,$Y = 0$, $dfbeta = -0.0299$)	213 ($X_1 = 9$, $X_2 = 70$,$Y = 0$, $dfbeta = 0.3575$) 224 ($X_1 = 5$, $X_2 = 25$,$Y = 0$, $dfbeta = -0.2106$)
Ejection fraction(X_2)	1 ($X_1 = 1.9$, $X_2 = 20$,$Y = 1$, $dfbeta = -0.0030$) 213 ($X_1 = 9$, $X_2 = 70$,$Y = 0$, $dfbeta = -0.0060$)	7 ($X_1 = 1.2$, $X_2 = 15$,$Y = 1$, $dfbeta = -0.0058$) 213 ($X_1 = 9$, $X_2 = 70$,$Y = 0$, $dfbeta = 0.0145$)
Interaction(X_1X_2)		44 ($X_1 = 4.4$, $X_2 = 20$,$Y = 0$, $dfbeta = 0.0035$) 224 ($X_1 = 5$, $X_2 = 25$,$Y = 0$, $dfbeta = 0.0036$)
	Task 2	
Variable	M_2	M_4
Age	24 ($Z_1 = 95$, $Z_2 = 138$,$Y = 1$, $dfbeta = 0.0045$) 51 ($Z_1 = 95$, $Z_2 = 132$,$Y = 0$, $dfbeta = -0.0048$)	28 ($Z_1 = 94$, $Z_2 = 134$,$Z_3 = 1.83$, $Z_4 = 0$ $Y = 1$, $dfbeta = 0.0020$) 51 ($Z_1 = 95$, $Z_2 = 132$,$Z_3 = 2$, $Z_4 = 1$ $Y = 0$, $dfbeta = 0.0003$)
Serum Sodium	5 ($Z_1 = 65$, $Z_2 = 116$,$Y = 1$, $dfbeta = -0.0194$) 195 ($Z_1 = 60$, $Z_2 = 113$,$Y = 0$, $dfbeta = 0.0242$)	19 ($Z_1 = 48$, $Z_2 = 121$,$Z_3 = 1.9$, $Z_4 = 1$ $Y = 1$, $dfbeta = -0.0175$) 195 ($Z_1 = 60$, $Z_2 = 113$,$Z_3 = 1.8$, $Z_4 = 0$ $Y = 0$, $dfbeta = 0.0207$)
Serum creatinine		26 ($Z_1 = 58$, $Z_2 = 134$,$Z_3 = 5.8$, $Z_4 = 1$ $Y = 1$, $dfbeta = 0.0657$) 213 ($Z_1 = 54$, $Z_2 = 137$,$Z_3 = 9$, $Z_4 = 1$ $Y = 0$, $dfbeta = -0.0989$)
Serum Anaemia		1 ($Z_1 = 75$, $Z_2 = 130$,$Z_3 = 1.9$, $Z_4 = 0$ $Y = 1$, $dfbeta = -0.0547$) 25 ($Z_1 = 70$, $Z_2 = 136$,$Z_3 = 1.3$, $Z_4 = 0$ $Y = 1$, $dfbeta = -0.0544$) 51 ($Z_1 = 95$, $Z_2 = 132$,$Z_3 = 2$, $Z_4 = 1$ $Y = 0$, $dfbeta = -0.0562$)

Table 12: Influential Observation index, Predictor Value, and Corresponding DFBETA

Finally, we can choose the best model based on the information criteria, performance, and the result from Hosmer-Lemeshow goodness of fit. From all of information, we choose model M_2 =`Age + Serum Sodium`, Equation 2, is better for predicting 30-days survival response variable.

We choose this model because this model has the lowest BIC among all models. Moreover, it fits the data well because we see that it gives a significant result on the Hosmer-Lemeshow goodness of fit test.

Conclusion

Our final model is not the same as in D and G (2020). Our model only based on the statistical standpoint. In D and G (2020), however, they may consider the view from the medical perspective. Like `ejection fraction`

could be useful for the doctor to assess one patient but in statistics field is not useful. Definitely, one can argue that our model is not the best since the AIC value does not show the same result and the sensitivity is not the highest. However, Dr. Charnigo said in the meeting that AIC favors the model with more predictors, while BIC penalizes it. The other reason is the interpretability of the model. M_2 only contains two predictors so is much easier to interpret than M_4 model. Additionally, the AIC and BIC across different models do not have significance difference. They only differ within single digit. So, a simpler model could be a better choice. Thus, we decide to use M_2 as our best model despite it is not perfect.

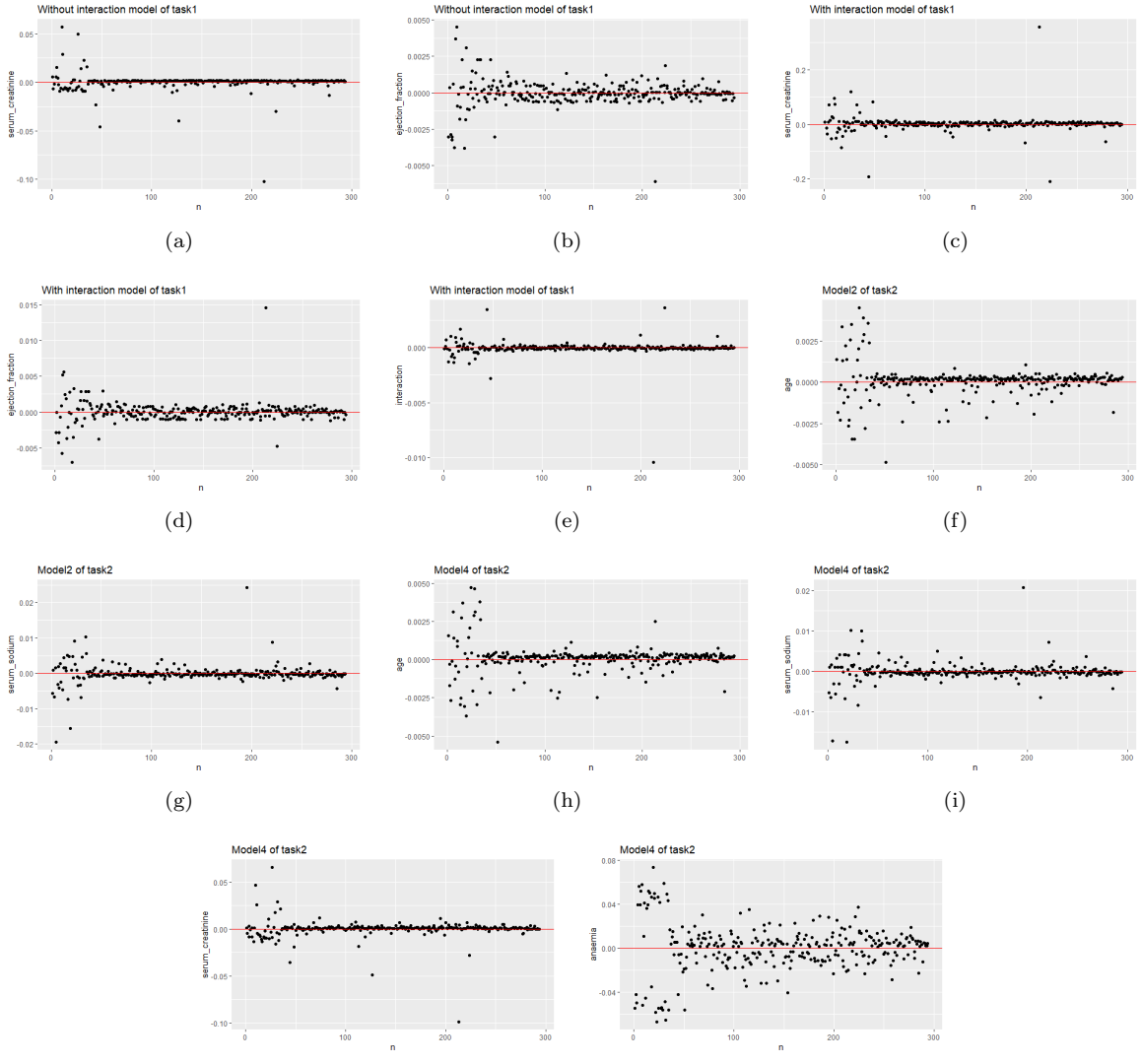

Figure 5: dfbeta values under different predictors and models. (a): Task 1, Model 1, X_1, (b): Task 1, Model 1, X_2, (c): Task 1, Model 2, X_1, (d): Task 1, Model 2, X_2, (e): Task 1, Model 2, X_1X_2, (f): Task 2, M_2, Z_1, (g): Task 2, M_2, Z_2, (h): Task 2, M_4, Z_1, (i): Task 2, M_4, Z_2, (j): Task 2, M_4, Z_3, (k): Task 2, M_4, Z_4

References

Chicco D and Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 20, 2020.