1. Group 23 (Group LGL)

Jinglun Li

Fuyu Guo

Xinhao Li

2. Background knowledge, feedback, and eyes on the field:

2.a. Background knowledge

In this study, we aim to explore the relationship between renal dysfunction and mortality rates among patients with heart failure. The question was firstly studied by Ahmad et al utilizing survival analysis. [1] In their work, Ahmad mainly identified potential risk factors for mortalities among patients with heart failure. Renal dysfunction (measured by serum creatinine) is reported as an important risk factor. A 1-unit increment in serum creatinine is associated with 2.24 hazard ratio with p-value less than 0.05. Chicco et al. re-analyzed the dataset using machine learning methods and reported that the serum creatinine level is the most important predictor for these patients' mortalities. [2] Analogous research topic has been studied by Van Domburg et al., using estimated glomerular filtration rate (eGFR) as the indicator of renal dysfunction in patients with known or suspected coronary artery disease. [3] Van Domburg et al. employed a multivariable adjusted Cox model and found the hazard ratios were 1.33, 1.67, and 3.38 among patients with mild, moderate, and severe renal impairment compared to their peers with normal renal function. In summary, although extent literature employed different biomarkers for renal dysfunction, increasing hazard ratios were reported for impaired renal function among patients with cardiovascular diseases.

After reviewing existing literature, we aimed to use a multivariable adjusted Cox model as the main model as it is commonly used in this research area. However, previous studies reported less on the modification of other biomarkers such as eject fraction. Potential effect modification by other biomarkers may reveal heterogenous effects of renal dysfunction among different subgroups. Thus, we aim to fill this gap in the current project.

2.b. feedback

We appreciate very much the teaching stuff and out peers' comments. These comments point out several very important questions which we will incorporate in our analysis. We aggregate the comments and show our response and analysis plan below (*for question i and ii*).

> In the primary question, you want to explore the association between serum creatine and mortality rates for patients with heart failure. I think the question is not specific enough. What kind of association you expect they should have? Which aspects you aim to explore? In model fitting part, how about add a model selection process that is to fit different models that we mentioned in class, can compare which models fit better?
> Response: It is a very helpful question. We will make it more clear that there are four sperate models and the association definitions are different: 1) The association between the expected survival time and serum creatine: the association is on average how many days will increase if the patient's serum creatinine level decreased by a 1 mg/dL in a linear regression model; 2) the association between death by 30 days and serum creatine: the association is the average odds ratio of death by 30 days if the patient's serum creatinine level decreased by a 1 mg/dL in a logistic regression model; 3).) the association between incident rates and serum creatine: the association is the average incidence rates ratio of death if the patient's serum creatinine level decreased by a 1

mg/dL in a Poisson regression model; and 4) the association between mortality rates and serum creatine: the association is the average hazard ratio if the patient's serum creatinine level decreased by a 1 mg/dL in a Cox model.

Regarding model selection, we will employ a Lasso regression to help us to determine the optimal covariate sets in the model. Candidate covariate sets will also be included in the model. The AIC will be used as a metric to compare the goodness of fit in models.

➢ The dataset is small. Overfitting is easier to happen. Simper model may be better.

Response: We appreciate this comment. As a solution to limited sample size, we plan to utilize a lasso regression in the models to help us identify key covariates influencing patients' morality. Based on the sparse results of the Lasso regression, we will decide which covariates to be included in our main analysis. In this study, we prefer lasso regression as a prediction selection method to stepwise regression, because stepwise regression is somehow time-consuming and may get inconsistent predictor sets when utilizing different directions. To determine the optimal penalty parameter in lasso, λ, we will utilize a cross-validation process in the "glmnet". After getting the most suitable covariate sets, we will delete and add a covariate according to the lasso results, and assess the goodness of fit in the model with these three different covariate sets. Detailed analysis please see the methods section below.

➢ My only advice is about your data set. Even though your dataset has only 299 cases, the dataset seems clean and complete, so after you check if there is any missing data and if it's not too much, maybe using complete cases with all data entries would be both easier and more accurate, Or at least the variable you are interested in is complete. If the

missing data reduce the number of cases a lot, the mice function in R could also be helpful.

Response: We appreciate this comment. Actually, after checking the dataset, luckily, we found there were no missing values. Thus, we think it is no need to make imputations. Thanks for the comment, and thanks for sharing this important package *MICE*.

2.b.iii.

As the research problem has been widely discussed in previous medicine literature, though the quantitative results are relatively limited, its clinical meaning is relatively clear. Given the limited time of this project and our limited resources, we admit we fail to contact a clinician expert during the past 1 month. We will keep reading related literature and try our best to seek domain experts' help.

3. Analysis Plan:

We make several amendments to our original analysis plan. We highlight these changes below.

a) Data cleaning: although the dataset has been elaborated by Davide Chicco, we will check any potential missing data in the outcome, exposure, and covariates. We will report the number of missing data and if the number is less than 10%, we will consider including a "missing" indicator for categorical variables and imputing continuous variables. We will report the final number of patients included in our study.

b) Checking the exposure:

The primary exposure, serum creatine, which is continuous, is usually categorized into two different levels ($\leq$ 1.5 mg/dL for the normal level vs, and > 1.5 mg/dL for the abnormal level). In this project, we will first treat the serum creatine as a continuous variable and calculate its sample mean, standard deviation, median, and range. Also, to make it comparable with previous studies, we will assess the serum creatine as a binary variable and report the proportion of normal and abnormal levels in patients.

c) Checking the outcomes:

We will calculate the average person-time until death in the normal serum creatinine group and the abnormal serum creatinine group. We will also calculate the Death 30-day in these two groups. Survival plots will be made to visualize the mortality rates in serum creatinine groups. Chis-squared tests will be applied to test the difference of Death 30-day in these groups.

d) Checking other covariates.

In this project, age (continuous), sex (male vs. female), anemia (yes vs. no), diabetes (yes vs. no), ejection fraction (≤ 30, 31-44, and ≥ 45), smoking (yes vs. no), platelets (continuous, kilo platelets/mL), and serum sodium (continuous, mEq/L) will be considered as covariates. The proportions for categorical covariates and mean (standard deviations) for continuous covariates in normal serum creatinine group and abnormal serum creatinine group will be calculated and compared using chi-squared tests and t-tests respectively.

e) Modeling analysis

First, we will do a simple linear regression do assess the association between survival time and serum creatine level in patients who died by the end of the study. The serum creatine level will model as a continuous and a categorical variable respectively. Second, the probability of Death 30-day will be modeled by logistic models. Third, we combine the incidence of deaths and time at risk among patients with identical covariate patterns. Poisson regression models will be employed to model the association between serum creatine level and incidence rate of deaths. Last, in our main analysis, we will perform survival analysis. A Kaplan-Merrier plot will be made for patients stratified by serum creatine (normal vs. abnormal). Then Cox proportional-hazards model will be performed, with outcome to be the survival time with the event (0 for censored and 1 for death). In addition to serum creatine level (both continuous and categorical variable will be assessed), appropriate covariates will be adjusted for.

f) Model selection

Considering the relatively small sample in this project, we decide to make our model as parsimonious as possible. In this sense, we will utilize lasso regressions to help us determine the optimal covariate sets in the models mentioned in subsection e), except the

Poisson regression. To justify the selection by the lasso regression, based on the covariates automatically determined by the software, we decide to include and delete a covariate. Thus, in each model we will have three sets of covariates. We will perform models using these three covariates sets and compare the model performance based on AIC.

g) Subgroup analysis for potential effect modification.

To check whether effect modification exists, we will include an interaction term between serum creatine and potential covariates finally decided by our model selection in subsection f) in the Cox model. The p-value of the interaction term as well as variance-deviance analysis will be used to determine if there are any effects modifications.

h) Checking nonlinearity

To check whether the relationship between serum creatine and mortality risks is linear, we will replace the linear term in the fully adjusted model with a natural spline of serum creatine. The knots and degrees of freedom of the spline will be determined during the following analysis.

4. Missing Data

After checking the data clearly, there was no missing in this dataset, partly because Davide Chicco et al. have elaborated the original dataset before uploading to the archive. Here we would like to answer the following questions as required supposing we did have some missing data in serum creatine and smoking.

4.a.

Because the biomarkers information was retrieved from blood reports. Smoking and drinking information was retrieved from clinicians' notes. If we had some missing data in serum creatine and smoking, it would be more likely be Missing Completely at Random (MCAR) as these documents should be recorded in the reports. The missing is more likely to be caused by system errors or notes missing. We can test the assumption by regressing the chance of missing events on the available covariates. If there are no significant associations, we deem it is a MCAR. Otherwise, we will deem it as Missing at Random (MAR). Missing Not at Random (MNAR) is not considered because of the underlying data generating process. If the data is MCAR, we will just use the complete cases in our analysis. If the data is MAR, complete case analysis is fine, but we will consider using a missing indicator.
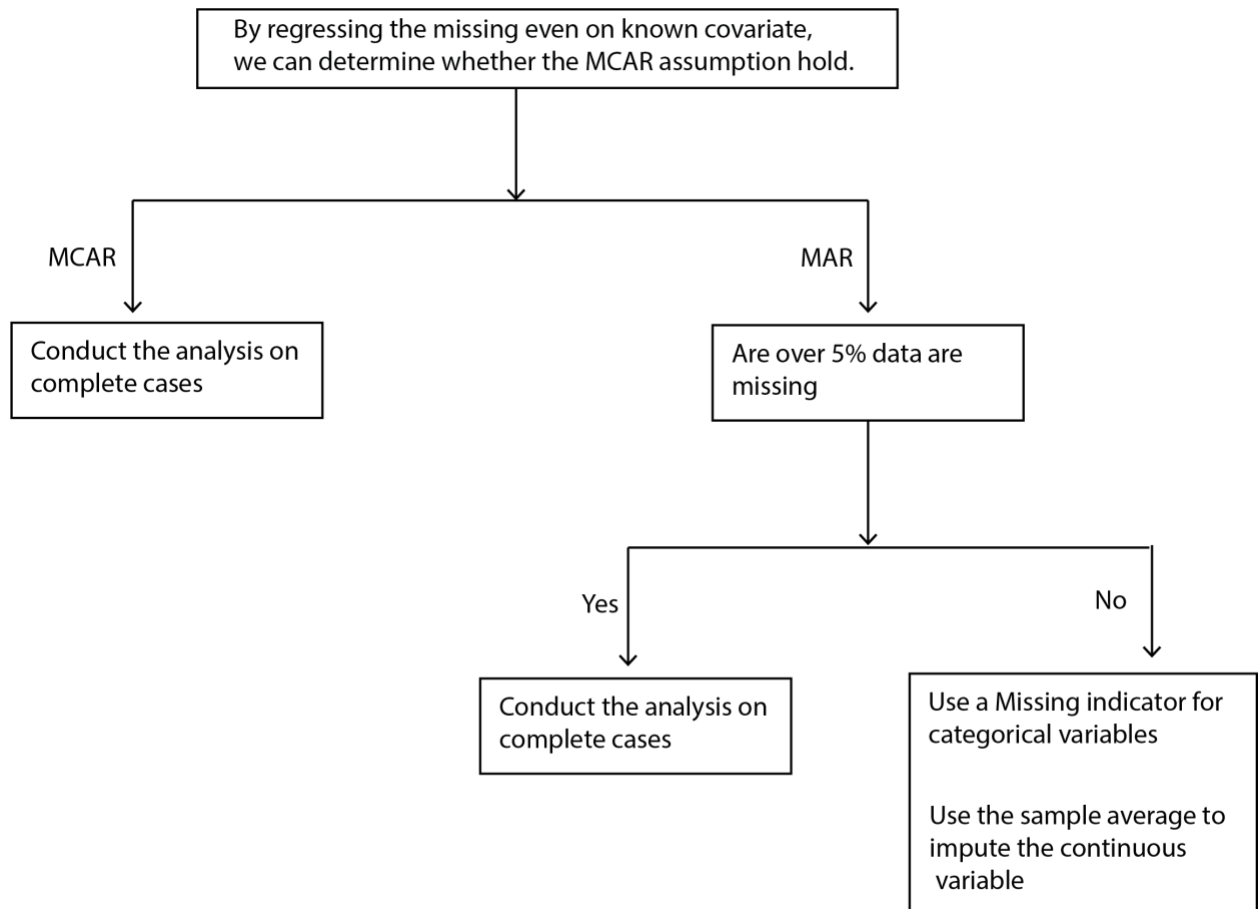
4.b. Please check the diagram below.

Figure 1. Flow chart for handling missing data

It is worth noting that although we discussed the potential scenarios that we had missing values

in the data, there is no missing data in the actual dataset. Thus, the steps above are just discussed

and will not be used in the real data analysis process.

5. Modeling

5.a. Linear, flexible/additive or other methods (LASSO, ridge) from this topic:

==Originally, the linear regression is not our most interested method because we assess the survival time for patients with heart failure. We generated a new variable which fits the linear regression framework.==

> ➤ New variable: the death time for people who died by the end. There were 96 people who actually died by the end of the study period. Their time to death can serve as a linear outcome in the linear model.

> ➤ We will assess the association between death time and the baseline serum creatine level among those who died by the end of the study (n = 96). It is a quite small sample size. We first fitted a crude a linear model

$$death\ time_i = \beta_0 + \beta_1 serum\ creatine_i + \epsilon_i$$

> ➤ The fitted results showed that the β_1 is 0.6687, with a p-value of 0.879. The results suggested that in the crude model there was no significant association between serum creatine level and death time among people who died by the end.

> ➤ We then decided to fit a multivariate linear model adjusting for other covariates. Since the sample size is relatively small, we utilized a Lasso to help determine the covariate sets. Unfortunately, according to the results of Lasso, it suggested that we should only include the intercept term in our model which is not very helpful.
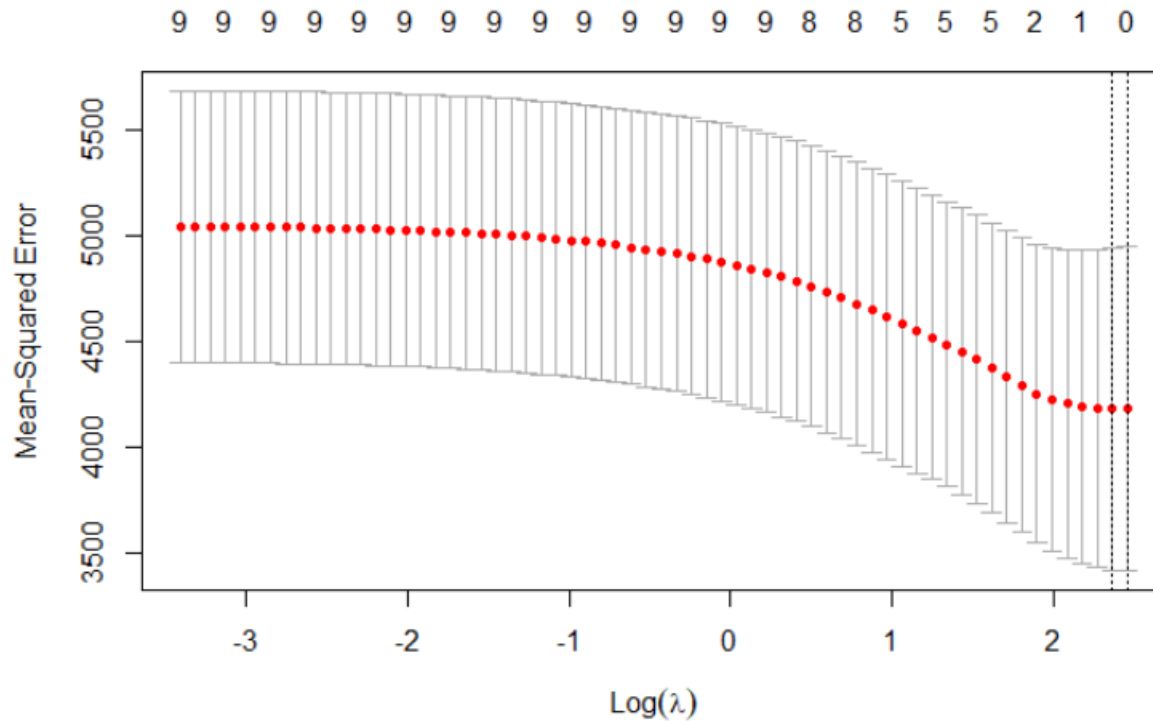
Figure 2. Cross-validation results of LASSO when choosing the most appropriate penalty parameter in the linear regression.

➢ Then, we just adjusted for patients' age and sex based on the background knowledge

$$death\ time_i = \beta_0 + \beta_1 serum\ creatine_i + \beta_2 age_i + \beta_3 Sex_i + \epsilon_i$$

Again, the results showed that there was no significant association between serum creatine level and death time, with β_1 equal to 1.1688, the p-value equal to 0.80.

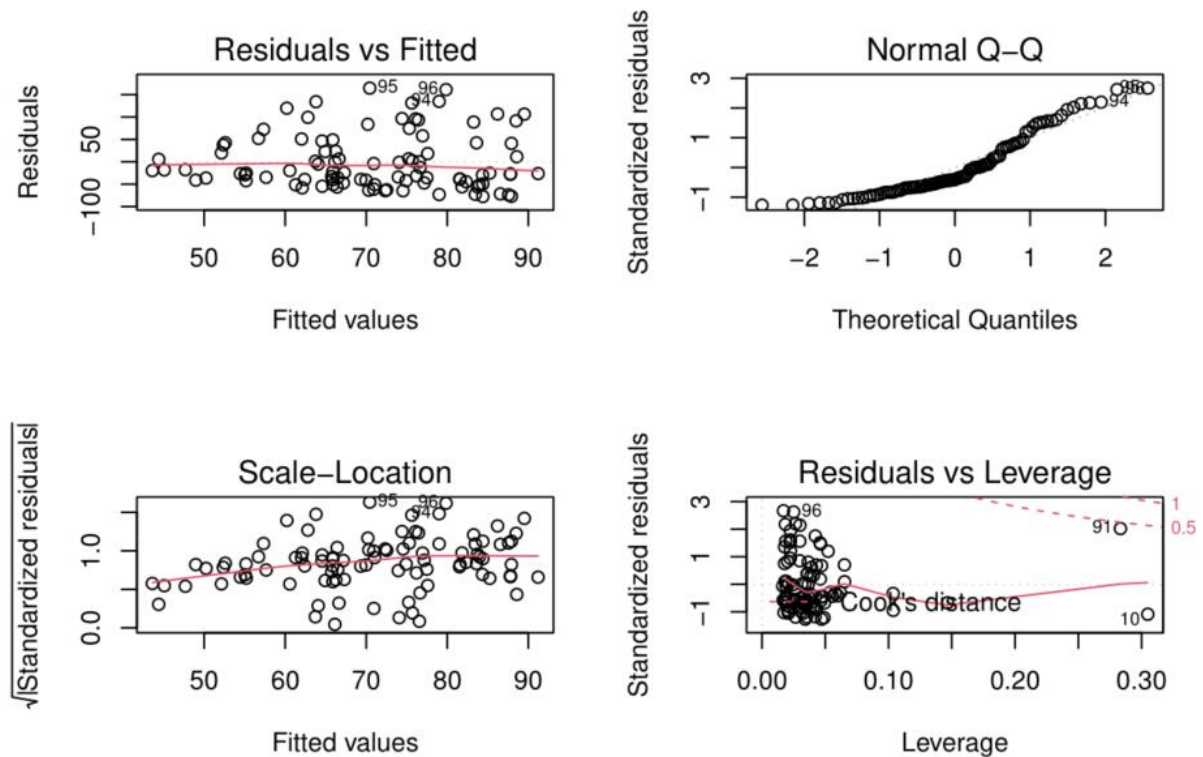➢ Based on this result, we performed some model diagnostics, using external studentized residuals.

Figure 3. Residual analysis plots

Based on the figures above we conclude that:

➢ The variance is not equal for different observations.

➢ The residual is not normally distributed.

➢ We cannot clearly teal if the linearity assumption is violated or not. It seems it still holds
in this model.

➢ Although there are three high-influencing points, they are still within the boundary of
Cook's distance. Thus, we conclude there is no large influence from outliers or leverage
values.

For the goodness-of-fit of the model. The adjusted-R^2 is 0.0045, indicating the model is not predicting the death time of patients well. This is in line with our findings of insignificant associations.

5.b. Logistic, multinomial, ordinal:

Originally, the logistic regression is not our most interested method because we assess the survival time for patients with heart failure. We generated a new variable which fits the logistic regression framework.

➢ We modeled the association between death by 30 days ("Death 30-day", binary variable) and serum creatine level. A total of 5 participants were lost to follow-ups before the 30 days. We excluded them from the data set in this analysis.

➢ Let p denote the probability of Death 30-day

In a crude model,

$$log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 serum\ creatine_i$$

The estimated OR for a 1 mg/dL increment in serum creatine level for Death 30-day is 1.44 (95% CI 1.12 to 1.89). It means that high level of serum creatine value is harmful for patients with heart failure.

➢ In multivariate adjusted logistic models, we first used Lasso regression to help us determine the appropriate covariate sets.
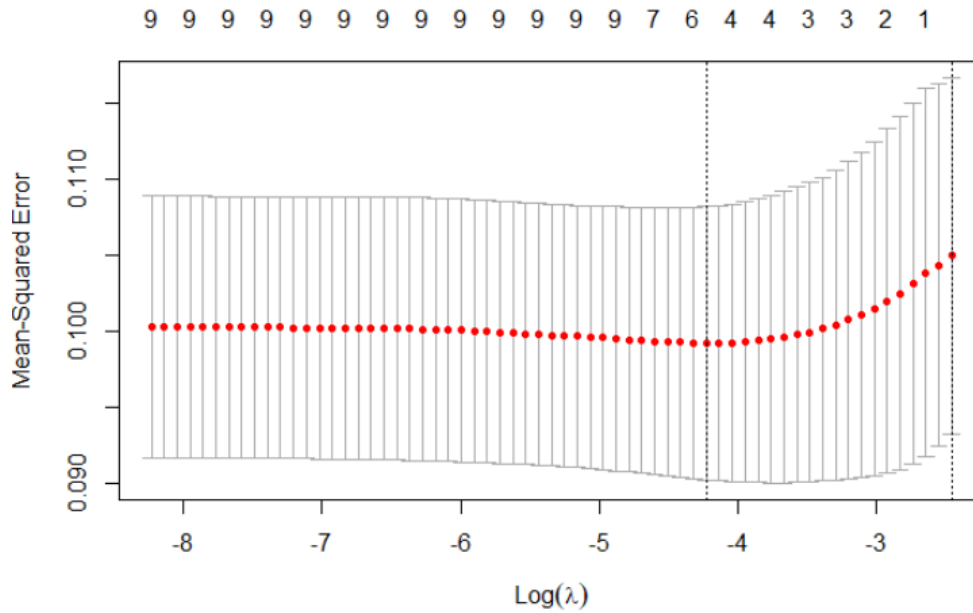
Figure 5. Cross-validation results of LASSO when choosing the most appropriate penalty parameter in the logistic regression.

➤ According to the cross-validation results using (MSE as the metric), the optimal number of covariates was 6 and were listed in the table below. To justify the lasso results, along the axis of λ, we put a covariate into the covariate set and exclude a covariate out of the set. Then we perform logistic regressions for those three covariates sets and compare their results.

Table 1. Logistic regressions for the associate between baseline serum creatine level and deaths by 30 days from the baseline (N= 294)

	Model 1	Model 2	Model 3
OR for 1 mg/dL in serum creatine	1.32 (0.98, 1.78)	1.32 (0.98, 1.78)	1.32 (0.98, 1.78)
Adjusted variables	age (continuous) + anemia (binary) +	+ sex (binary)	-platelets (continuous)

	ejection fraction (continuous) + platelets (continuous) + serum sodium (continuous)		
λ in Lasso	0.01329066	0.010050	0.016010
AIC	198.6465	198.6465	198.6465

> According to the results of those three logistic regressions, 1 mg/dL increment in serum creatine is associated with 1.32 (95% CI 0.98 to 1.78) OR for death by 30-days, on average, holding other covariates constant. The point estimated OR suggests that the serum creatine is harmful for patients' survival. However, since the 95% confidence interval contains the null association, we need more samples to reach the conclusion.

> We will do some diagnostics. The line of residuals against predictors in the model are approximately horizontal, with p-values > 0.05 from the t-test, suggesting the assumption underlying the logistic models are meeting.
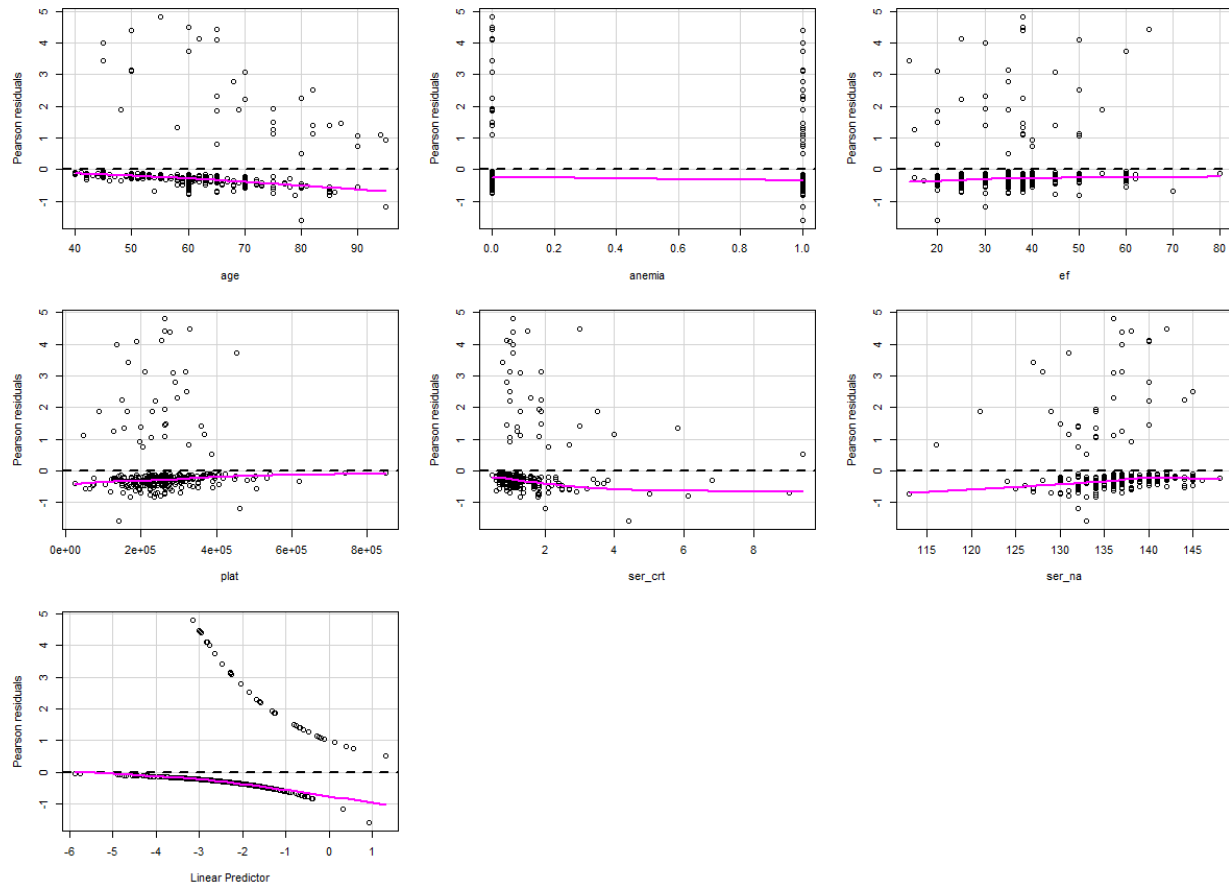
Figure 5. Plots for residuals against predictors in the logistic regression

➢ Because most predictors in the model are linear, it is hard to evaluate the model by calibration. We will evaluate its goodness to fit by discrimination. The AUC is 0.7629, which is somewhat acceptable.
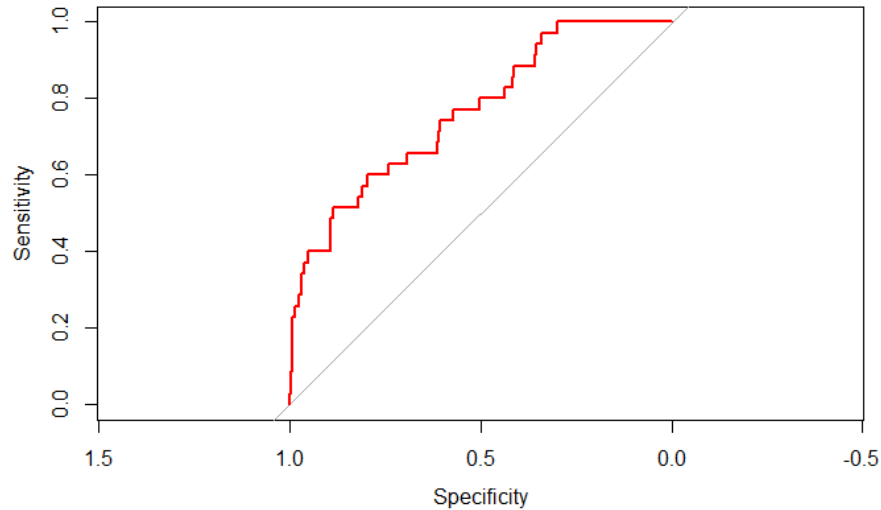
Figure 6. ROC curve for the logistic regression

5.c. Poisson

Originally, the Poisson regression is not our most interested method because we assess the survival time for patients with heart failure. We generated a new variable which fits the Poisson regression framework.

➢ We used a Poisson regression to assess the incidence rate to death during the whole follow-up period. The original dataset includes individual observations. To assess the "count" in the data, we collapsed the data into group data with the total cases of death and total time at risk based on the covariate's combinations above.

➢ We categorized persons by serum creatine (normal vs. abnormal), age (<65 vs >= 65), sex (male vs. female), eject fraction ($\leq$ 30, 31-44, and $\geq$ 45), platelets (lower than the median vs. higher than the median), and serum sodium (lower than the median vs. higher than the median). After collapse, we had 126 combinations i.e., 126 rows in the group dataset.

ID	Serum Creatine	...	Death	Time at risk
1	Normal	...	1	30
2	Normal	...	1	180
3	Abnormal	...	0	210
4	Abormal	...	1	30

↓

Group ID	Serum Creatine	...	Death cases	Time at risk
1	Normal	...	2	$30 + 180 = 210$
2	Abnormal	...	1	$210 + 30 = 240$

Figure 6. Schema for data collapse for Po

➢ We applied a Poisson regression to test the association between incidence rate of death and serum creatine level. The offset was the total time at risk in that group.

Let λ denote the incidence rate of death, $\mu = \lambda t$ denotes the mean incidence cased of death during the risk at time.

$$\mu_i = \beta_0 + \beta_1 serumcreatine(abnormal) + \sum_j \beta_j X_j, i + log(t_i)$$

$X_{j,i}$ represents the adjusted variables (including dummy variables) above.

t_i represents the total risk at time in that group

➢ According to the results of the Poisson regression, compared to people with normal level of serum creatine, the incidence rate ratio of death among people with abnormal serum creatine level is 2.60 (95% 1.68 to 4.00) on average.

➢ To account for potential over-dispersion in the variance which may violate the Poisson assumption, we conducted a quasi-Poisson regression. In the quasi-Poisson regression's assumption, the variance is $\theta\mu$. The point estimate stayed the same, but the 95% CI was wider from 1.27 to 5.52.

5.d. Survival analysis

i. We decide to take survival analysis into out project. Actually it will be the main analysis.

ii. The outcome is the survival time with the end point status for the patients with heart failure during the follow-up period. The interested exposure is the patients' serum creatine value. The adjusted covariates are age (continuous), anemia (binary), ejection fraction (continuous), platelets (continuous), serum sodium (continuous). The covariate set is just the same as in the logistic regression and Poisson regression. In the next stage of analysis, we will use a lasso regression in the survival analysis framework.

6. Writing

6.a. Abstract

Background Renal dysfunction serves is a complication of heart failure through multiple

mechanisms. Studies have shown that the serum level of creatinine, a substance readily filtered

out by healthy kidneys, acts as an indicator of kidney function could help predict mortality for

patients of heart failure. However, the potential effect modifications by other demographic and

biomarkers were not clearly assessed.

Aims This project hence sought to investigate the association between serum creatinine level and

mortality rates among patients with heart failures.

Methods A dataset consisting of 299 patients of heart failure enrolled from April 2015 to

December 2015 was studied. In the main analysis, serum creatinine level was first treated as a

continuous variable then classified into two different categories (normal vs. abnormal) with

multivariable-adjusted Cox proportional hazards models used to estimate hazard ratios (HR) and

95% confidence intervals (CI). Linear regression, logistic regression, and Poisson regression

were performed as exploratory analysis. Lasso regression was utilized to help determine the

optimal covariate sets.

Results

Conclusion

Keywords: Cardiovascular disease Serum creatinine; mortality for heart failure; model selection.

6.b. Background

Heart failure (HF) is a serious medical condition that develops when the heart doesn't pump enough blood for the whole body's needs. HF is often caused by morbidities that damage the heart like coronary heart disease, diabetes, and high blood pressure. As a serious condition requiring medical care and treatment, HF affects more than 6 million patients and their families in the United States, which brings a high public health cost and burden.[4]

Renal dysfunction is a common complication of HF, which can lead to mortalities. The reduced cardiac output and the consequently renal under-perfusion is the main pathophysiology cause of renal dysfunction because of the low renal blood flow and increased renal venous pressure.[5] Besides, neurohormonal activation (renin–angiotensin–aldosterone and sympathetic nervous system), inflammatory activation and diuretic treatment are also mechanisms leading to renal dysfunction in patients with HF. [5] Serval study has demonstrated that renal dysfunction can lead to higher mortality rates among patients with cardiovascular diseases. [1,2]

Serum creatinine is an important and commonly-used biomarker to indicator of the presence kidney dysfunction. [6] Serum creatinine level is tested through using venous blood is included in routine clinical easy procedures. Serum creatinine greater than 1.5 mg/dL will be regarded as abnormal.[1]

In this study, we aimed to employ serum creatinine as an indicator for renal dysfunction and explore its association with mortality rates among patients with HF. Ahmad et al. [1] and Chicco et al. have demonstrated the harmful associations in their previous studies. [2] However, these studies concerned less on the modification of other biomarkers such as eject fraction and serum sodium, which are also related to HF through diverse mechanisms.[8] Therefore, this study aimed to fill the

current research gap by studying potential effect modifications by other important factors and fill the vacancy of existing studies.

6.c. Methods

6.c.1. Material sources and data description

The dataset was retrieved from UCI Machine Learning Repository. It was first collected and analyzed by Ahmad et al. [1] based on medical records of 299 patients with HF at NYHA class III and IV, the most two severe HF stages. We downloaded this dataset from the UCI Machine Learning Repository and will use it under the same Attribution 4.0 International (CC BY 4.0) copyright.

A total of 13 clinical features are collected in the dataset, including age (years), sex (binary), anaemia (binary), high blood pressure (binary), creatinine phosphokinase level (mcg/L), diabetes (binary), ejection fraction (percentage), platelets (kiloplatelets/mL), serum creatinine (mg/dL), serum sodium (mEq/L), smoking (binary), follow-up period (days) and death event (binary).

In our project, serum creatinine level was first treated as a continuous variable and then categorized into two different levels (≤ 1.5 mg/dL for the normal level vs, > 1.5 mg/dL for the abnormal level).

6.c.2. Exposure

The serum creatinine was the exposure of interest in this project. It originally assessed as a continuous variable in the dataset. In this project, we first treated the serum creatinine as a continuous variable and then categorized it into two different levels (≤ 1.5 mg/dL for the normal

level vs, and > 1.5 mg/dL for the abnormal level). The sample mean, standard deviation, median, range and the proportion of two different levels were calculated.

6.c.3. Outcome

Checking the outcomes:

In this survival analysis project, the primary outcome is the time to event (i.e., survival time and the end status) for each patient. To make our outcome compatible with the linear regression, logistic regression, and Poisson regression framework, we generated several secondary outcomes and did exploratory analysis. First, among patients who died by the end of the study, their survival time until death was calculate. Second, the status of each patient by the 30 days from the start of the follow-up was identified and transformed into a binary variable (alive vs., deceased). Last, we collapsed the individual-based dataset into group-based dataset and calculated the cases of death and the total time at risk. The collapse was conducted by groups with the identical covariate patterns

6.c.4. Covariates

In this project, age (continuous), sex (male vs. female), anemia (yes vs. no), diabetes (yes vs. no), ejection fraction ($\leq$ 30, 31-44, and $\geq$ 45), smoking (yes vs. no), platelets (continuous, kilo platelets/mL), and serum sodium (continuous, mEq/L) were considered as potential covariates.

6.c.5. Statistical analysis

We first checked potential missing data in this dataset. After careful check, there were no missing values in exposure, outcome, or covariates. Then descriptive analysis was conducted exploring the

baseline characteristics for patients with normal and abnormal serum creatine. T-tests and Chi-squared tests were conducted for continuous and categorical variables respectively.

In the modeling stage, first, the simple linear regression was used to evaluate the association between survival time and serum creatinine level in patients who died by the end of the study period. Second, the probability of deaths by the 30 days was modeled using logistic models. Third, we combined the incidence of deaths and time at risk among patients with identical covariate patterns. Poisson regression models were employed to model the association between serum creatinine level and incidence rate of deaths. Last, in our main analysis, we performed survival analysis. A Kaplan-Merrier plot was made for patients stratified by serum creatinine (normal vs. abnormal). Then Cox proportional-hazards model was performed, with the outcome to be the survival time with the event (0 for censored and 1 for death). Appropriate covariates were adjusted for these models.

6.c.6. Model selection

Since the dataset is of relatively small size (289 samples in total), we utilized lasso regressions to determine the optimal covariate sets in the models mentioned above, except the Poisson regression. To justify the selection by the lasso regression, we decided to include and delete a covariate based on the covariates automatically calculated by the software. Thus, in each model we had three sets of covariates. We performed models using these three covariates sets and compared the model performance based on AIC.

To check whether effect modification exists, we included an interaction term between serum creatinine and selected potential covariates in Cox models. The potential effect modifiers were determined by p-value of the interaction term as well as by deviance analysis.

References

1.   Ahmad, T., et al., *Survival analysis of heart failure patients: A case study.* PLoS One, 2017. **12**(7): p. e0181001.

2.   Chicco, D. and G. Jurman, *Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone.* BMC Medical Informatics and Decision Making, 2020. **20**(1): p. 16.

3.   Van Domburg RT, Hoeks SE, Welten GM, Chonchol M, Elhendy A, Poldermans D. Renal insufficiency and mortality in patients with known or suspected coronary artery disease. *J Am Soc Nephrol.* 2008;19(1):158-163. doi:10.1681/ASN.2006101112

4.   https://www.nhlbi.nih.gov/health-topics/heart-failure

5.   Núñez, J., et al., *Early serum creatinine changes and outcomes in patients admitted for acute heart failure: the cardio-renal syndrome revisited.* European Heart Journal. Acute Cardiovascular Care, 2017. **6**(5): p. 430-440.

6.   Butler, J., et al., *Renal function, health outcomes, and resource utilization in acute heart failure: a systematic review.* Circ Heart Fail, 2010. **3**(6): p. 726-45.

7.   Chicco, D. and G. Jurman, *Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone.* BMC Medical Informatics and Decision Making, 2020. **20**(1): p. 16.

8.   Patel, Y.R., et al., *Prognostic Significance of Baseline Serum Sodium in Heart Failure With Preserved Ejection Fraction.* J Am Heart Assoc, 2018. **7**(12).

7. Appendix

```r
library(tidyverse)
library(survival)
library(ggpubr)
library(survminer)
library(glmnet)
library(car)
library(pROC)
```

```r
############################
# load data
dta <- read.csv("D:/BST_210_Heart_failure/heart_f.csv")

# select variables we will use
dta <- dplyr::select(dta,
                     "age", "sex", "anaemia",
                     "diabetes", "ejection_fraction", "smoking",
                     "platelets", "serum_creatinine", "serum_sodium",
                     "time", "DEATH_EVENT")

# rename the variables to make our work easier
names(dta) <- c("age", "sex", "anemia",
                "dbt", "ef", "smoking",
                "plat", "ser_crt", "ser_na",
                "time", "death")
dta$ser_crt_ab <- if_else(dta$ser_crt <= 1.5, 0, 1)
# check sample size
dim(dta)
```

```
## [1] 299  12
```

```r
# check if there is any missing value in variables
complete.cases(dta) %>% all()
```

```
## [1] TRUE
```

```r
# no missing value
```

Linear, flexible/additive or other methods (LASSO, ridge)

```r
# choose patients who died by the end of the study
dta_line <- dplyr::filter(dta, death == 1)
# there are only 96 people died by the end

#################################################
```

```
# a crude analysis
fit_lin_1 <- lm(time~ser_crt, data = dta_line)
summary(fit_lin_1)
```

```
##
## Call:
## lm(formula = time ~ ser_crt, data = dta_line)
##
## Residuals:
##     Min      1Q Median      3Q     Max
## -66.93 -46.63 -26.23  31.36 170.12
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.6578    10.2774   6.778 1.06e-09 ***
## ser_crt       0.6687     4.3805   0.153    0.879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.7 on 94 degrees of freedom
## Multiple R-squared:  0.0002478,  Adjusted R-squared:  -0.01039
## F-statistic: 0.0233 on 1 and 94 DF,  p-value: 0.879
```
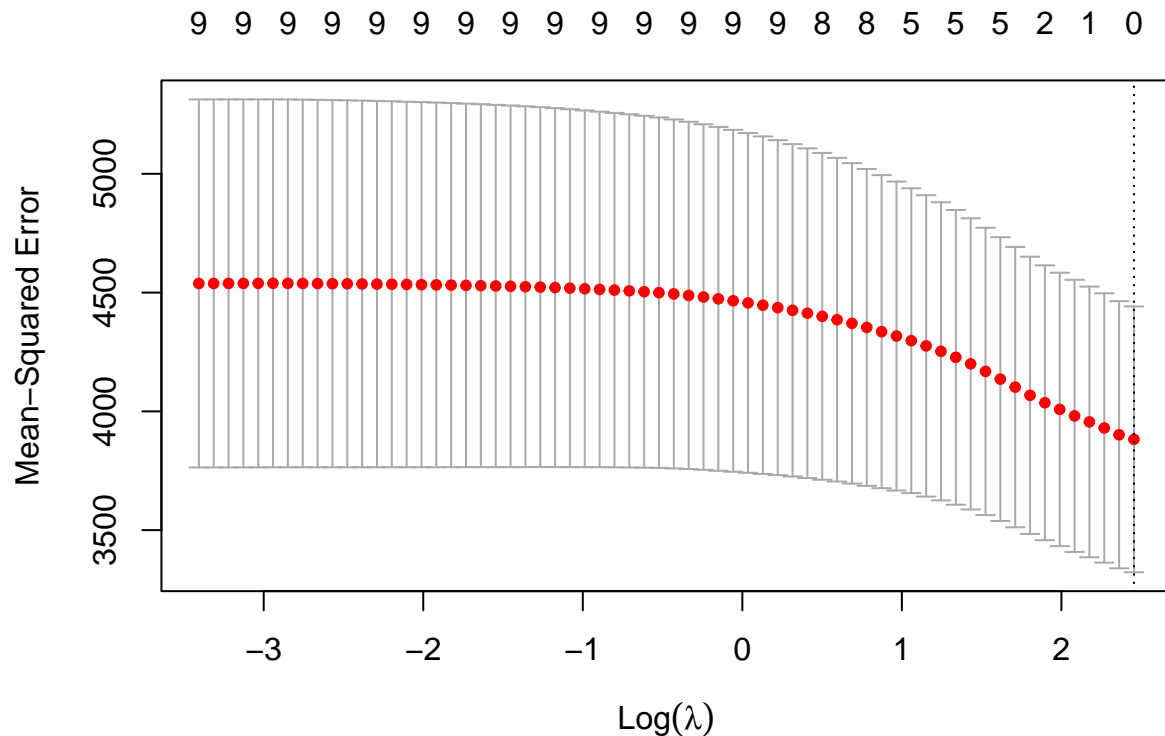
```
#################################################
# conduct a lasso regression to choose covariate sets for linear serum creatine as a continuous variabl
x <- dplyr::select(dta_line, -death, -time, -ser_crt_ab) %>%
        as.matrix()
y <- dta_line$time %>% as.vector()
cv <-  cv.glmnet(x, y, type.measure = "mse", nfolds = 4)
plot(cv)
```

The results of the lasso tells us that we should not include any covariates in the model. Only intercept is enough. This result is not helpful in guiding covariates selection, partly because of the small sample size and null association.

```
# we will adjust for age and sex based on the background knowledge
fit_lin_2 <- lm(time~ser_crt + age + sex, data = dta_line)
summary(fit_lin_2)
```

```
##
## Call:
## lm(formula = time ~ ser_crt + age + sex, data = dta_line)
##
## Residuals:
##     Min     1Q Median     3Q    Max
## -77.39 -42.70 -24.74  41.42 164.57
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 127.1234    32.9043   3.863 0.000208 ***
## ser_crt       1.1688     4.3668   0.268 0.789563
## age          -0.8914     0.4918  -1.813 0.073147 .
## sex          -0.3834    13.5115  -0.028 0.977425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.24 on 92 degrees of freedom
## Multiple R-squared:  0.03597,    Adjusted R-squared:  0.004537
```

```
## F-statistic: 1.144 on 3 and 92 DF,  p-value: 0.3355
```

```
#########################################################
#try use serum creatine as a binary variable

fit_lin_3 <- lm(time~ser_crt_ab, data = dta_line)
summary(fit_lin_3)
```
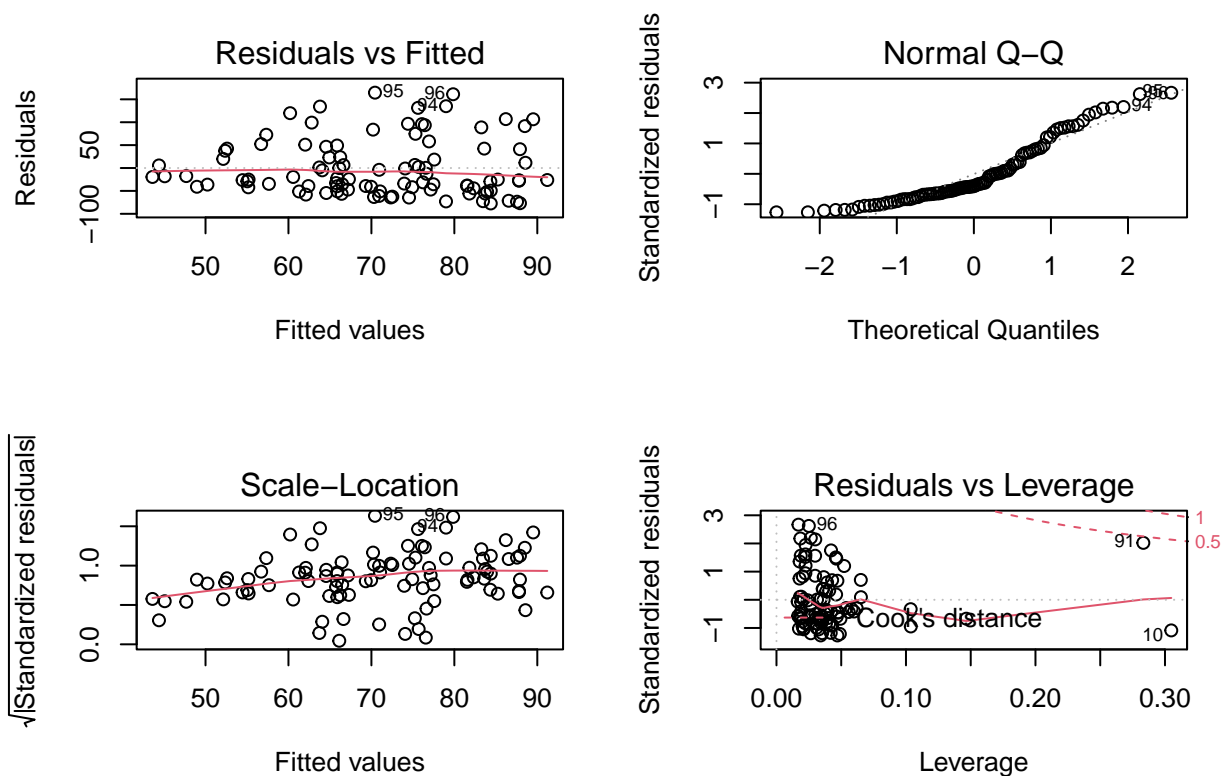
```
##
## Call:
## lm(formula = time ~ ser_crt_ab, data = dta_line)
##
## Residuals:
##     Min     1Q Median     3Q    Max
## -76.14 -49.57 -20.38  27.68 171.62
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   63.377      8.536   7.425 5.04e-11 ***
## ser_crt_ab    16.762     12.754   1.314    0.192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.14 on 94 degrees of freedom
## Multiple R-squared:  0.01804,    Adjusted R-squared:  0.007598
## F-statistic: 1.727 on 1 and 94 DF,  p-value: 0.1919
```

```
fit_lin_4 <- lm(time~ser_crt_ab + age + sex, data = dta_line)
summary(fit_lin_2)
```

```
##
## Call:
## lm(formula = time ~ ser_crt + age + sex, data = dta_line)
##
## Residuals:
##     Min     1Q Median     3Q    Max
## -77.39 -42.70 -24.74  41.42 164.57
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 127.1234    32.9043   3.863 0.000208 ***
## ser_crt       1.1688     4.3668   0.268 0.789563
## age          -0.8914     0.4918  -1.813 0.073147 .
## sex          -0.3834    13.5115  -0.028 0.977425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.24 on 92 degrees of freedom
## Multiple R-squared:  0.03597,    Adjusted R-squared:  0.004537
## F-statistic: 1.144 on 3 and 92 DF,  p-value: 0.3355
```

```
# we will do some model diagnostics with the most spares model
# residual analysis

par(mfrow = c(2,2))
plot(fit_lin_2)
```

- Variance is not equal.
- The residual is not normally distributed.
- Although there are three high influence values, they are still within the Cook's distance boundary. Thus, we conclude there is no high influence from outliers or leverage values in the model.

```
# adjusted R^2 in the model above
summary(fit_lin_2)
```

```
##
## Call:
## lm(formula = time ~ ser_crt + age + sex, data = dta_line)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -77.39 -42.70 -24.74  41.42 164.57
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 127.1234    32.9043   3.863 0.000208 ***
## ser_crt       1.1688     4.3668   0.268 0.789563
## age          -0.8914     0.4918  -1.813 0.073147 .
## sex          -0.3834    13.5115  -0.028 0.977425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.24 on 92 degrees of freedom
## Multiple R-squared:  0.03597,    Adjusted R-squared:  0.004537
```

```
## F-statistic: 1.144 on 3 and 92 DF,  p-value: 0.3355
```

Logistic, multinomial, ordinal:

```r
# assess the death by the 30-days
dta$death_30 <- NA
dta$death_30[dta$death == 1& dta$time <= 30] <- "Yes"
dta$death_30[(dta$death == 1& dta$time > 30) |
                (dta$death == 0& dta$time > 30)] <- "No"
dta$death_30[dta$death == 0& dta$time <= 30] <- "Censored"
table(dta$death_30)
```

```
##
## Censored       No      Yes
##        5      259       35
```

```r
# there are 5 patients censored at the 30 days
# we will just drop them

dta_log <- dplyr::filter(dta, death_30 != "Censored")
dta_log$death_30 <- if_else(dta_log$death_30 == "Yes", 1, 0)
############################################################
# a crude logistic model
fit_log_1 <- glm(death_30~ser_crt, family = "binomial", data = dta_log)
summary(fit_log_1)
```

```
##
## Call:
## glm(formula = death_30 ~ ser_crt, family = "binomial", data = dta_log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5005  -0.4824  -0.4579  -0.4422   2.1788
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.5695     0.2829  -9.083  < 2e-16 ***
## ser_crt       0.3670     0.1300   2.822  0.00477 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 214.63  on 293  degrees of freedom
## Residual deviance: 206.92  on 292  degrees of freedom
## AIC: 210.92
##
## Number of Fisher Scoring iterations: 4
```

```r
fit_log_1$coefficients %>% exp
```

```
## (Intercept)      ser_crt
##  0.07657599   1.44340585
```

```r
confint(fit_log_1) %>% exp # check 95% CI
```

```
## Waiting for profiling to be done...
##                   2.5 %    97.5 %
## (Intercept) 0.04271699 0.1305243
## ser_crt      1.12006891 1.8922657
```

```r
# use lasso to help determine the covariate sets
x <- dplyr::select(dta_log, -death, -time, -death_30, -ser_crt_ab) %>%
        as.matrix()
y <- dta_log$death_30 %>% as.vector()
cv <-  cv.glmnet(x, y, type.measure = "mse", nfolds = 4)
plot(cv)
```
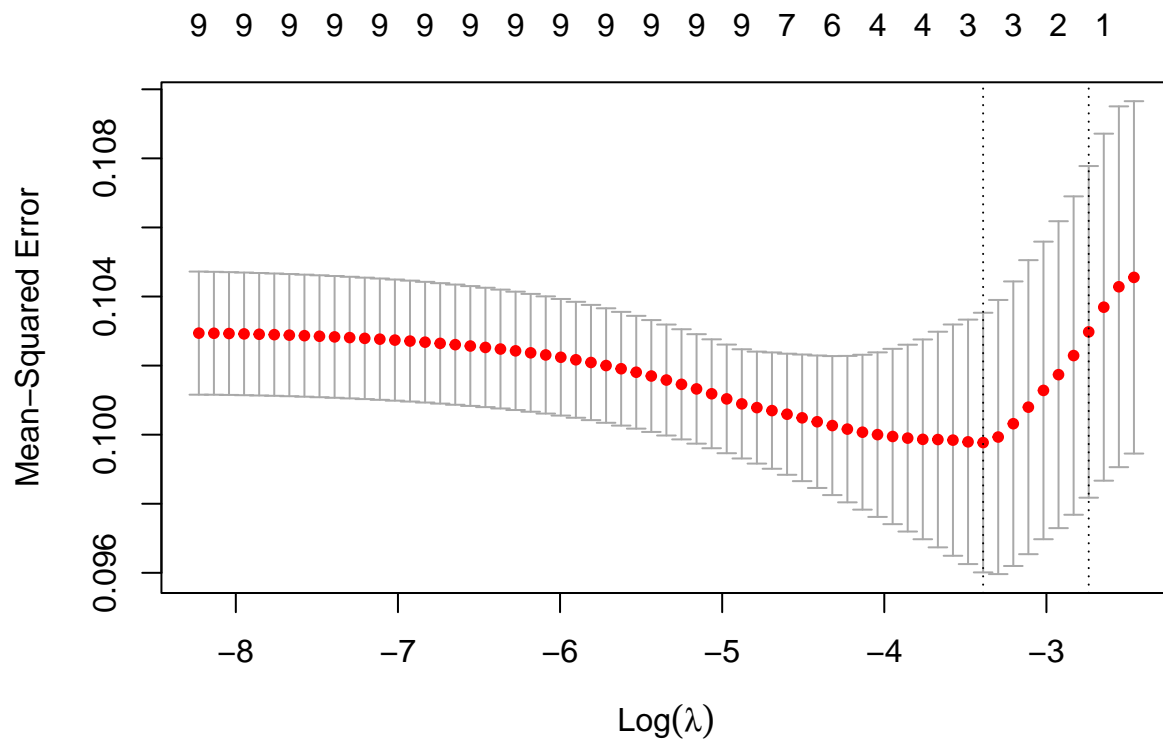


```r
cv$lambda.min
```

```
## [1] 0.03369666
```

```r
cv$glmnet.fit
```

```
##
## Call:  glmnet(x = x, y = y)
##
##      Df  %Dev    Lambda
## 1    0   0.00  0.085430
## 2    1   1.18  0.077840
## 3    1   2.16  0.070930
## 4    1   2.98  0.064630
## 5    1   3.65  0.058890
## 6    2   4.61  0.053650
```

```
## 7   3   5.58 0.048890
## 8   3   6.50 0.044550
## 9   3   7.27 0.040590
## 10  3   7.91 0.036980
## 11  3   8.43 0.033700
## 12  3   8.87 0.030700
## 13  4   9.28 0.027980
## 14  4   9.69 0.025490
## 15  4 10.03 0.023230
## 16  4 10.31 0.021160
## 17  4 10.55 0.019280
## 18  4 10.74 0.017570
## 19  5 10.92 0.016010
## 20  6 11.09 0.014590
## 21  6 11.26 0.013290
## 22  6 11.39 0.012110
## 23  6 11.50 0.011030
## 24  7 11.60 0.010050
## 25  7 11.69 0.009161
## 26  9 11.80 0.008347
## 27  9 11.89 0.007605
## 28  9 11.98 0.006930
## 29  9 12.04 0.006314
## 30  9 12.10 0.005753
## 31  9 12.15 0.005242
## 32  9 12.18 0.004776
## 33  9 12.22 0.004352
## 34  9 12.24 0.003965
## 35  9 12.27 0.003613
## 36  9 12.28 0.003292
## 37  9 12.30 0.003000
## 38  9 12.31 0.002733
## 39  9 12.32 0.002490
## 40  9 12.33 0.002269
## 41  9 12.34 0.002068
## 42  9 12.34 0.001884
## 43  9 12.35 0.001717
## 44  9 12.35 0.001564
## 45  9 12.36 0.001425
## 46  9 12.36 0.001299
## 47  9 12.36 0.001183
## 48  9 12.36 0.001078
## 49  9 12.37 0.000982
## 50  9 12.37 0.000895
## 51  9 12.37 0.000816
## 52  9 12.37 0.000743
## 53  9 12.37 0.000677
## 54  9 12.37 0.000617
## 55  9 12.37 0.000562
## 56  9 12.37 0.000512
## 57  9 12.37 0.000467
## 58  9 12.37 0.000425
## 59  9 12.37 0.000387
## 60  9 12.37 0.000353
```

```
## 61  9 12.37 0.000322
## 62  9 12.37 0.000293
## 63  9 12.37 0.000267
```

```
coef(cv, s = "lambda.min")
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept)  0.299914132
## age          0.004033112
## sex              .
## anemia           .
## dbt              .
## ef               .
## smoking          .
## plat             .
## ser_crt      0.018167964
## ser_na      -0.003308001
```

```
best_cov <- c("age", "anemia", "ef", "plat", "ser_crt", "ser_na")
coef(cv, s = 0.010050)
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept)  7.850451e-01
## age          5.613910e-03
## sex          3.373783e-04
## anemia       3.544657e-02
## dbt              .
## ef          -5.075034e-04
## smoking          .
## plat        -3.718755e-08
## ser_crt      3.344088e-02
## ser_na      -7.619702e-03
```

```
include_cov <- c("sex", "age", "anemia", "ef", "plat", "ser_crt", "ser_na")
coef(cv, s = 0.016010)
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept)  0.6699599884
## age          0.0052104089
## sex              .
## anemia       0.0242794930
## dbt              .
## ef          -0.0000444681
## smoking          .
## plat             .
## ser_crt      0.0296350188
## ser_na      -0.0067224731
```

```
exclude_cov <- c("age", "anemia", "ef",  "ser_crt", "ser_na")
```

According to the lasso results, we will perform three logitstic regressions and compare their results.

```
log_fun <- function(set){
        x_matrix <- dta_log[set,]
```

```
        y <- dta_log$death_30
        fit <- glm(y~x, family = "binomial")
        OR <- coef(fit)["xser_crt"] %>% exp()
        CI <- confint(fit)["xser_crt",] %>% exp()
        aic <- fit$aic
        return(list(OR, CI, aic))
}
lapply(list(best_cov, exclude_cov, include_cov), log_fun)
```

```
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...

## [[1]]
## [[1]][[1]]
## xser_crt
## 1.324913
##
## [[1]][[2]]
##     2.5 %     97.5 %
## 0.9826464 1.7762613
##
## [[1]][[3]]
## [1] 198.6465
##
##
## [[2]]
## [[2]][[1]]
## xser_crt
## 1.324913
##
## [[2]][[2]]
##     2.5 %     97.5 %
## 0.9826464 1.7762613
##
## [[2]][[3]]
## [1] 198.6465
##
##
## [[3]]
## [[3]][[1]]
## xser_crt
## 1.324913
##
## [[3]][[2]]
##     2.5 %     97.5 %
## 0.9826464 1.7762613
##
## [[3]][[3]]
## [1] 198.6465
```
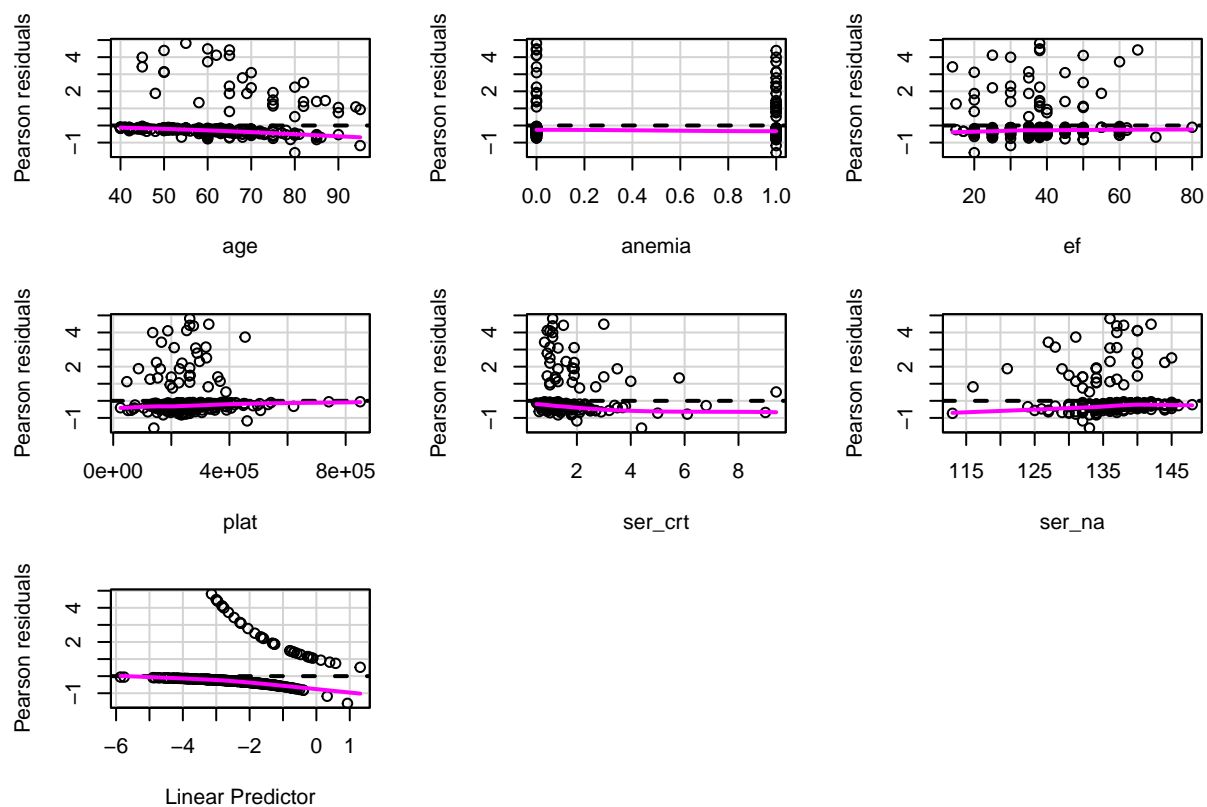
Diagnostics for the model

```
log_fit <- glm(death_30~age+anemia+ef+plat+ser_crt+ser_na, family = "binomial", data = dta_log)
residualPlots(log_fit)
```

```
##          Test stat Pr(>|Test stat|)
## age        2.6482            0.1037
## anemia     0.0000            1.0000
## ef         0.3016            0.5829
## plat       0.1064            0.7443
## ser_crt    0.1761            0.6748
## ser_na     0.7059            0.4008
```
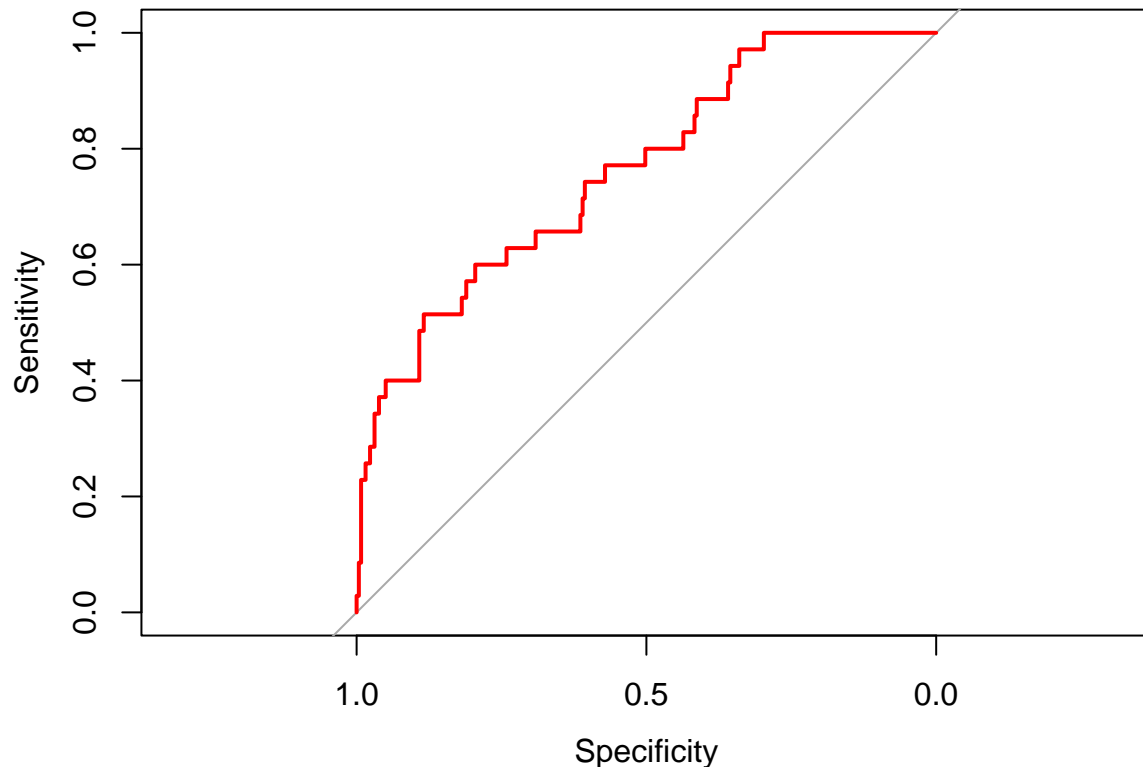
Discrimination

```
library(pROC)
predprob <- predict(log_fit,type=c("response"))
roccurve <- roc(dta_log$death_30 ~ predprob)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(roccurve,col="red")
```

```
auc(roccurve)
```

```
## Area under the curve: 0.7629
```

Poisson Regression

We will collapse the individual observations into total groups.

```
dta$age_65 <- if_else(dta$age >= 65, 1, 0)
dta$ef_group <- case_when(dta$ef <= 30 ~ "low",
                          dta$ef >30 & dta$ef < 45 ~ "normal",
                          dta$ef >=45 ~ "high")

dta$plat_group <- if_else(dta$plat >= median(dta$plat), "high", "low")
dta$ser_na_group <- if_else(dta$ser_na >= median(dta$ser_na), "high", "low")

dta_pos <- dta %>% group_by(age_65, sex, ef_group, plat_group, anemia, ser_na_group, ser_crt_ab) %>%
        summarise(death = sum(death),
                  time = sum(time))
```

```
## `summarise()` has grouped output by 'age_65', 'sex', 'ef_group', 'plat_group', 'anemia', 'ser_na_grou
```

```
names(dta_pos)
```

```
## [1] "age_65"       "sex"          "ef_group"     "plat_group"   "anemia"
## [6] "ser_na_group" "ser_crt_ab"   "death"        "time"
```

```
pois_fit <- glm(death~ser_crt_ab  + sex+ age_65+ef_group+plat_group+anemia + ser_na_group, offset = log
dim(dta_pos)
```

```
## [1] 126    9
```

```
pois_fit %>% summary()
```

```
##
## Call:
## glm(formula = death ~ ser_crt_ab + sex + age_65 + ef_group +
##     plat_group + anemia + ser_na_group, family = poisson(), data = dta_pos,
##     offset = log(time))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1568  -0.9484  -0.4188   0.7748   2.5915
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -6.9959     0.3348 -20.898  < 2e-16 ***
## ser_crt_ab         0.9566     0.2210   4.328  1.5e-05 ***
## sex               -0.1124     0.2156  -0.521 0.602233
## age_65             0.7562     0.2058   3.675 0.000238 ***
## ef_grouplow        0.6838     0.2849   2.400 0.016377 *
## ef_groupnormal    -0.2848     0.3050  -0.934 0.350550
## plat_grouplow     -0.1605     0.2080  -0.771 0.440472
## anemia             0.5014     0.2121   2.364 0.018092 *
## ser_na_grouplow    0.2712     0.2311   1.174 0.240547
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 230.62  on 125  degrees of freedom
## Residual deviance: 159.13  on 117  degrees of freedom
## AIC: 329.94
##
## Number of Fisher Scoring iterations: 6
```

```
coef(pois_fit) %>% exp()
```

```
##      (Intercept)       ser_crt_ab              sex          age_65     ef_grouplow
##     0.0009156486     2.6028034324     0.8937042734     2.1301637981     1.9813948239
##   ef_groupnormal    plat_grouplow           anemia ser_na_grouplow
##     0.7521940404     0.8517420535     1.6511077859     1.3114984728
```

```
confint(pois_fit) %>% exp()
```

```
## Waiting for profiling to be done...
```

```
##                      2.5 %       97.5 %
## (Intercept)     0.0004611462 0.001716743
## ser_crt_ab      1.6823610137 4.009301799
## sex             0.5897161610 1.377366846
## age_65          1.4227862654 3.196016045
## ef_grouplow     1.1520367551 3.539387224
## ef_groupnormal  0.4157020757 1.386028096
## plat_grouplow   0.5658328669 1.282171634
```

```
## anemia            1.0869053000 2.502671572
## ser_na_grouplow 0.8357405639 2.072758933
```

```
residualPlots(pois_fit)
```
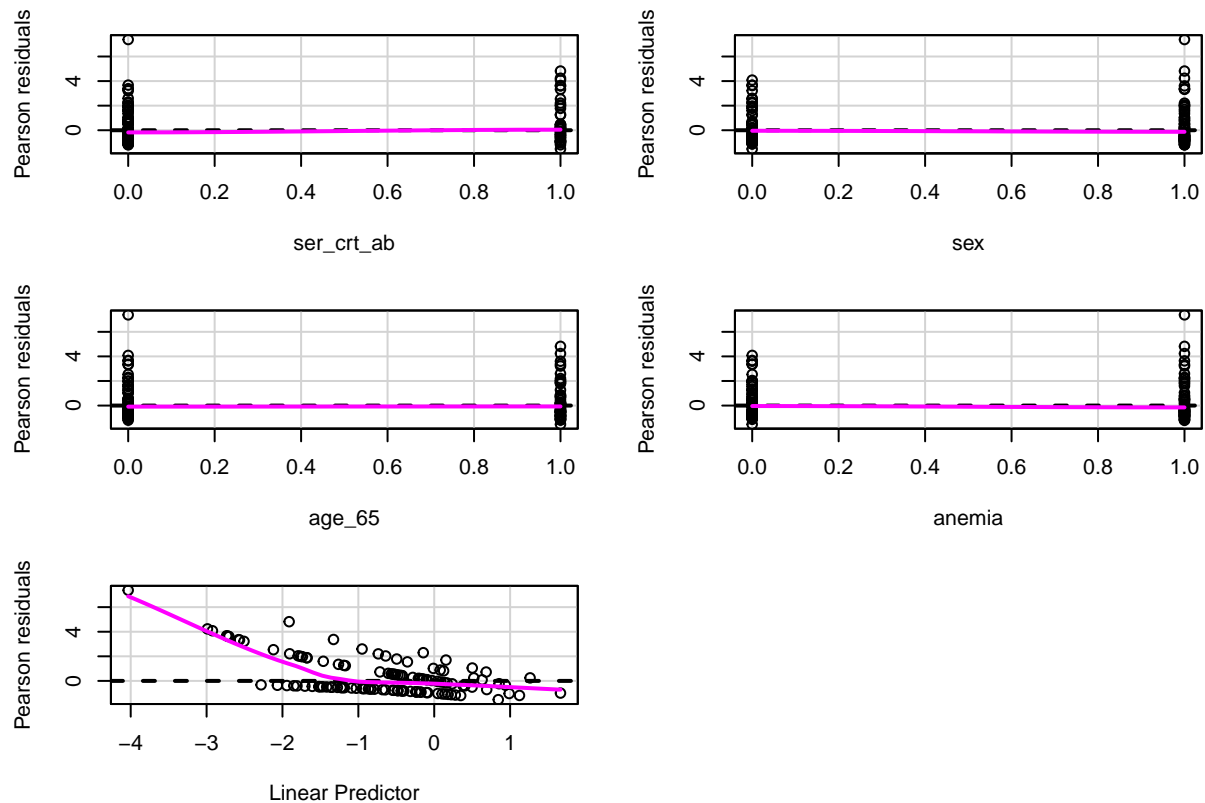


```
##              Test stat Pr(>|Test stat|)
## ser_crt_ab         0                1
## sex                0                1
## age_65             0                1
## anemia             0                1
```

For a quasi-Poisson

```
quai_pois_fit <- glm(death~ser_crt_ab  + sex+ age_65+ef_group+plat_group+anemia + ser_na_group, offset =
quai_pois_fit %>% summary()
```

```
##
## Call:
## glm(formula = death ~ ser_crt_ab + sex + age_65 + ef_group +
##     plat_group + anemia + ser_na_group, family = quasipoisson(),
##     data = dta_pos, offset = log(time))
##
## Deviance Residuals:
##     Min       1Q    Median       3Q       Max
## -2.1568  -0.9484   -0.4188    0.7748    2.5915
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      -6.9959     0.5435 -12.872  < 2e-16 ***
## ser_crt_ab         0.9566     0.3588   2.666  0.00876 **
## sex               -0.1124     0.3501  -0.321  0.74877
## age_65             0.7562     0.3341   2.263  0.02546 *
## ef_grouplow        0.6838     0.4625   1.478  0.14197
## ef_groupnormal    -0.2848     0.4953  -0.575  0.56641
## plat_grouplow     -0.1605     0.3377  -0.475  0.63559
## anemia             0.5014     0.3444   1.456  0.14810
## ser_na_grouplow    0.2712     0.3751   0.723  0.47121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 2.636011)
##
##      Null deviance: 230.62  on 125  degrees of freedom
## Residual deviance: 159.13  on 117  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

```
coef(quai_pois_fit) %>% exp()
```

```
##      (Intercept)        ser_crt_ab                  sex          age_65      ef_grouplow
##     0.0009156486     2.6028034324      0.8937042734    2.1301637981     1.9813948239
##   ef_groupnormal    plat_grouplow               anemia ser_na_grouplow
##     0.7521940404     0.8517420535      1.6511077859    1.3114984728
```

```
confint(quai_pois_fit) %>% exp()
```

```
## Waiting for profiling to be done...
```

```
##                         2.5 %       97.5 %
## (Intercept)      0.0002910743 0.002474161
## ser_crt_ab       1.2749822700 5.251777025
## sex              0.4571383160 1.826980824
## age_65           1.1035485975 4.135106963
## ef_grouplow      0.8316029813 5.224560757
## ef_groupnormal   0.2866752941 2.071872562
## plat_grouplow    0.4369758856 1.660372661
## anemia           0.8337866477 3.250599757
## ser_na_grouplow  0.6312388771 2.777590426
```

8. Intentions

At this stage, we have not prepared to publish this project. Now we mainly focus on making better analysis. If our work is qualified enough, we will consider format it into a publishable manuscript.