AMS 572 Project

Group 1 - Yichun Hao, Ben Marantz, Paul Vespe, Zhe Zhou

11/18/2021

Import Packages

```
library(dplyr) # summarise() and %>%
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(mice) # deal with missing value  md.pattern() and mice()
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
library(MASS) # glm() function
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(caret) # Data train and  predict
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(pROC) # AUC or Draw ROC curve
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(leaps) # regsubsets()
library(knitr) # kable() and R Markdown conversion
```

Load Data

```r
set.seed(123)
# Data Provided by CDC Case Surveillance Task Force
# Reference: https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4
data.original <- read.csv("COVID-19_Case_Surveillance_Public_Use_Data_with_Geography.csv")
data.original <- data.original[sample(nrow(data.original), 100000), ]

# Convert "Unknown" and "Missing" character values into NA
data.original[data.original == "Unknown"] <- NA
data.original[data.original == "Missing"] <- NA
data.original[data.original == ""] <- NA

# The number of missing values
sum(is.na(data.original))
```

```
## [1] 382729
```

```r
# Modify case month to a numeric variable - case month since start of 2020 (Jan 2020 -> 1)
d <- c()
for (i in data.original$case_month) {
  year <- (as.integer(substr(i, 0, 4)) - 2020) * 12
  year
  month <- as.integer(substr(i, 6, 8))
  month
  d <- append(d, year + month)
}
data.original$case_month <- d

# Inspect distributions
summary(data.original$case_onset_interval)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##  -1.000   0.000   0.000   0.188   0.000  70.000   22328
```

```r
summary(data.original$case_positive_specimen_interval)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
## -66.000   0.000   0.000   0.278   0.000  70.000   19679
```

```
# Both are almost always zero but have clear outliers
# Both also have a lot of missing values

state <- as.factor(data.original$res_state)
str(state)
```

```
##  Factor w/ 43 levels "AK","AL","AZ",..: 31 22 31 29 22 36 34 31 36 26 ...
```

```
summary(state)
```

```
##    AK    AL    AZ    CA    CO    DC    FL    GU    IA    ID    IL    IN    KS 
##    14     1     1     2   123   102     3     3  4441    79     2     1 10683 
##    KY    LA    MA    MI    MN    MO    MS    MT    NC    ND    NE    NH    NJ 
##    59   212     1     1   579    12     8     1 31948    17     1    28   178 
##    NV    NY    OH    OK    PA    PR    RI    TN    TX    UT    VA    VI    VT 
##  3522   310 17197     3 15123     8    41  3038     3 11070     2   129    22 
##    WA    WI    WV    WY  NA's 
##    18     2     3  1005     4
```

```
# Most patients come from a handful of states
# Most states are underrepresented.
```

Process Reported Data Set

```
# Data Provided by National Center for Health Statistics
# Reference: https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku
data.reported <- read.csv("Provisional_COVID-19_Deaths_by_Sex_and_Age.csv")
data.reported <- data.reported[c(3, 7, 9, 11, 13, 15, 16, 17), ]

for (i in rownames(data.reported)) {
  if (data.reported[i, "Age.Group"] == "18-29 years" |
           data.reported[i, "Age.Group"] == "30-39 years" |
           data.reported[i, "Age.Group"] == "40-49 years") {
    data.reported[i, "Age.Group"] = "18 to 49 years"
  }
  else if (data.reported[i, "Age.Group"] == "65-74 years" |
           data.reported[i, "Age.Group"] == "75-84 years" |
           data.reported[i, "Age.Group"] == "85 years and over") {
    data.reported[i, "Age.Group"] = "65+ years"
  }
}

data.reported <- summarise(group_by(data.reported, Age.Group),
                           COVID.19.Deaths = sum(COVID.19.Deaths, na.rm = TRUE),
                           Death.Proportion = COVID.19.Deaths / sum(data.reported$COVID.19.Deaths))
data.reported
```

	Age.Group	COVID.19.Deaths	Death.Proportion
	<chr>	<int>	<dbl>
1	0-17 years	605	0.000790519
2	18 to 49 years	49886	0.065183191
3	50-64 years	139761	0.182617729
4	65+ years	575068	0.751408561

4 rows

```
# Death proportion for each age group
prop.reported <- c(data.reported$Death.Proportion)
```

Part 1 : Do Not Generate Missing Values # Copy Data but Do Not Generate Missing Values

```
data.inc <- data.original
```

First Hypothesis: Are the reported death proportions for varying age groups the same as those in surveillance data?

```
# Group data set by "age_group" variable
data.table <- summarise(group_by(data.inc, age_group),
                        Death = length(which(death_yn == "Yes")),
                        Survive = length(which(death_yn == "No")),
                        Total = sum(Death, Survive))
# Delete NA group
data.table <- data.table[-5, ]
data.table
```

age_group <chr>	Death <int>	Survive <int>	Total <int>
0 - 17 years	0	13954	13954
18 to 49 years	1	43056	43057
50 to 64 years	20	14906	14926
65+ years	983	9469	10452

4 rows

```
# Total Deaths
sum(data.table$Death)
```

```
## [1] 1004
```

```
# Total # of Observations (Not missing death or age group)
sum(data.table$Total)
```

```
## [1] 82389
```

```
# Expected # of deaths by age group
sum(data.table$Death) * prop.reported
```

```
## [1]   0.7936811  65.4439241 183.3481994 754.4141954
```

```
# Expected count for 0-17 group is < 1, so we merge 0-17 and 18-49
prop.reported.final <- c(prop.reported[1] + prop.reported[2], prop.reported[3], prop.reported[4])
obs <- c(data.table$Death[1] + data.table$Death[2], data.table$Death[3], data.table$Death[4])
prop.reported.final
```

```
## [1] 0.06597371 0.18261773 0.75140856
```

```
obs
```

```
## [1]   1  20 983
```

```
# Perform the chi-square test
# Null Hypothesis: the reported death proportions are the same as those in surveillance data
# Alternative Hypothesis: the reported death proportions are not the same as those in surveillance data
chisq.test(x = obs, p = prop.reported.final)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obs
## X-squared = 279.04, df = 2, p-value < 2.2e-16
```

```
# According to the result of the test, we are able to reject the null hypothesis since p-value is less than 5%.
# Hence, the reported death proportions for each age group from NCHS is significantly different from that of
# the surveillance data.
```

First Hypothesis: Is the true death rate of surveyed U.S. COVID patients equal to the reported U.S. COVID death rate?

```
# According to the data from Johns Hopkins Coronavirus Resourse Center, the case fatality rate in the United States is 1.6%
# Reference: https://coronavirus.jhu.edu/data/mortality

# Due to the nature of the data collection, there are reasons to suggest the death rate may
# be lower (people who visit their doctor or get tested may be less likely to die since they may care about their health more,
people surveyed are likely to come from areas where healthcare is more widely available - perhaps better)
# or higher (people who visit their doctor or get tested may have a more serious case since they are more likely to be symptomatic,
# some cases may have been reported from hospitalized COVID patients which are likely more severe)
# As a result, we will perform a two-sided test

# Null Hypothesis: True death rate among surveyed U.S. COVID patients is 1.6%, the same as the reported U.S. death rate.
# Alternative Hypothesis: True death rate among surveyed U.S. COVID patients is not 1.6%.

# Perform the proportion test
prop.test(sum(data.table$Death), sum(data.table$Total), 0.016)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(data.table$Death) out of sum(data.table$Total), null probability 0.016
## X-squared = 75.877, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.016
## 95 percent confidence interval:
##  0.01145345 0.01296459
## sample estimates:
##          p
## 0.01218609
```

```
# p-value < 2.2e-16
# Since p-value is very close to 0, we are able to reject the null hypothesis.
# Therefore, we can conclude that the reported death proportion is inaccurate.
# Due to this, when imputing missing values for death rate, we should use our data's rate instead of the reported rate.
```

Second Hypothesis: Are age group, sex, race, etc. significant predictors of COVID-19 death?

```
# Pick variables can be used to build a logistic regression model
data.inc <- data.inc[ , c("case_month", "age_group", "sex", "race", "current_status", "symptom_status", "hosp_yn", "death_y
n")]

# To deal with character value to factor
data.inc$age_group <- as.factor(data.inc$age_group)
data.inc$sex <- as.factor(data.inc$sex)
data.inc$race <- as.factor(data.inc$race)
data.inc$current_status <- as.factor(data.inc$current_status)
data.inc$symptom_status <- as.factor(data.inc$symptom_status)
data.inc$hosp_yn <- as.factor(data.inc$hosp_yn)
data.inc$death_yn <- as.factor(data.inc$death_yn)

# Statistics for data.inc
str(data.inc) # the number of observations before deleting all missing values
```

```
## 'data.frame':    100000 obs. of  8 variables:
##  $ case_month    : num  6 16 13 13 13 12 12 14 9 4 ...
##  $ age_group     : Factor w/ 4 levels "0 - 17 years",..: 4 2 4 2 3 2 2 4 1 3 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 2 1 1 1 1 2 ...
##  $ race          : Factor w/ 6 levels "American Indian/Alaska Native",..: 6 6 6 6 3 6 6 NA 6 NA ...
##  $ current_status: Factor w/ 2 levels "Laboratory-confirmed case",..: 1 1 1 2 1 1 1 1 1 1 ...
##  $ symptom_status: Factor w/ 2 levels "Asymptomatic",..: NA 2 2 2 NA 2 2 2 2 2 ...
##  $ hosp_yn       : Factor w/ 2 levels "No","Yes": NA 1 1 1 NA 1 1 NA 1 1 ...
##  $ death_yn      : Factor w/ 2 levels "No","Yes": NA 1 1 1 1 1 1 NA 1 1 ...
```

```
summary(data.inc)
```

```
##    case_month             age_group         sex
##  Min.   : 1.00   0 - 17 years  :16428   Female:51172
##  1st Qu.: 9.00   18 to 49 years:49194   Male  :44121
##  Median :12.00   50 to 64 years:17985   NA's  : 4707
##  Mean   :11.75   65+ years     :14722
##  3rd Qu.:14.00   NA's          : 1671
##  Max.   :21.00
##
##                                      race
##  American Indian/Alaska Native         :  355
##  Asian                                 : 1237
##  Black                                 : 9481
##  Multiple/Other                        : 3008
##  Native Hawaiian/Other Pacific Islander:  229
##  White                                 :61147
##  NA's                                  :24543
##                    current_status       symptom_status  hosp_yn
##  Laboratory-confirmed case:86610   Asymptomatic: 5370   No  :81067
##  Probable Case            :13390   Symptomatic :84684   Yes : 6137
##                                    NA's        : 9946   NA's:12796
##
##
##
##
##  death_yn
##  No  :82590
##  Yes : 1009
##  NA's:16401
##
##
##
##
```

```
sum(is.na(data.inc))
```

```
## [1] 70064
```

```
# Delete all missing values
data.inc <- data.inc[complete.cases(data.inc), ]
str(data.inc) # the number of observations after deleting all missing values
```

```
## 'data.frame':    58796 obs. of  8 variables:
##  $ case_month    : num  16 13 13 12 12 9 6 8 10 13 ...
##  $ age_group     : Factor w/ 4 levels "0 - 17 years",..: 2 4 2 2 2 1 2 2 2 2 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 2 2 ...
##  $ race          : Factor w/ 6 levels "American Indian/Alaska Native",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ current_status: Factor w/ 2 levels "Laboratory-confirmed case",..: 1 1 2 1 1 1 1 1 1 1 ...
##  $ symptom_status: Factor w/ 2 levels "Asymptomatic",..: 2 2 2 2 2 2 2 2 1 2 ...
##  $ hosp_yn       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ death_yn      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
# Separate data set into training set and test set
set.seed(123)
split <- data.inc$death_yn %>% createDataPartition(p = 0.80, list = FALSE)
train <- data.inc[split, ]
test <- data.inc[-split, ]
str(train) # the number of observations in training set
```

```
## 'data.frame':    47038 obs. of  8 variables:
##  $ case_month    : num  13 13 12 9 6 8 10 13 13 18 ...
##  $ age_group     : Factor w/ 4 levels "0 - 17 years",..: 4 2 2 1 2 2 2 2 2 2 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 2 2 1 1 ...
##  $ race          : Factor w/ 6 levels "American Indian/Alaska Native",..: 6 6 6 6 6 6 6 6 6 4 ...
##  $ current_status: Factor w/ 2 levels "Laboratory-confirmed case",..: 1 2 1 1 1 1 1 1 1 1 ...
##  $ symptom_status: Factor w/ 2 levels "Asymptomatic",..: 2 2 2 2 2 2 1 2 2 2 ...
##  $ hosp_yn       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ death_yn      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
str(test) # the number of observations in test set
```

```
## 'data.frame':    11758 obs. of  8 variables:
##  $ case_month    : num  16 12 7 13 20 12 16 13 10 13 ...
##  $ age_group     : Factor w/ 4 levels "0 - 17 years",..: 2 2 3 2 3 2 3 2 2 2 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 2 1 2 2 2 2 1 ...
##  $ race          : Factor w/ 6 levels "American Indian/Alaska Native",..: 6 6 6 6 6 6 6 3 6 3 ...
##  $ current_status: Factor w/ 2 levels "Laboratory-confirmed case",..: 1 1 1 1 2 1 1 2 1 1 ...
##  $ symptom_status: Factor w/ 2 levels "Asymptomatic",..: 2 2 1 2 2 2 2 2 2 2 ...
##  $ hosp_yn       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ death_yn      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
# Build a logistic regression model
M.raw <- glm(death_yn ~ (.), family = binomial("logit"), data = train)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(M.raw)
```

```
##
## Call:
## glm(formula = death_yn ~ (.), family = binomial("logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1173  -0.0279  -0.0084  -0.0053   4.0447
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                 -36.54200 1564.21940  -0.023
## case_month                                   -0.12749    0.01449  -8.796
## age_group18 to 49 years                      11.06892  330.46971   0.033
## age_group50 to 64 years                      13.78106  330.46841   0.042
## age_group65+ years                           17.96019  330.46820   0.054
## sexMale                                       0.17629    0.10371   1.700
## raceAsian                                    16.15384 1528.91250   0.011
## raceBlack                                    15.82483 1528.91238   0.010
## raceMultiple/Other                           14.92332 1528.91254   0.010
## raceNative Hawaiian/Other Pacific Islander    1.31508 2491.38730   0.001
## raceWhite                                    16.33155 1528.91237   0.011
## current_statusProbable Case                   0.76273    0.16229   4.700
## symptom_statusSymptomatic                     0.10522    0.23903   0.440
## hosp_ynYes                                    2.08281    0.10950  19.021
##                                            Pr(>|z|)
## (Intercept)                                  0.9814
## case_month                                  < 2e-16 ***
## age_group18 to 49 years                      0.9733
## age_group50 to 64 years                      0.9667
## age_group65+ years                           0.9567
## sexMale                                      0.0892 .
## raceAsian                                    0.9916
## raceBlack                                    0.9917
## raceMultiple/Other                           0.9922
## raceNative Hawaiian/Other Pacific Islander   0.9996
## raceWhite                                    0.9915
## current_statusProbable Case                2.61e-06 ***
## symptom_statusSymptomatic                    0.6598
## hosp_ynYes                                  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5126.5  on 47037  degrees of freedom
## Residual deviance: 2796.7  on 47024  degrees of freedom
## AIC: 2824.7
##
## Number of Fisher Scoring iterations: 20
```

```
# According to the results, case month, current status, and hospitalization are significant predictors of death at alpha =
 0.05.
# Case month and hospitalization both have p-values < 2e-16, so very significant

# Using step wise method to find the best model
step.model <- step(M.raw, direction = "both", trace = FALSE)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(step.model)
```

```
##
## Call:
## glm(formula = death_yn ~ case_month + age_group + sex + race + 
##     current_status + hosp_yn, family = binomial("logit"), data = train)
##
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.1162  -0.0278  -0.0083  -0.0053   4.0442  
##
## Coefficients:
##                                             Estimate Std. Error z value
## (Intercept)                                 -36.4441  1564.8332  -0.023
## case_month                                   -0.1276     0.0145  -8.803
## age_group18 to 49 years                      11.0721   330.4839   0.034
## age_group50 to 64 years                      13.7838   330.4826   0.042
## age_group65+ years                           17.9595   330.4823   0.054
## sexMale                                       0.1786     0.1036   1.724
## raceAsian                                    16.1567  1529.5374   0.011
## raceBlack                                    15.8203  1529.5373   0.010
## raceMultiple/Other                           14.9227  1529.5375   0.010
## raceNative Hawaiian/Other Pacific Islander    1.3231  2491.6079   0.001
## raceWhite                                    16.3311  1529.5373   0.011
## current_statusProbable Case                   0.7614     0.1623   4.692
## hosp_ynYes                                    2.0890     0.1087  19.225
##                                            Pr(>|z|)    
## (Intercept)                                  0.9814    
## case_month                                  < 2e-16 ***
## age_group18 to 49 years                      0.9733    
## age_group50 to 64 years                      0.9667    
## age_group65+ years                           0.9567    
## sexMale                                      0.0847 .  
## raceAsian                                    0.9916    
## raceBlack                                    0.9917    
## raceMultiple/Other                           0.9922    
## raceNative Hawaiian/Other Pacific Islander   0.9996    
## raceWhite                                    0.9915    
## current_statusProbable Case                 2.7e-06 ***
## hosp_ynYes                                  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5126.5  on 47037  degrees of freedom
## Residual deviance: 2796.9  on 47025  degrees of freedom
## AIC: 2822.9
##
## Number of Fisher Scoring iterations: 20
```

```
# The model selected includes the variables case month, age group, sex, race, current status, and hospitalization.
# Same as before, the p-values for case month, current status, and hospitalization are less than 0.05
# The other variables may have been selected because it believes them to be significant, but their error is too high
# for them to be statistically significant.

# Use subset method to find the best model
M <- regsubsets(death_yn ~ ., data=train, nbest = 1 , nvmax=5, method = 'forward', intercept = TRUE)
temp <- summary(M)
Var <- colnames(model.matrix(M.raw))
M_select <- apply(temp$which, 1, function(x) paste0(Var[x], collapse='+'))
kable(data.frame(cbind(model = M_select, BIC = temp$bic)), caption='Model Summary')
```

Model Summary

model	BIC
(Intercept)+age_group65+ years	-3058.67416567003
(Intercept)+age_group65+ years+hosp_ynYes	-4794.58185024012
(Intercept)+case_month+age_group65+ years+hosp_ynYes	-4845.98243275054
(Intercept)+case_month+age_group65+ years+current_statusProbable Case+hosp_ynYes	-4853.09912349678
(Intercept)+case_month+age_group50 to 64 years+age_group65+ years+current_statusProbable Case+hosp_ynYes	-4856.55308964

```
# The models include the variables age group, case month, hospitalization, and race.
# The BIC decreases (improves) with more variables included, but the decrease from adding age_group50 to 64 years is negligible.
# Therefore, the best model seems to be the fourth, which includes case month, age group, current status, and hospitalization
# These are the same as when we ignored missing values.
```

```
# We decided to select case month, age group, current status, and hospitalization. The same as the fourth model from regsubsets.
# We did this because they were in the best models of both methods.

# Final model
M.final <- glm(death_yn ~ (case_month + age_group + current_status + hosp_yn), family = binomial("logit"), data = train)
summary(M.final)
```

```
## 
## Call:
## glm(formula = death_yn ~ (case_month + age_group + current_status + 
##     hosp_yn), family = binomial("logit"), data = train)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.0438  -0.0280  -0.0085  -0.0059   4.0987  
## 
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                -20.17149  332.76211  -0.061    0.952    
## case_month                  -0.12353    0.01434  -8.615  < 2e-16 ***
## age_group18 to 49 years     11.06510  332.76358   0.033    0.973    
## age_group50 to 64 years     13.80831  332.76229   0.041    0.967    
## age_group65+ years          18.01801  332.76208   0.054    0.957    
## current_statusProbable Case  0.75318    0.16173   4.657 3.21e-06 ***
## hosp_ynYes                   2.06587    0.10769  19.184  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 5126.5  on 47037  degrees of freedom
## Residual deviance: 2816.4  on 47031  degrees of freedom
## AIC: 2830.4
## 
## Number of Fisher Scoring iterations: 20
```

```
# Make predictions, see relationship, and classify using threshold of 0.5
prob = predict(M.final, test, type = "response")
head(subset(prob, test$death_yn == "Yes"))
```

```
##     1761415     1668182      263760     1459910     1754683     1181196 
## 0.172218301 0.025684227 0.025684227 0.172218301 0.003080174 0.066136616 
```

```
pred = ifelse(prob > 0.5, 1, 0)

# Confusion matrix
table(test$death_yn, pred)
```

```
##      pred
##           0
##   No  11645
##   Yes   113
```

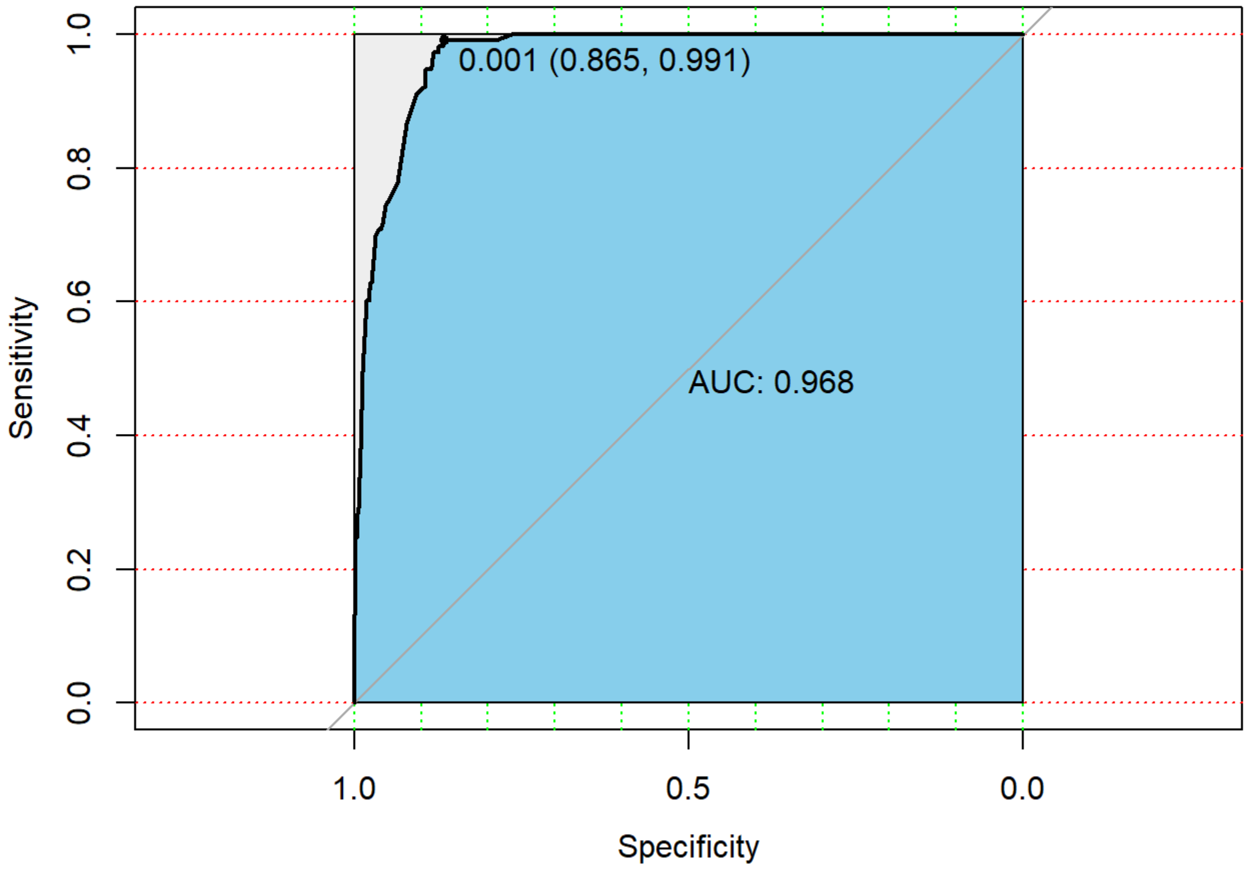
```
# We end up predicting that pretty much all of the patients will not die.
# This shows that we can't predict whether or not a patient will die.
# In addition, this result is not useful in determining patient's risk of death.

# Sensitivity vs. Specificity
modelroc <- roc(test$death_yn, prob)
```

```
## Setting levels: control = No, case = Yes
```

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

```
plot(modelroc,
     print.auc = TRUE,
     auc.polygon = TRUE,
     grid = c(0.1, 0.2),
     grid.col = c("green", "red"),
     max.auc.polygon = TRUE,
     auc.polygon.col = "skyblue",
     print.thres = TRUE)
```



```
# Found pretty good specificity and sensitivity with a threshold of 0.001.
# AUC of 0.968 shows that our model predicts death well.
# Can conclude that using this model, we should consider a patient to have a
# high (relative) risk of death if the predicted prob is > 0.001.
```

Part 2 : Generate Missing Values #Process Data Set that Assists in Generating Missing Values

```
# Data Provided by CDC
# Reference: https://covid.cdc.gov/covid-data-tracker/#demographics
case.assist <- read.csv("cases_by_age_group.csv", header = TRUE, skip = 2)
death.assist <- read.csv("deaths_by_age_group.csv", header = TRUE, skip = 2)
data.assist <- merge(case.assist, death.assist, by = "Age.Group")

data.assist[data.assist == "<0.1"] <- 0.3 / 4
data.assist[ , "Percentage.of.deaths"] <- as.numeric(data.assist[ , "Percentage.of.deaths"])

for (i in rownames(data.assist)) {
  if (data.assist[i, "Age.Group"] == "0-4 Years" |
      data.assist[i, "Age.Group"] == "5-11 Years" |
      data.assist[i, "Age.Group"] == "12-15 Years" |
      data.assist[i, "Age.Group"] == "16-17 Years" ) {
    data.assist[i, "Age.Group"] = "0 - 17 years"
  }
  else if (data.assist[i, "Age.Group"] == "18-29 Years" |
           data.assist[i, "Age.Group"] == "30-39 Years" |
           data.assist[i, "Age.Group"] == "40-49 Years") {
    data.assist[i, "Age.Group"] = "18 to 49 years"
  }
  else if (data.assist[i, "Age.Group"] == "65-74 Years" |
           data.assist[i, "Age.Group"] == "75-84 Years" |
           data.assist[i, "Age.Group"] == "85+ Years") {
    data.assist[i, "Age.Group"] = "65+ years"
  }
  else {
    data.assist[i, "Age.Group"] = "50 to 64 years"
  }
}

data.assist <- summarise(group_by(data.assist, Age.Group),
                           Percent.of.cases = sum(Percent.of.cases, na.rm = TRUE) / 100,
                           Count.of.cases = sum(Count.of.cases, na.rm = TRUE),
                           Percentage.of.deaths = sum(Percentage.of.deaths, na.rm = TRUE) / 100,
                           Count.of.deaths = sum(Count.of.deaths, na.rm = TRUE),
                           Case.fatality.rate = Count.of.deaths / Count.of.cases)
data.assist
```

	Age.Group <chr>	Percent.of.cases <dbl>	Count.of.cases <int>	Percentage.of.deaths <dbl>	Count.of.deaths <int>
1	0 - 17 years	0.152	5721775	0.003	902
2	18 to 49 years	0.529	19964162	0.060	38510
3	50 to 64 years	0.192	7260497	0.169	106856
4	65+ years	0.128	4821717	0.768	487028

4 rows | 1-6 of 7 columns

```
# According to the result, the proportion of infection for "0 - 17 years" is 15.2%, the proportion of infection for "18 to 49 years" is 52.9%, the proportion of infection for "50 to 64 years" is 19.2%, and the proportion of infection for "65+ years" is 12.8%.
infection.rate <- c(data.assist$Percent.of.cases)

# According to the result, the death rate for "0 - 17 years" is 0.016%, the death rate for "18 to 49 years" is 0.193%, the death rate for "50 to 64 years" is 1.472%, and the death rate for "65+ years" is 10.101%.
death.rate <- c(data.assist$Case.fatality.rate)
```

Copy Data

```
data.imp <- data.original
```

Generate MNAR

```
# The number of missing values in columns, "age_group" and "death_yn" before generating missing values.
length(rownames(data.imp[is.na(data.imp$age_group), ]))
```

```
## [1] 1671
```

```
length(rownames(data.imp[is.na(data.imp$death_yn), ]))
```

```
## [1] 16401
```

```r
length(rownames(data.imp[is.na(data.imp$underlying_conditions_yn), ]))
```

```
## [1] 71463
```

```
# According to previous result, the proportion of infection for "0 - 17 years" is 15.2%, the proportion of infection for "18 to 49 years" is 52.9%, the proportion of infection for "50 to 64 years" is 19.2%, and the proportion of infection for "65+ years" is 12.8%.
for (individual in rownames(data.imp[is.na(data.imp$age_group), ])) {
  data.imp[individual, "age_group"] = sample(c("0 - 17 years", "18 to 49 years", "50 to 64 years", "65+ years"), 1, prob = infection.rate)
}

# According to previous result, the death rate for "0 - 17 years" is 0.016%, the death rate for "18 to 49 years" is 0.193%, the death rate for "50 to 64 years" is 1.472%, and the death rate for "65+ years" is 10.101%.
for (individual in rownames(data.imp[is.na(data.imp$death_yn), ])) {
  if (data.imp[individual, "age_group"] == "0 - 17 years") {
    x <- runif(1)
    if (x < death.rate[1]) {
      data.imp[individual, "death_yn"] = "Yes"
    }
    else {
      data.imp[individual, "death_yn"] = "No"
    }
  }
  else if (data.imp[individual, "age_group"] == "18 to 49 years") {
    x <- runif(1)
    if (x < death.rate[2]) {
      data.imp[individual, "death_yn"] = "Yes"
    }
    else {
      data.imp[individual, "death_yn"] = "No"
    }
  }
  else if (data.imp[individual, "age_group"] == "50 to 64 years") {
    x <- runif(1)
    if (x < death.rate[3]) {
      data.imp[individual, "death_yn"] = "Yes"
    }
    else {
      data.imp[individual, "death_yn"] = "No"
    }
  }
  else {
    x <- runif(1)
    if (x < death.rate[4]) {
      data.imp[individual, "death_yn"] = "Yes"
    }
    else {
      data.imp[individual, "death_yn"] = "No"
    }
  }
}

# The number of missing values in columns, "age_group" and "death_yn" after generating missing values.
length(rownames(data.imp[is.na(data.imp$age_group), ]))
```

```
## [1] 0
```

```
length(rownames(data.imp[is.na(data.imp$death_yn), ]))
```

```
## [1] 0
```

```
length(rownames(data.imp[is.na(data.imp$underlying_conditions_yn), ]))
```

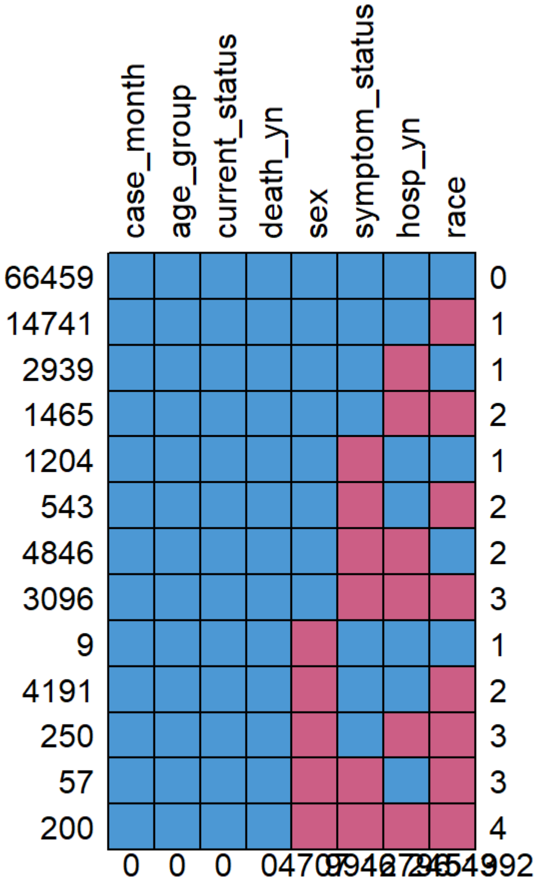
```
## [1] 71463
```

Generate MCAR

```r
# Pick variables that can be used in this project
data.imp <- data.imp[ , c("case_month", "age_group", "sex", "race", "current_status", "symptom_status", "hosp_yn", "death_yn")]

# To deal with character value to factor
data.imp$age_group <- as.factor(data.imp$age_group)
data.imp$sex <- as.factor(data.imp$sex)
data.imp$race <- as.factor(data.imp$race)
data.imp$current_status <- as.factor(data.imp$current_status)
data.imp$symptom_status <- as.factor(data.imp$symptom_status)
data.imp$hosp_yn <- as.factor(data.imp$hosp_yn)
data.imp$death_yn <- as.factor(data.imp$death_yn)

# Applying "mice"
md.pattern(data.imp, rotate.names = TRUE)
```



	case_month	age_group	current_status	death_yn	sex	symptom_status	hosp_yn	race	
66459	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	0
14741	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	1
2939	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟦	1
1465	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥	2
1204	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟦	1
543	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟥	2
4846	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟦	2
3096	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟥	3
9	🟦	🟦	🟦	🟦	🟥	🟦	🟦	🟦	1
4191	🟦	🟦	🟦	🟦	🟥	🟦	🟦	🟥	2
250	🟦	🟦	🟦	🟦	🟥	🟦	🟥	🟥	3
57	🟦	🟦	🟦	🟦	🟥	🟥	🟦	🟥	3
200	🟦	🟦	🟦	🟦	🟥	🟥	🟥	🟥	4
	0	0	0	0	4707	9946	12796	24543	51992

```
##       case_month age_group current_status death_yn  sex symptom_status hosp_yn
## 66459          1         1              1        1    1              1       1
## 14741          1         1              1        1    1              1       1
## 2939           1         1              1        1    1              1       0
## 1465           1         1              1        1    1              1       0
## 1204           1         1              1        1    1              0       1
## 543            1         1              1        1    1              0       1
## 4846           1         1              1        1    1              0       0
## 3096           1         1              1        1    1              0       0
## 9              1         1              1        1    0              1       1
## 4191           1         1              1        1    0              1       1
## 250            1         1              1        1    0              1       0
## 57             1         1              1        1    0              0       1
## 200            1         1              1        1    0              0       0
##                0         0              0        0 4707           9946   12796
##        race      
## 66459     1     0
## 14741     0     1
## 2939      1     1
## 1465      0     2
## 1204      1     1
## 543       0     2
## 4846      1     2
## 3096      0     3
## 9         1     1
## 4191      0     2
## 250       0     3
## 57        0     3
## 200       0     4
##       24543 51992
```

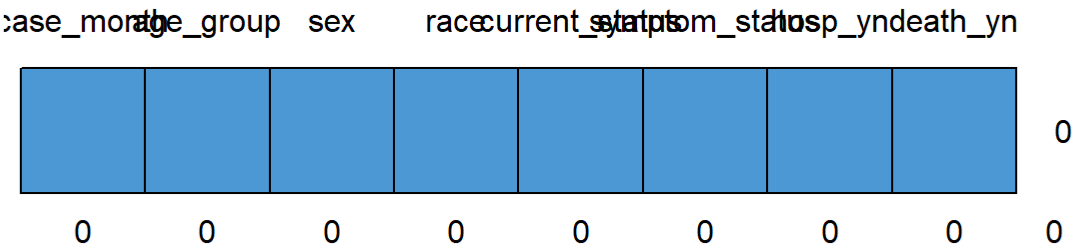
```
imp <- mice(data.imp, method = "pmm", seed = 123) # m = 5
```

```
##
##  iter imp variable
##   1   1  sex  race  symptom_status  hosp_yn
##   1   2  sex  race  symptom_status  hosp_yn
##   1   3  sex  race  symptom_status  hosp_yn
##   1   4  sex  race  symptom_status  hosp_yn
##   1   5  sex  race  symptom_status  hosp_yn
##   2   1  sex  race  symptom_status  hosp_yn
##   2   2  sex  race  symptom_status  hosp_yn
##   2   3  sex  race  symptom_status  hosp_yn
##   2   4  sex  race  symptom_status  hosp_yn
##   2   5  sex  race  symptom_status  hosp_yn
##   3   1  sex  race  symptom_status  hosp_yn
##   3   2  sex  race  symptom_status  hosp_yn
##   3   3  sex  race  symptom_status  hosp_yn
##   3   4  sex  race  symptom_status  hosp_yn
##   3   5  sex  race  symptom_status  hosp_yn
##   4   1  sex  race  symptom_status  hosp_yn
##   4   2  sex  race  symptom_status  hosp_yn
##   4   3  sex  race  symptom_status  hosp_yn
##   4   4  sex  race  symptom_status  hosp_yn
##   4   5  sex  race  symptom_status  hosp_yn
##   5   1  sex  race  symptom_status  hosp_yn
##   5   2  sex  race  symptom_status  hosp_yn
##   5   3  sex  race  symptom_status  hosp_yn
##   5   4  sex  race  symptom_status  hosp_yn
##   5   5  sex  race  symptom_status  hosp_yn
```

```
data.com <- complete(imp)

# Check
md.pattern(data.com)
```

```
##  /\     /\
## {  `---'  }
## {  O   O  }
## ==>  V <==  No need for mice. This data set is completely observed.
##  \  \|/  /
##   `-----'
```



```
##        case_month age_group sex race current_status symptom_status hosp_yn
## 100000          1         1   1    1              1              1       1
##                 0         0   0    0              0              0       0
##        death_yn
## 100000        1 0
##               0 0
```

```
sum(is.na(data.com))
```

```
## [1] 0
```

First Hypothesis: Are the reported death proportions for varying age groups the same as those in surveillance data?

```
# Group data set by "age_group" variable
data.table.com <- summarise(group_by(data.com, age_group),
                        Death = length(which(death_yn == "Yes")),
                        Survive = length(which(death_yn == "No")),
                        Total = sum(Death, Survive))
data.table.com
```

	age_group <fct>	**Death** <int>	**Survive** <int>	**Total** <int>
1	0 - 17 years	0	16664	16664
2	18 to 49 years	21	50053	50074
3	50 to 64 years	77	18237	18314
4	65+ years	1407	13541	14948

4 rows

```
# Total Deaths
sum(data.table.com$Death)
```

```
## [1] 1505
```

```
# Total # of Observations
sum(data.table.com$Total)
```

```
## [1] 100000
```

```
# Expected # of deaths by age group
sum(data.table.com$Death) * prop.reported
```

```
## [1]    1.189731   98.100703  274.839681 1130.869884
```

```
# To be consistent with the previous version of the test, we merge 0-17 and 18-49 groups
obs.com <- c(data.table.com$Death[1] + data.table.com$Death[2], data.table.com$Death[3], data.table.com$Death[4])
prop.reported.final
```

```
## [1] 0.06597371 0.18261773 0.75140856
```

```
obs.com
```

```
## [1]   21   77 1407
```

```
# Perform the chi-square test
# Null Hypothesis: the reported death proportions are the same as those in surveillance data
# Alternative Hypothesis: the reported death proportions are not the same as those in surveillance data
chisq.test(x = obs.com, p = prop.reported.final)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obs.com
## X-squared = 271.57, df = 2, p-value < 2.2e-16
```

```
# According to the result of the test, we are able to reject the null hypothesis since p-value is less than 5%.
# Hence, the reported death proportions for each age group from NCHS is significantly different from that of
# the surveillance data.
```

First Hypothesis: Is the true death rate of surveyed U.S. COVID patients equal to the reported U.S. COVID death rate?

```
# Null Hypothesis: True death rate among surveyed U.S. COVID patients is 1.6%, the same as the reported U.S. death rate.
# Alternative Hypothesis: True death rate among surveyed U.S. COVID patients is not 1.6%.

# Perform the proportion test
prop.test(sum(data.table.com$Death), sum(data.table.com$Total), 0.016)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(data.table.com$Death) out of sum(data.table.com$Total), null probability 0.016
## X-squared = 5.6722, df = 1, p-value = 0.01724
## alternative hypothesis: true p is not equal to 0.016
## 95 percent confidence interval:
##  0.01430893 0.01582858
## sample estimates:
##       p
## 0.01505
```

```
# Since p-value is less than 5%, we are able to reject the null hypothesis.
# Therefore, we can conclude that the reported death rate is significantly different from that of surveyed U.S. COVID patien
ts.
```

Second Hypothesis: Are age group, sex, race, etc. significant predictors of COVID-19 death?

```
# Statistics for data.inc
str(data.com)
```

```
## 'data.frame':    100000 obs. of  8 variables:
##  $ case_month    : num  6 16 13 13 13 12 12 14 9 4 ...
##  $ age_group     : Factor w/ 4 levels "0 - 17 years",..: 4 2 4 2 3 2 2 4 1 3 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 2 1 1 1 1 2 ...
##  $ race          : Factor w/ 6 levels "American Indian/Alaska Native",..: 6 6 6 6 3 6 6 6 6 6 ...
##  $ current_status: Factor w/ 2 levels "Laboratory-confirmed case",..: 1 1 1 2 1 1 1 1 1 1 ...
##  $ symptom_status: Factor w/ 2 levels "Asymptomatic",..: 1 2 2 2 2 2 2 2 2 2 ...
##  $ hosp_yn       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 2 1 1 ...
##  $ death_yn      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 2 1 1 ...
```

```
summary(data.com)
```

```
##    case_month             age_group         sex
##  Min.   : 1.00   0 - 17 years  :16664   Female:53577
##  1st Qu.: 9.00   18 to 49 years:50074   Male  :46423
##  Median :12.00   50 to 64 years:18314
##  Mean   :11.75   65+ years     :14948
##  3rd Qu.:14.00
##  Max.   :21.00
##                                       race
##  American Indian/Alaska Native         :  450
##  Asian                                 : 1571
##  Black                                 :12748
##  Multiple/Other                        : 3872
##  Native Hawaiian/Other Pacific Islander:  233
##  White                                 :81126
##                    current_status       symptom_status  hosp_yn     death_yn
##  Laboratory-confirmed case:86610   Asymptomatic: 5941   No :92350   No :98495
##  Probable Case            :13390   Symptomatic :94059   Yes: 7650   Yes: 1505
##
##
##
##
```

```
sum(is.na(data.com))
```

```
## [1] 0
```

```
# Separate data set into training set and test set
set.seed(123)
split.com <- data.com$death_yn %>% createDataPartition(p = 0.80, list = FALSE)
train.com <- data.com[split.com, ]
test.com <- data.com[-split.com, ]
str(train.com) # the number of observations in training set
```

```
## 'data.frame':    80000 obs. of  8 variables:
##  $ case_month    : num  6 16 13 13 13 12 12 14 9 4 ...
##  $ age_group     : Factor w/ 4 levels "0 - 17 years",..: 4 2 4 2 3 2 2 4 1 3 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 2 1 1 1 1 2 ...
##  $ race          : Factor w/ 6 levels "American Indian/Alaska Native",..: 6 6 6 6 3 6 6 6 6 6 ...
##  $ current_status: Factor w/ 2 levels "Laboratory-confirmed case",..: 1 1 1 2 1 1 1 1 1 1 ...
##  $ symptom_status: Factor w/ 2 levels "Asymptomatic",..: 1 2 2 2 2 2 2 2 2 2 ...
##  $ hosp_yn       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 2 1 1 ...
##  $ death_yn      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 2 1 1 ...
```

```
str(test.com) # the number of observations in test set
```

```
## 'data.frame':    20000 obs. of  8 variables:
##  $ case_month    : num  16 13 3 13 4 12 20 12 6 17 ...
##  $ age_group     : Factor w/ 4 levels "0 - 17 years",..: 2 2 4 2 4 3 2 4 2 1 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 2 1 1 2 2 1 2 2 2 1 ...
##  $ race          : Factor w/ 6 levels "American Indian/Alaska Native",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ current_status: Factor w/ 2 levels "Laboratory-confirmed case",..: 2 1 2 2 1 1 1 1 1 2 ...
##  $ symptom_status: Factor w/ 2 levels "Asymptomatic",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ hosp_yn       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ death_yn      : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 1 1 1 ...
```

```
# Build a logistic regression model
M.raw.com <- glm(death_yn ~ (.), family = binomial("logit"), data = train.com)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(M.raw.com)
```

```
##
## Call:
## glm(formula = death_yn ~ (.), family = binomial("logit"), data = train.com)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9137  -0.0773  -0.0319  -0.0232   4.1947
##
## Coefficients:
##                                             Estimate Std. Error z value
## (Intercept)                               -2.085e+01  1.521e+02  -0.137
## case_month                                -8.019e-02  8.175e-03  -9.809
## age_group18 to 49 years                    1.278e+01  1.521e+02   0.084
## age_group50 to 64 years                    1.478e+01  1.521e+02   0.097
## age_group65+ years                         1.756e+01  1.521e+02   0.116
## sexMale                                    7.269e-02  6.248e-02   1.163
## raceAsian                                  3.065e-01  1.125e+00   0.272
## raceBlack                                  8.020e-01  1.031e+00   0.778
## raceMultiple/Other                         8.224e-01  1.046e+00   0.786
## raceNative Hawaiian/Other Pacific Islander -1.339e+01  1.131e+03  -0.012
## raceWhite                                  1.133e+00  1.025e+00   1.105
## current_statusProbable Case                5.274e-01  9.810e-02   5.377
## symptom_statusSymptomatic                  7.901e-02  1.429e-01   0.553
## hosp_ynYes                                 1.453e+00  6.515e-02  22.307
##                                           Pr(>|z|)
## (Intercept)                                  0.891
## case_month                                 < 2e-16 ***
## age_group18 to 49 years                      0.933
## age_group50 to 64 years                      0.923
## age_group65+ years                           0.908
## sexMale                                      0.245
## raceAsian                                    0.785
## raceBlack                                    0.436
## raceMultiple/Other                           0.432
## raceNative Hawaiian/Other Pacific Islander    0.991
## raceWhite                                    0.269
## current_statusProbable Case               7.59e-08 ***
## symptom_statusSymptomatic                    0.580
## hosp_ynYes                                 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12494.7  on 79999  degrees of freedom
## Residual deviance:  7953.7  on 79986  degrees of freedom
## AIC: 7981.7
##
## Number of Fisher Scoring iterations: 19
```

```
# According to the results, case month, current status, and hospitalization
# are significant predictors of death at alpha = 0.05. Same as when we ignored missing values.
# Case month and hospitalization both have p-values < 2e-16, so very significant

# Using step wise method to find the best model
step.model.com <- step(M.raw.com, direction = "both", trace = FALSE)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(step.model.com)
```

```
##
## Call:
## glm(formula = death_yn ~ case_month + age_group + race + current_status +
##     hosp_yn, family = binomial("logit"), data = train.com)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8978  -0.0785  -0.0312  -0.0234   4.2058
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                -2.074e+01  1.521e+02  -0.136
## case_month                                 -8.031e-02  8.176e-03  -9.823
## age_group18 to 49 years                     1.278e+01  1.521e+02   0.084
## age_group50 to 64 years                     1.478e+01  1.521e+02   0.097
## age_group65+ years                          1.756e+01  1.521e+02   0.115
## raceAsian                                   3.016e-01  1.125e+00   0.268
## raceBlack                                   7.991e-01  1.030e+00   0.776
## raceMultiple/Other                          8.252e-01  1.045e+00   0.789
## raceNative Hawaiian/Other Pacific Islander -1.340e+01  1.131e+03  -0.012
## raceWhite                                   1.134e+00  1.025e+00   1.106
## current_statusProbable Case                 5.260e-01  9.808e-02   5.363
## hosp_ynYes                                  1.460e+00  6.461e-02  22.593
##                                            Pr(>|z|)
## (Intercept)                                   0.892
## case_month                                  < 2e-16 ***
## age_group18 to 49 years                       0.933
## age_group50 to 64 years                       0.923
## age_group65+ years                            0.908
## raceAsian                                     0.789
## raceBlack                                     0.438
## raceMultiple/Other                            0.430
## raceNative Hawaiian/Other Pacific Islander    0.991
## raceWhite                                     0.269
## current_statusProbable Case                 8.2e-08 ***
## hosp_ynYes                                  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12494.7  on 79999  degrees of freedom
## Residual deviance:  7955.4  on 79988  degrees of freedom
## AIC: 7979.4
##
## Number of Fisher Scoring iterations: 19
```

```
# The model selected includes the variables case month, age group, race, current status,  and hospitalization.
# Same as before, the p-values for case month, current status, and hospitalization are less than 0.05
# Sex was not included, unlike that of when we ignored missing values.

# Use subset method to find the best model
M.com <- regsubsets(death_yn ~ ., data=train.com, nbest = 1 , nvmax=5, method = 'forward', intercept = TRUE)
temp.com <- summary(M.com)
Var.com <- colnames(model.matrix(M.raw.com))
M_select.com <- apply(temp.com$which, 1, function(x) paste0(Var.com[x], collapse='+'))
kable(data.frame(cbind(model = M_select.com, BIC = temp.com$bic)), caption='Model Summary')
```

Model Summary

model	BIC
(Intercept)+age_group65+ years	-6056.71695541999
(Intercept)+age_group65+ years+hosp_ynYes	-7915.17733179056
(Intercept)+case_month+age_group65+ years+hosp_ynYes	-7976.92660037921
(Intercept)+case_month+age_group65+ years+current_statusProbable Case+hosp_ynYes	-7988.77782491967
(Intercept)+case_month+age_group65+ years+raceWhite+current_statusProbable Case+hosp_ynYes	-7990.18757522276

```
# The models include the variables age group, hospitalization, case month, race, and current status.
# The BIC decreases (improves) with more variables included, but the decrease from adding raceWhite is negligible.
# Therefore, the best model seems to be the fourth, which includes age group, hospitalization, case month, and current statu
s.
```

```
# We decided to select case month, age group, current status, and hospitalization. The same as the fourth model from regsubs
ets.
# We did this because they were in the best models of both methods.

# Final model
M.final.com <- glm(death_yn ~ (age_group + case_month + hosp_yn + current_status), family = binomial("logit"), data = train.
com)
summary(M.final.com)
```

```
##
## Call:
## glm(formula = death_yn ~ (age_group + case_month + hosp_yn +
##     current_status), family = binomial("logit"), data = train.com)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8805  -0.0782  -0.0313  -0.0246   4.1422
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -19.673901 152.354965  -0.129    0.897
## age_group18 to 49 years      12.770400 152.355099   0.084    0.933
## age_group50 to 64 years      14.789214 152.354986   0.097    0.923
## age_group65+ years           17.593937 152.354941   0.115    0.908
## case_month                   -0.079771   0.008166  -9.769  < 2e-16 ***
## hosp_ynYes                    1.435174   0.064189  22.358  < 2e-16 ***
## current_statusProbable Case   0.535466   0.098010   5.463 4.67e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12494.7  on 79999  degrees of freedom
## Residual deviance:  7972.6  on 79993  degrees of freedom
## AIC: 7986.6
##
## Number of Fisher Scoring iterations: 19
```

```
# Make predictions, see relationship, and classify using threshold of 0.5
prob.com = predict(M.final.com, test.com, type = "response")
head(subset(prob.com, test.com$death_yn == "Yes"))
```

```
##         39        875       1160       1272       1325       2658
## 0.08324546 0.17912805 0.26044773 0.27610359 0.26044773 0.16769713
```

```
pred.com = ifelse(prob.com > 0.5, 1, 0)

# Confusion matrix
table(test.com$death_yn, pred.com)
```

```
##      pred.com
##           0
##   No  19699
##   Yes   301
```

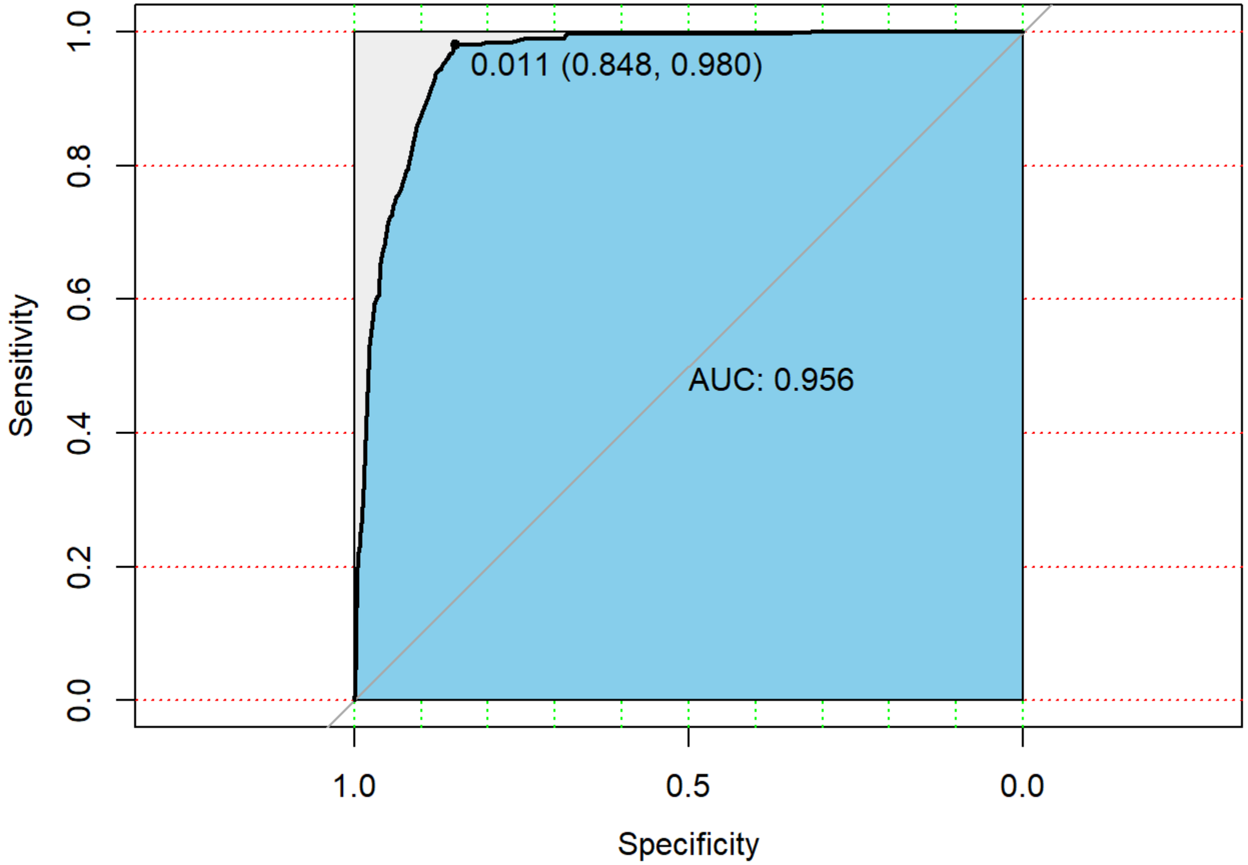
```
# We end up predicting that pretty much all of the patients will not die.
# This shows that we can't predict whether or not a patient will die.
# In addition, this result is not useful in determining patient's risk of death.

# Sensitivity vs. Specificity
modelroc.com <- roc(test.com$death_yn, prob.com)
```

```
## Setting levels: control = No, case = Yes
```

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

```
plot(modelroc.com,
     print.auc = TRUE,
     auc.polygon = TRUE,
     grid = c(0.1, 0.2),
     grid.col = c("green", "red"),
     max.auc.polygon = TRUE,
     auc.polygon.col = "skyblue",
     print.thres = TRUE)
```



```
# Found pretty good specificity and sensitivity with a threshold of 0.011.
# AUC of 0.956 shows that our model predicts death well.
# Can conclude that using this model, we should consider a patient to have a
# high (relative) risk of death if the predicted prob is > 0.011.
# Our threshold is higher than the model we got when we ignored missing values,
# and our accuracy metrics decreased by a bit.
```