AMS 572 Term Project

Professor: Dr. Pei Fen Kuan

Report By: Ji Hun Kim, Chi-Sheng Lo,
Sam van der Poel, Jinglin Yang

Due: Thursday, December 2, 2021

1 Introduction

The objective of the project is to conduct regression analysis on the heart failure dataset which has been commonly used to model mortality after taking account of age, ejection fraction, serum creatine, serum sodium, anemia, platelets, creatinine phosphokinase, blood pressure gender, diabetes, and smoking status. In addition to running conventional statistical analysis, we will also simulate two types of missing value methods: missing value completely at random (MCAR) and missing value not at random (MNAR).

2 Dataset

The heart failure clinical dataset was assembled by Davide Chicco at the Krembil Research Institute in Toronto, Canada. We downloaded this dataset from the machine learning repository at UC Irvine [DG17]. This dataset has 299 observations that are patients; in addition, it has 13 mixed continuous and binary predictors, and one binary outcome. Among the 299 patients, 105 are women and 194 are men. The age of every patient must be over 40 years. The continuous variables are age, and creatinine phosphokinase (CPK), ejection fraction (EF), platelets, serum creatinine, serum sodium, and time. The categorical variables are anaemia, diabetes, high blood pressure, sex, and smoking. More details about the variables are summarized in Table 1.

Variable	Description [Range]
Age	Age of patient [40, 95]
CPK	Level in the blood [0, 1]
EF	% blood leaving heart at each contraction [14, 80]
Platelets	Amount in the blood [25.01, 850.00]
Serum creatinine	Level in the blood [0.50, 9.40]
Serum sodium	Level in the blood [114, 148]
Time	Follow-up period [4, 285]
Anaemia	Decrease of blood cells [0, 1]
Diabetes	Whether patient has diabetes [0, 1]
Death event	Whether patient died during the follow-up period [0, 1]
High blood pressure	Whether patient has hypertension [0, 1]
Sex	Gender [0, 1]
Smoking	Whether the patient smokes [0, 1]

Table 1: Variables in Heart Failure Data Set

3 Literature Review

The heart failure clinical records dataset was collected in 2015 and has been analyzed in several publications such as Ahmad et al. [AMB$^+$17] and Chicco and Jurman [CJ20]. Both papers used the survival analysis but with different

statistical approaches. Ahmad et al. [AMB+17] adopted cox regression to conduct the study on the survival analysis of a sample of 299 heart failure patients in Pakistan and maintained that the high risk of death among heart failure patients can be attributed to growing age, renal dysfunction, high BP, high anaemia and low ejection fraction. On the contrary, they also found that high level of serum sodium can reduce the likelihood of death. Chicco and Jurman [CJ20] implemented the machine learning binary classification method to make prediction on the survival of patients and to rank risk factors. They suggested that serum creatinine and ejection fraction are the most crucial in determining the risk of mortality of heart failure patients

In the later sections, we will be doing missing data analysis with MCAR and MNAR methods. Han [Han18] pointed out that missingness in MNAR depends on both the observed and the missing value and missingness in MCAR depends on neither the observed nor the missing values. There have been many articles such as Haitovsky [Hai68], Rubin [Rub76], and Efron [Efr94] which have made influential contributions on the theoretical aspects of missing data. Since our second test runs with the GLM, we also found some relevant references in regarding the missing data with the GLM. For instance, Ibrahim et al [ICLH05] reviewed four approaches including maximum likelihood (ML), multiple imputation (MI), fully Bayesian (FB), and weighted estimating equations (WEES) for GLM with missing data.

4 Test 1: Two Sample Mean Test

4.1 Normality Assumption

We will test the null hypothesis that the average serum sodium levels are the same for subjects who were deceased during the trial and subjects who were not deceased during the trial. We first test the normality assumption to choose the test method. The most common ways to examine the normality are QQ plot (quantile-quantile plot) and the Shapiro-Wilk test. The QQ plots are displayed in Figure 1.

Group x is the average serum sodium levels are the same for subjects who were not deceased and group y is the subjects who were deceased.

We also fit the Shapiro-Wilk test model

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{1}$$

The test results are following:

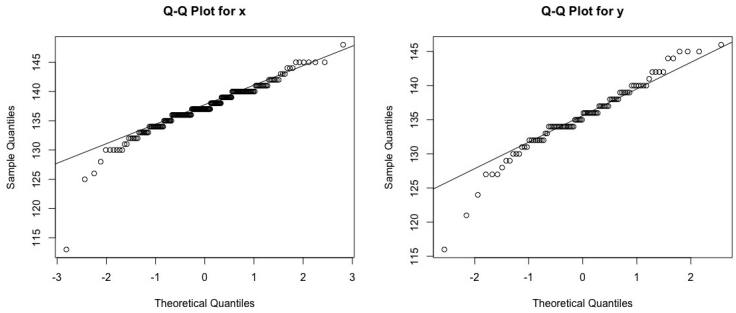

Figure 1: QQ Plots for Serum Sodium Samples

x		y	
Statistics (W)	p-value	Statistics (W)	p-value
0.92477	1.088×10^{-8}	0.95821	0.003841

Table 2: Results of Shapiro-Wilk Test

We can assume that the data is normally distributed if the data points on QQ plot fall along the reference line and the p-value of Shapiro-Wilk test is larger than the significance level. The interpretation of figure 1 above shows that the data points of each group does not fit the reference line at the first and the last some quantiles. In addition, the results of Shapiro-Wilk test of the average Serum Sodium levels of the both groups calculated p-value as 0.003841 and $1.088e-08$. Since p-values of both groups are smaller than the level of significance, we reject H_0 and conclude that they do not follow the normal distribution. Therefore, Wilcoxon Rank Sum Test should be used to do a two-sample mean test using data that are not normally distributed.

4.2 Wilcoxon Rank Sum Test

Normality of the data is assumed and each group is randomly chosen independent sample. We will test whether the level of Serum Sodium is a significant predictor of the death events by comparing the mean of each group. Then the test hypothesis is,

$$H_0 : \mu_x = \mu_y \qquad \text{vs.} \qquad H_a : \mu_x \neq \mu_y,$$

where μ_x be the mean of the Serum Sodium levels of group x and μ_y be the mean of the Serum Sodium levels of group y. The test follows 95% significance level. In the summary of the test, p-value = 0.00029. Since p-value is smaller than significance level = 0.05, we reject H_0. Thus, there is a significant evidence that the average levels of Serum Sodium are different in each group. The test result indicates that Serum Sodium levels is a significant indicator of heart failure.

5 Test 2: Multiple Logistic Regression

We fit a multiple logistic regression model

$$\log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{12} x_{12}, \tag{2}$$

where x_1 is the age variable, x_2 is the anaemia indicator variable, and so on. We will test whether the variable serum_creatinine is a significant predictor of the variable DEATH_EVENT in model (2), that is,

$$H_0 : \beta_8 = 0 \qquad \text{vs.} \qquad H_a : \beta_8 \neq 0,$$

where β_8 is the coefficient corresponding with serum_creatinine. In the summary of this model, R reports that $\hat{\beta}_8 = 0.6661$ with estimated standard error 0.1815. It follows that $(0.1986, 1.1336)$ is a 99% confidence interval for β_8, and since this interval does not contain zero, we reject $H_0 : \beta_8 = 0$ at the $\alpha = 0.01$ level of significance. The interpretation is that the level of serum creatinine is a statistically significant predictor of death occurrence in model (2). Based on the confidence interval for β_8, we say with 99% certainty that a one-unit increase in serum_creatinine increases the odds of death by a multiplicative factor between $e^{0.1986} = 1.2197$ and $e^{1.1336} = 3.1068$.

The type III analysis of deviance table associated with model (2) is given in Table 5. The likelihood ratio chi-squared statistic associated with variable serum_creatinine reported in Table 5 is 14.024 with p-value 0.00018, providing additional confidence that $\beta_8 \neq 0$.

Variable	LR χ^2	df	p-value
age	9.820	1	0.0017264
anaemia	0.000	1	0.9834653
creatinine_phosphokinase	1.775	1	0.1827789
diabetes	0.171	1	0.6793508
ejection_fraction	27.146	1	1.887×10^{-7}
high_blood_pressure	0.082	1	0.7743756
platelets	0.411	1	0.5212489
serum_creatinine	14.024	1	0.0001805
serum_sodium	2.845	1	0.0916393
sex	1.684	1	0.1943846
smoking	0.001	1	0.9739150
time	74.727	1	$< 2.2 \times 10^{-16}$

Table 3: Type III Analysis of Deviance Table

6 Missing Data

6.1 MCAR

The definition of MCAR, Missing Completely At Random, is that the probability of being missing is the same for all case. This hypothesis implies that the causes of missing data are unrelated to the data, which means the missing data is not dependent of the data itself. Therefore, employing a discrete uniform distribution to pick 50 missing data is viable.Then we choose to ignore the missing observation and proceed the first test.

x		y	
Statistics (W)	p-value	Statistics (W)	p-value
0.92314	9.323×10^{-9}	0.95368	0.002175

Table 4: Results of Shapiro-Wilk Test with MCAR (Ignoring missing data)

Though the statistics changes for Shapiro-Wilk test, the conclusion remains the same at significant level = 0.05 because p-values are still small.In this MCAR scenario, we can say that MCAR and ignoring the missing data don't have much impact on our conclusion.

6.2 MNAR

MNAR, Missing Not At Random, implies that the causes of missing data might be related to the data it self. In this data set, we presume that the probability of missing data is related to the death event. In particular, those who are dead might possess higher probability to missing data than those who survive. This makes sense, because some data for the survival might not be more available than that for the dead. We let the probability of missing a dead person's data to be 0.3 and that of missing a survived person's data to be 0.1. We implemented data imputation using random forests supplied by the imputeMissings R package. Thereby, we can then proceed to run test 2: the variable serum creatinine is a significant predictor of the variable DEATH EVENT:

$$H_0 : \beta_8 = 0 \qquad \text{vs.} \qquad H_a : \beta_8 \neq 0.$$

Variable	LR χ^2	df	p-value
age	5.472	1	0.019326
anaemia	0.120	1	0.728879
creatinine_phosphokinase	0.598	1	0.439246
diabetes	1.278	1	0.258203
ejection_fraction	8.976	1	0.002736
high_blood_pressure	0.010	1	0.918482
platelets	0.570	1	0.450325
serum_creatinine	2.496	1	0.114112
serum_sodium	2.010	1	0.156306
sex	0.057	1	0.811625
smoking	0.151	1	0.697258
time	35.193	1	$< 2.987 \times 10^{-9}$

Table 5: Type III Analysis of Deviance Table with MNAR(Imputation)

Nevertheless, the p-value for serum creatinine is 0.1141, so we fail to reject the null hypothesis at the $\alpha = 0.05$ level of significance. This implies that our MNAR data after imputation doesn't suggest that serum creatinine is an indicator of death event which is contrary to our previous conclusion. The reason is that our method of choosing MNAR makes more data of the dead to be NA than that of the survival; moreover, the imputation does not affect the distribution significantly. Some extreme values of serum creatinine might be substituted with mean or median. The data of the survival is overwhelmed in the new test since only 10 percent of data has changed. Therefore, we make the reverse conclusion compared with that in section 5.

7 Conclusion

Our project is based on the heart failure clinical dataset which contains continuous and categorical data from 299 patients. There are two main objectives. The first objective is to determine whether there is significant relationship between the serum creatinine level or serum sodium and heart failure death. The second objective is to conduct the same regression method under two kinds of missing data schemes: MCAR and MNAR. We first examine the normality assumption with the two sample mean test on the serum sodium levels for both patients who passed away and patients who survived during the trail; since they are not normally distributed, we then move on to the Wilcoxon rank sum test and we find that the serum sodium level is a significant indicator of heart failure death. Moreover, in the next multiple regression with the logistic model, we still can confirm that the serum creatinine level is a significant predictor of the death event. Thereafter, we focus on the missing data analysis which covers both MCAR and MNAR. In MCAR, we randomly pick 50 missing data based on discrete uniform distribution and we find that the conclusion remains intact

from section 4. In MNAR, we assume that the probability of missing data and death event are linked; in addition, we use the imputation method to fill in missing values with estimates obtained from random forest models (imputeMissings package). Interestingly, MNAR turns out to reverse our previous conclusion in test 2 and states that serum creatinine is not significant enough to be an indicator of death event. This result is caused by the imputation which replaces extreme values with mean or median.

In the future, we can extend this research by incorporating a mixture of causal inference technique and machine learning and see whether the conclusion is consistent.

References

[AMB+17] Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza. Survival analysis of heart failure patients: A case study. *PloS one*, 12(7):e0181001, 2017.

[CJ20] Davide Chicco and Giuseppe Jurman. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, 20(1):1–16, 2020.

[DG17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[Efr94] Bradley Efron. Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89(426):463–475, 1994.

[Hai68] Yoel Haitovsky. Missing data in regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(1):67–82, 1968.

[Han18] Peisong Han. Calibration and multiple robustness when data are missing not at random. *Statistica Sinica*, 28(4):1725–1740, 2018.

[ICLH05] Joseph G Ibrahim, Ming-Hui Chen, Stuart R Lipsitz, and Amy H Herring. Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346, 2005.

[Rub76] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

8 Appendix: R Code

```r
library(car) # Type III analysis
library(imputeMissings) # data imputation

# PREPARE DATA
data = read.csv('heart_failure.csv', header = TRUE)
data$DEATH_EVENT = as.factor(data$DEATH_EVENT)
data$anaemia = as.factor(data$anaemia)
data$diabetes = as.factor(data$diabetes)
data$high_blood_pressure = as.factor(data$high_blood_pressure)
data$sex = as.factor(data$sex)
data$smoking = as.factor(data$smoking)


# TEST 1
x = data[data$DEATH_EVENT==0,c('serum_sodium')]
y = data[data$DEATH_EVENT==1,c('serum_sodium')]

# Test of normality
par(mfrow=c(1,2))
qqnorm(x, main = 'Q-Q Plot for x')
qqline(x)
qqnorm(y, main = 'Q-Q Plot for y')
qqline(y)
shapiro.test(x)
shapiro.test(y)

# Wilcoxon test
wilcox.test(x, y, alternative = "two.sided")


# TEST 2
model = glm(DEATH_EVENT ~ ., data = data, family=binomial('logit'))
summary(model)
Anova(model, type="III")


# MCAR
m = 50
for (i in sample(1:299, size=m, replace=FALSE)) {
     data[i, sample(names(data), size=1, replace=FALSE)] = NA
}
# Now repeat TEST 1 and TEST 2 to check difference in results


# MNAR
# Missing values in variable DEATH_EVENT are more likely when
# DEATH_EVENT=1
data.mnar = data
for (i in 1:nrow(data.mnar)) {
     r = runif(1)
     if (r < 0.1 && data.mnar[i, c('DEATH_EVENT')] == 0) {
          data.mnar[i, c('DEATH_EVENT')] = NA
     } else if (r < 0.3 && data.mnar[i, c('DEATH_EVENT')] == 1) {
          data.mnar[i, c('DEATH_EVENT')] = NA
     }
}
miss.0 = data$DEATH_EVENT == 0 & is.na(data.mnar$DEATH_EVENT)
```

```r
miss.1 = data$DEATH_EVENT == 1 & is.na(data.mnar$DEATH_EVENT)
num.miss.0 = sum(miss.0)
num.miss.1 = sum(miss.1)

# Proportion of subjects with DEATH_EVENT == 0  with missing data
sum(miss.0) / sum(data$DEATH_EVENT == 0)
# Proportion of subjects with DEATH_EVENT == 1  with missing data
sum(miss.1) / sum(data$DEATH_EVENT == 1)

# Impute missing values
data.impute = impute(data)

# Proportion of subjects with DEATH_EVENT=0 correctly classified by
# imputation
sum(miss.0 & data.impute$DEATH_EVENT == 0) / num.miss.0

# Proportion of subjects with DEATH_EVENT=1 correctly classified by
# imputation
sum(miss.1 & data.impute$DEATH_EVENT == 1) / num.miss.1

# Now repeat TEST 1 and TEST 2 to check difference in results
```