ICMR* in patients under 18 years old

Author: MD Hung Dung Van **Start date** : July 17, 2020 **Last Update**: August 16, 2020

*ICMR - Isolated Congenital Mitral Regurgitation

Design Study Objectives

Population: N = 119 patients under 18 years old with either a CE ring implant, a Band implant or neither of those implants.

Theories:

- For patients with **< 26mm (small size) CE ring**, there is a higher chance of relapse (which leads to re-operation/redo) caused by mitral stenosis (MS).
- For patients with **band**, there is a higher chance of relapse caused by mitral regurgiation (MR)

Aim:

- to determine whether the above are true by comparing the MS rate and the MR rate between the Band and CE ring groups in the population.
- to caculate the relapse (MS or MR) probability by sex.

Methods used :

- Multinomial Logistic regression
- Survival analysis

Descriptive Statistics

Flow chart of the study

About the data

The data set contains variables on 119 patients. The outcome variable is `Cause_of_redo` (qualitative information).

- 3 levels
  - "NONE"
  - "MS"
  - "MR"

The data set has a total of 45 variables (after processing). Some of the predictor variables which are going to be used in the multinomial regression model are:

1. **Age_group**, four-level categorical variable describing the age groups:
    - 0-4 yrs
    - 5-9 yrs
    - 10-14 yrs
    - 16-18 yrs
2. **GROUP**, three-level categorical variable describing the three groups of interest, patients with band, patients with CE ring and patients with neither:
    - NONE
    - BAND
    - RING
3. **Ring_group**, three-level categorical variable describing the CE ring sub groups, patients with small size ring (<26mm), patients with large size ring (>26mm) and patients without CE ring:
    - NONE
    - <26mm
    - $\geq$ 26mm

Data Visualization

Multinomial Logistic Regresssion

Package used: `nnet`

Step 1: Relevel the baseline groups

```
data$Cause_of_redo <- factor(data$Cause_of_redo, levels = c("NONE", "MR", "MS"))
data$Age_group <- factor(data$Age_group, levels = c('0-4 yrs', '5-9 yrs', '10-14 yrs', '15-18 yrs'))
data$GROUP <- factor(data$GROUP, levels = c("NONE", "BAND", "RING"))
data$Ring_Group <- factor(data$Ring_Group, levels = c("NONE", "<26", ">=26"))
```

Step 2: Split the data into random training and testing sets.

The model will attempt to learn the relationship on the training data and be evaluated on the test data. In this case, 80% of the data is used for training and 20% for testing.

```
set.seed(12)
train <- sample_frac(data, 0.8)
sample_id <- as.numeric(rownames(train))
test <- data[-sample_id,]
```

Step 3: Fit the model and obtain the results

```
model <- multinom(Cause_of_redo ~ Age_group + Sex + GROUP + Ring_Group, data = train)
```

```
summary(model)$coefficients
```

```
##    (Intercept) Age_group5-9 yrs Age_group10-14 yrs Age_group15-18 yrs      SexM
## MR   -28.79478         1.386831          -25.70965          -18.36790 0.1247472
## MS   -33.80229         2.572523          -32.35526          -13.59761 1.6581460
##    GROUPBAND GROUPRING Ring_Group<26 Ring_Group>=26
## MR  26.53103  17.15680      8.551873       8.604928
## MS -11.91916  20.27758     11.654593       8.622985
```

Step 4: Find the p-value of each coefficient

```
z <- summary(model)$coefficients/summary(model)$standard.errors
p <- (1 - pnorm(abs(z), 0, 1)) * 2
p
```

```
##    (Intercept) Age_group5-9 yrs Age_group10-14 yrs Age_group15-18 yrs
## MR           0       0.17945973                  0                  0
## MS           0       0.01161087                  0                  0
##          SexM GROUPBAND GROUPRING Ring_Group<26 Ring_Group>=26
## MR 0.90353658         0         0             0              0
## MS 0.09261432         0         0             0              0
```

Comment: Overall all coefficients (except for **SexM** variable and **Age_group5-9**) are statistically significant based on their p-values being smaller than 0.05. The model equation for the first row is then:

For MR patients:

- β_{15} The log odds of having to re-operate due to **MR** vs. having not to re-operate will increase by 26.53103 if moving from **GROUP = "None"** to **GROUP = "Band"**.
- β_{16} The log odds of having to re-operate due to **MR** vs. having not to re-operate will increase by 17.15680 if moving from **GROUP = "None"** to **GROUP = "Ring"**.
- β_{17} The log odds of having to re-operate due to **MR** vs. having not to re-operate will increase by 8.551873 if moving from **Ring_grp = "None"** to **Ring_grp = "<26 mm"**.
- β_{18} The log odds of having to re-operate due to **MR** vs. having not to re-operate will increase by 8.604928 if moving from **Ring_grp = "None"** to **Ring_grp = ">=26 mm"**.

The model equation for the second row is:

For MS patients:

- β_{25} The log odds of having to re-operate due to **MS** vs. having not to re-operate will decrease by 11.91916 if moving from **GROUP = "None"** to **GROUP = "Band"**.
- β_{26} The log odds of having to re-operate due to **MS** vs. having not to re-operate will increase by 20.27758 if moving from **GROUP = "None"** to **GROUP = "Ring"**.
- β_{27} The log odds of having to re-operate due to **MS** vs. having not to re-operate will increase by 11.654593 if moving from **Ring_grp = "None"** to **Ring_grp = "<26 mm"**.
- β_{28} The log odds of having to re-operate due to **MS** vs. having not to re-operate will increase by 8.622985 if moving from **Ring_grp = "None"** to **Ring_grp = ">=26 mm"**.

For both MR and MS,

- β_{11} & β_{21} The log odds of having to re-operate vs. having not to re-operate will increase if moving from **age = "0-4"** to **age = "5-9"**.
- β_{12} & β_{22} The log odds of having to re-operate vs. having not to re-operate will decrease if moving from **age = "0-4"** to **age = "10-14"**.
- β_{13} & β_{23} The log odds of having to re-operate vs. having not to re-operate will decreaseif moving from **age = "0-4"** to **age = "15-18"**.

Comments: Based on the above statistical findings, it is safe to assume that:

- For patients with **< 26mm (small size) CE ring**, there is a higher chance of relapse caused by mitral stenosis (MS).

- For patients with **band**, there is a higher chance of relapse caused by mitral regurgiation (MR)

Relative Risk

The ratio of the probability of choosing one outcome category over the probability of choosing the baseline category is often referred as relative risk. The output coefficients are represented in the log of odds, hence relative risk can be computed by taking the exponential of the intercepts from the linear equation.

```
exp(coef(model))
```

```
##     (Intercept) Age_group5-9 yrs Age_group10-14 yrs Age_group15-18 yrs     SexM
## MR 3.123102e-13         4.002148       6.830311e-12       1.054195e-08 1.132862
## MS 2.088588e-15        13.098830       8.877505e-15       1.243459e-06 5.249569
##       GROUPBAND GROUPRING Ring_Group<26 Ring_Group>=26
## MR 3.328751e+11  28255587      5176.443       5458.495
## MS 6.661528e-06 640383534    115219.358       5557.951
```

A few comments:

- The relative risk ratio switching from **Age_group = 0-4 yrs** to **5-9 yrs** is 4.002148 for redo caused by MR vs. no redo at all.
- The relative risk ratio switching from **Age_group = 0-4 yrs** to **10-14 yrs** 8.88×10^{-9} for redo caused by MS vs. no redo at all.

Step 5: Look at the predicted probabilities

```
head(pp <- fitted(model))
```

```
##        NONE           MR           MS
## 1 1.0000000 3.727184e-13 3.464353e-16
## 2 0.9058297 9.417031e-02 1.260298e-20
## 3 1.0000000 3.727184e-13 3.464353e-16
## 4 0.7061825 2.938175e-01 1.286993e-19
## 5 1.0000000 3.290060e-13 6.599310e-17
## 6 1.0000000 3.727184e-13 3.464353e-16
```

The probability of n^{th} obs being "NONE", "MR" or "MS" shown in the above table can be interpreted in percentage. For example, the probability of the first observation being "NONE" is 100%, it being "MR" is 0.0% and and it being "MS" is 0.0%. Thus we can conclude that the patient from this observation did not have to have a re-operation.

Training vs Test sets

```
train$predicted <- predict(model, newdata = train, "class")

# Building classification table
ctable <- table(train$Cause_of_redo, train$predicted)

# Calculating accuracy - sum of diagonal elements divided by total obs
round((sum(diag(ctable))/sum(ctable))*100, 2)
```

```
## [1] 86.32
```

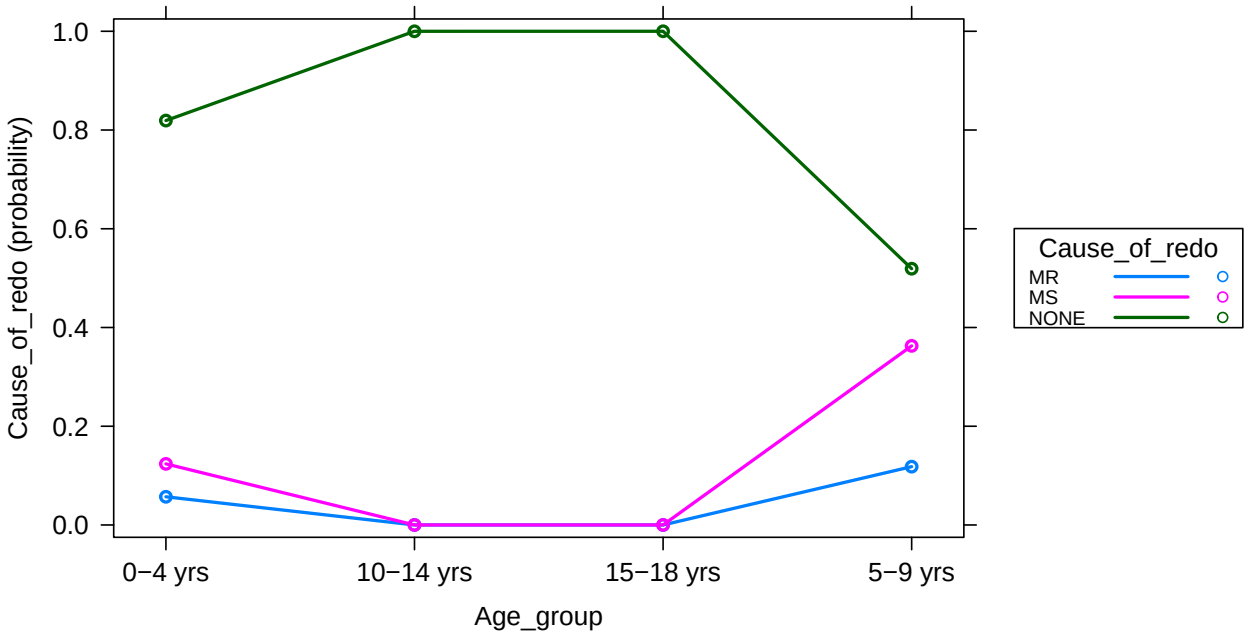
```r
test$predicted <- predict(model, newdata = test, "class")

# Building classification table
ctable2 <- table(test$Cause_of_redo, test$predicted)

# Calculating accuracy - sum of diagonal elements divided by total obs
round((sum(diag(ctable2))/sum(ctable2))*100, 2)
```

```
## [1] 87.5
```

Comment: Accuracy in training dataset is 86.32% and the accuracy of the test set is 87.5% , which is slightly higher than that of the training data. This is not ideal since test accuracy should not be higher than that of training as the model is optimized for the latter. The model performs well overall because it performs well not only on the training data but also on the test (unseen) data.

Summary

Age group effect plot



The plot shows the difference in the average age trajectory between the "NONE", "MR" and "MS" groups, with the fitted response line for the "NONE" group being significantly above the latter. The fitted probability lines for for "MR" and "MS" are identical in 10-14 yrs and 15-18 yrs age groups, and different in 0-4 yrs and 5-9 where the MS line is higher.

Survival Analysis

Packages used:

- `survival`
- `survminer`

Method: Kaplan-Meier non-parametric survival estimate. Kaplan-Meier curves are especially useful when the predictor variable is categorical (e.g.: treatment A vs treatment B; males vs females).

Measures of interest: time to relapse

- status: censoring status 1 = censored, 2 = relapsed
- sex: male = M, female = F
- time: disease-free time in months

Results

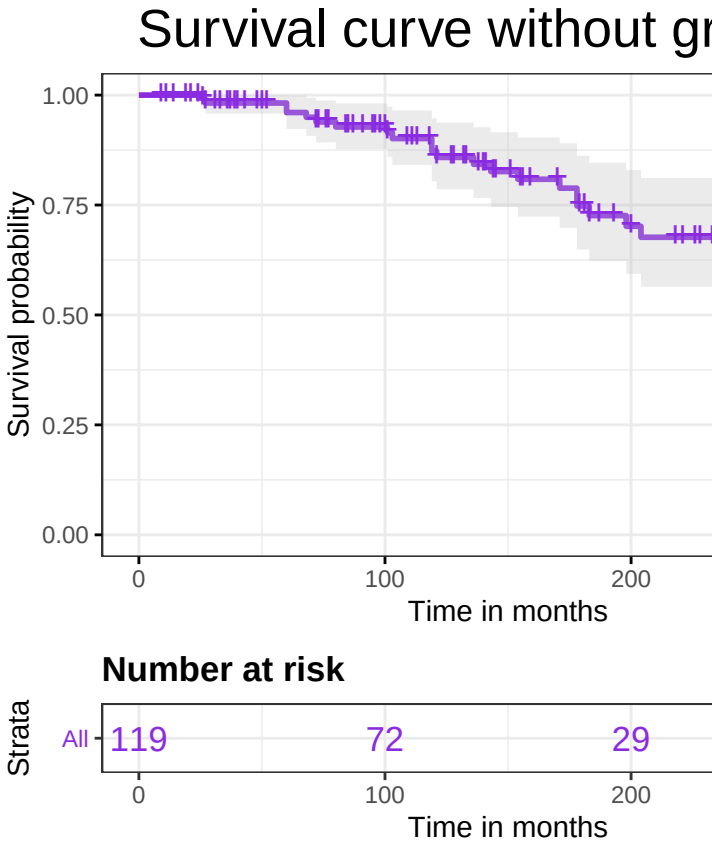
Redo variable is a two level (Yes and No) categorical variable - whether there is a re-operation done for each individual patient. It is used as an indicator to determine if an individual is censored:

$$\text{Redo} \begin{cases} *\textbf{Yes} \text{ to re-operation} : \textbf{not censored} \\ *\textbf{No} \text{ to re-operation} : \textbf{censored} \end{cases}$$

Data with a `plus` sign are censored data and otherwise.

```
##  [1]  26+  31+ 242  240+ 228+  96+  27+  37+ 111+ 154  141+ 179+ 170+ 271+ 237+
## [16] 247+ 226+ 242+ 261+ 270+
```

The following visualizations depict the survival rate (or disease-free rate in this case) without grouping and with grouping.



Visualizing the estimated distribution of survival times

```
summary(fit)$table
```

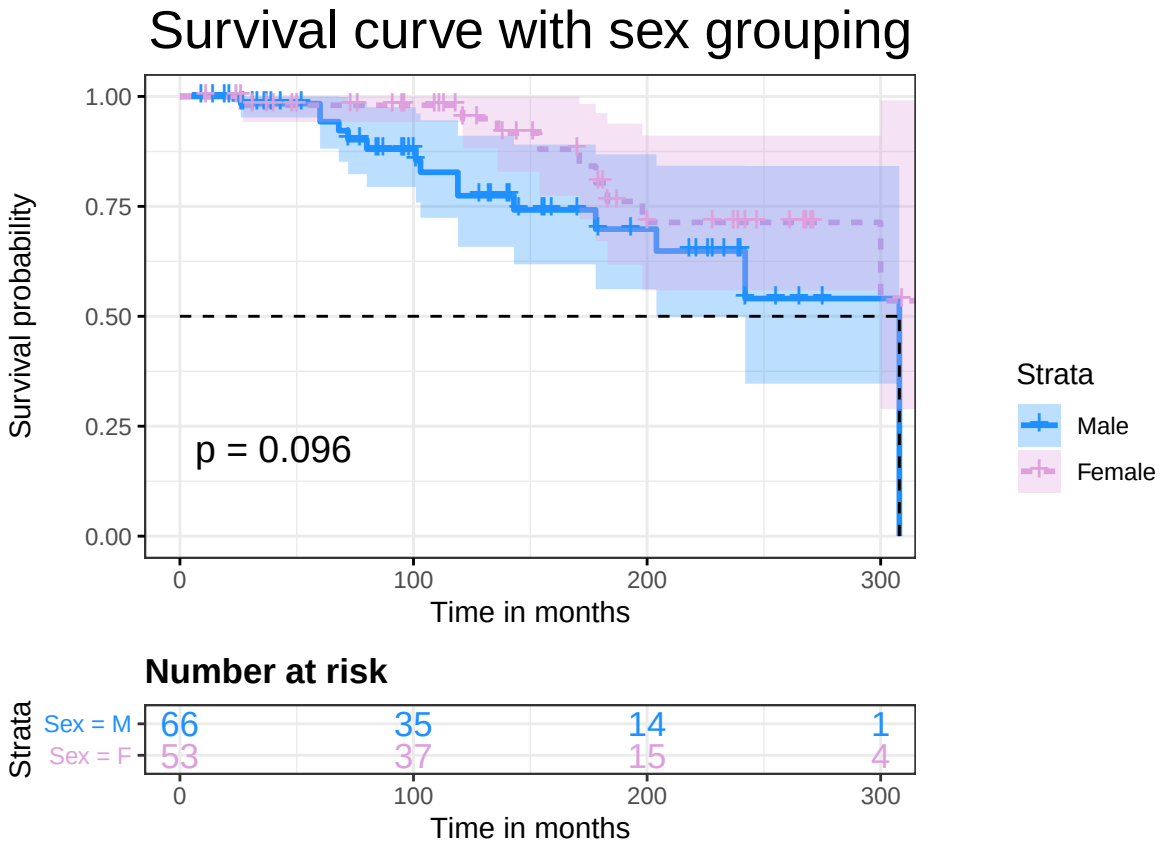
```
##    records      n.max    n.start     events     *rmean *se(rmean)     median
##  119.00000  119.00000  119.00000   24.00000  256.57462   11.71914  308.00000
##    0.95LCL    0.95UCL
##  300.00000         NA
```

Comment: The horizontal axis (x-axis) represents time in months, and the vertical axis (y-axis) shows the probability of having no disease or the proportion of people having no disease. The line represent survival curves of the population. A vertical drop in the curves indicates an event. The vertical tick mark on the curves means that a patient was censored at this time.

- At time zero, the survival probability is 1.0 (or 100% of the participants are disease-free).
- At time 300, the probability of survival is approximately 0.50 (or 50%).
- The median survival time for the population is 308 months.

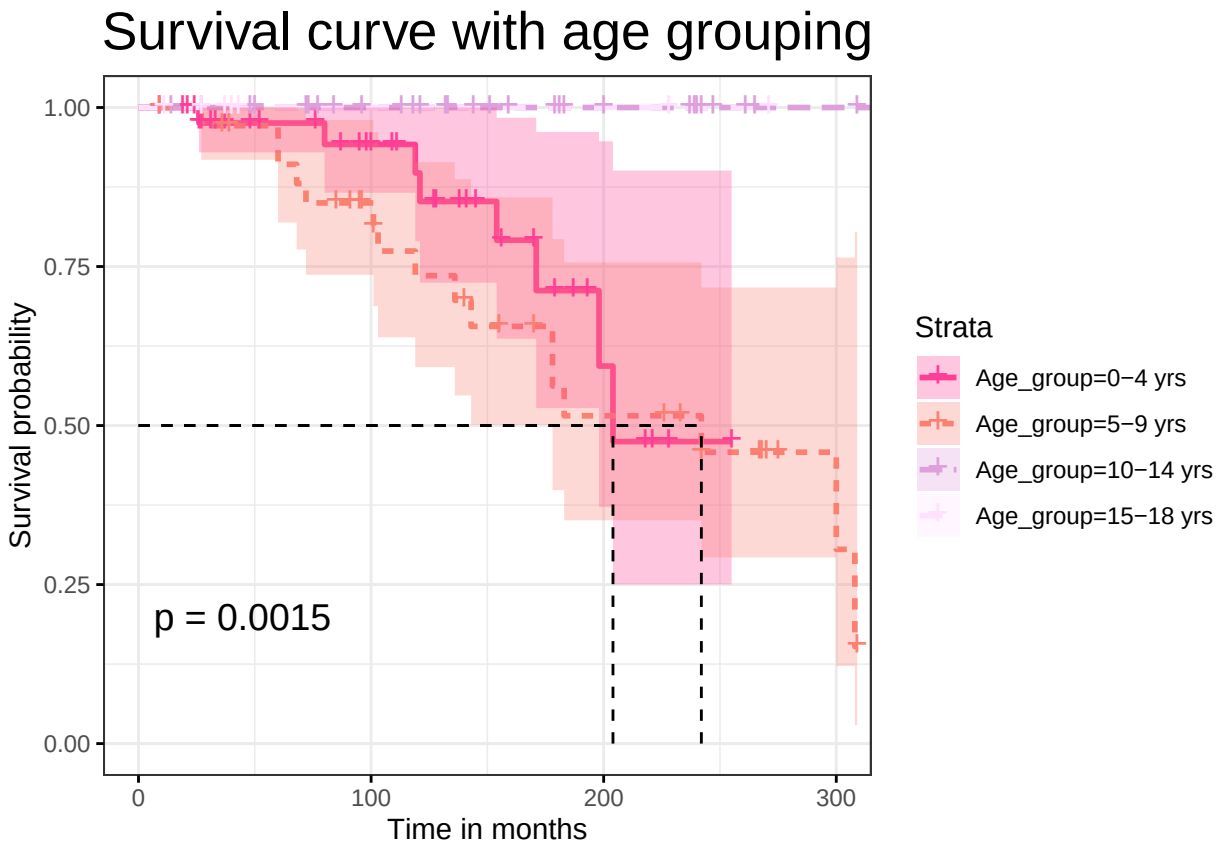
Table of survival analysis showing the first 10 observations

time	n.risk	n.event	n.censor	surv	upper	lower
9	119	0	2	1	1	1
11	117	0	1	1	1	1
14	116	0	1	1	1	1
19	115	0	1	1	1	1
21	114	0	1	1	1	1
24	113	0	1	1	1	1
26	112	1	2	0.9911	1	0.9738
27	109	1	4	0.982	1	0.9575
31	104	0	2	0.982	1	0.9575
33	102	0	1	0.982	1	0.9575



Comment: The log-rank p-value for **Sex** variable is lower than significance level $\alpha = 0.5$ as shown in the plot above. This is also supported by the sex variable's p-value in multinomial logistic regression model.

Therefore Sex is not a significant predictor hence has no effect on the survival/disease-free curve.




```
##                       records n.max n.start events   *rmean *se(rmean) median
## Age_group=0-4 yrs          45    45      45      8 222.4189   18.22248    204
## Age_group=5-9 yrs          36    36      36     16 204.0747   17.20604    242
## Age_group=10-14 yrs        31    31      31      0 290.0000    0.00000     NA
## Age_group=15-18 yrs         7     7       7      0 290.0000    0.00000     NA
##                       0.95LCL 0.95UCL
## Age_group=0-4 yrs         198      NA
## Age_group=5-9 yrs         178      NA
## Age_group=10-14 yrs        NA      NA
## Age_group=15-18 yrs        NA      NA
```

Comment: The log-rank p-value for **Age** variable is statistically significant. The survival curves show the 0-4 yrs age group to have less survival/disease-free advantage in comparison to the other groups. The median survival time for age_group = 0-4 yrs is 204 months, as opposed to 242 months for age_group = 5-9 yrs.

Further analysis to evaluate whether the differences between the age groups are statistically different can be done using a Log-Rank test.

```
surv_diff <- survdiff(Surv(time, status) ~ Age_group, data = data)
surv_diff
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ Age_group, data = data)
##
```

```
##                         N Observed Expected (O-E)^2/E (O-E)^2/V
## Age_group=0-4 yrs      45        8    6.367     0.419     0.608
## Age_group=5-9 yrs      36       16    8.745     6.020     9.730
## Age_group=10-14 yrs    31        0    7.987     7.987    12.143
## Age_group=15-18 yrs     7        0    0.902     0.902     0.950
##
##  Chisq= 15.4  on 3 degrees of freedom, p= 0.001
```

The log rank test for difference in survival gives a p-value of p = 0.001, indicating that the age groups differ significantly in survival.

References

Sharma, Mohit. 2018. *MULTINOMIAL LOGISTIC REGRESSION USING R*. https://datasciencebeginners.com/2018/12/20/multinomial-logistic-regression-using-r/.

n.d. *Survival Analysis Basics*. STHDA. http://www.sthda.com/english/wiki/survival-analysis-basics.

n.d. *MULTINOMIAL LOGISTIC REGRESSION | R DATA ANALYSIS EXAMPLES*. UCLA:Statistical Consulting Group. https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/.