Survival Analysis:

Recovery Times for COVID-19 Patients

Erick Castillo

A project presented for STAT 560

Department of Mathematics and Statistics
California State University, Long Beach
April 27, 2022

Contents

1 Introduction

At the time of the creation of this report, it has been two years since the COVID-19 virus swept the world. This topic of research holds special meaning to me, as my family and I have gotten sick from the pandemic a total of three times. Thanks to our age and our health, we were able to survive the disease. I'm certain it is a universal experience across all people who have been sick to yearn for the days when they weren't going through their ailment. This research focuses on the recovery times for hospital patients sick with the first variant of the coronavirus pandemic.

2 Background

According to the CDC, males are more likely to die due to complications from COVID-19 [1]. The same paper claims age is a significant predictor for whether the disease will be fatal towards a patient. Exposure to this paper primed me on what to look out for when conducting my analysis.

3 About the Data

The data was obtained from Kaggle, a website which hosts a community of data scientists that publicly share data. The raw data contains 1085 observations and 26 columns. Python and its data manipulation library, pandas, were used to clean the dataset, reducing its size to 140 usable observations, and 8 columns. The final 8 columns were:

- Patient ID: categorical variable.
- Age: continuous variable of age.
- Symptoms: description of patient symptoms.
- Days sick: discrete variable.
- Status: binary variable indicating patient survival.
- Age group: categorical variable indicating age group. (adolescent, young adult, middle aged, senior)

It is worth noting that the days to recovery vary widely, with a few observations extending past 30 days. There are two potential explanations for this: the first being that the data was collected early in the spread of the pandemic. At this point, doctors were still trying to understand how to qualify a recovery. Second, the data originates from hospitals, meaning that patients whose recovery times were recorded may have been so ill from the disease that they required external care. Their recovery times, in turn, would be much longer than the average case.

4 Results

Statistical softwares, SAS and R, were used for this analysis. All their output agreed for the following results.

A Kaplan-Meier (KM) Estimate on the time to recover for all members of the dataset was initially programmed into SAS. The output indicates that by day 18, 50% of patients recovered. SAS's output agrees with the output produced by R.

When stratified by age group, it is interesting to see that adolescents tend to recover much faster when compared to seniors. There is much overlap with the KM estimates of older age groups, meaning there is no statistical significance in the time to recovery for patients in the "young adult", "middle aged", and "senior" categories. A log-rank test on all groups yields a p-value of 10%. This implies there is weak evidence to suggest that the KM curves of the groups are different from one another. The R output agreed on both these points as well.

When stratifying by sex, the female KM curve consistently remained underneath and close to the male KM curve. This implies females recover slightly faster from the disease

in relation to men. A log-rank test on sex yields a p-value of 20% indicating there is little evidence to suggest that the curves for each sex are significantly different from one another.

Finally, a Cox-Proportional Hazard Model (CPHM) was fit to the data, such that recovery rates were regressed on age and gender. The output indicates that age is significant at the 0.01% level, and that gender is significant at the 10% level. The output indicated the following:

- Males have a hazard function that is 432.23% of that for females.
- A one-year increase in age increases the hazard of recovering by 7.519%, given the sex variable is held constant.

The above male percentage was calculated as follows: $100\% \times e^{1.4638}$. The percent increase in age was calculated as follows: $100\% \times (e^{0.0725} - 1)$

Predictions on the probability of an 80 year old patient going 20 days without recovering were made:

- A male patient had a 41.5% chance of getting to this point.
- A female patient had a 81.4% chance of getting to this point.

5 Conclusion

This research indicates age as a suitable predictor in determining how quickly patients recover from a COVID-19 infection.

Furthermore, it is interesting to note the discrepancy that exists for the sex variable. When analyzed from the KM perspective, females were noted to have a slightly faster recovery time; however, from the CPHM perspective, it is more likely for females to take longer to recover. This disparity exists because the CPH model simultaneously accounts for the age variable, changing how the variable is interpreted.

If time were not a constraint, it would be interesting to study the recovery times of patients in relation to their symptoms. This was a column, which to my regret, was not used or explored for this study.

6 Appendix

6.1 Python Code

```
[ ]  import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns
     import zipfile
```

```
[ ]  # importing data, extracting specific csv from zip
     zip = zipfile.ZipFile('/content/covid19-csea.zip')
     df = pd.read_csv(zip.open('COVID19_line_list_data.csv'))
     print(df.shape)
     df[:2]
```

```
[ ]  # counting missing values
     newdf.isna().sum()

     # i noticed death and recovered columns have dates

     #newdf.recovered.value_counts()
     #newdf.death.value_counts()
```

```
id                   0
reporting date       1
country              0
gender             183
age                242
symptom_onset      522
hosp_visit_date    578
exposure_start     957
exposure_end       744
death                0
recovered            0
symptom            815
dtype: int64
```

```
[ ]  death = newdf[(newdf.death != '0') & (newdf.death != '1')]
     recovered = newdf[(newdf.recovered != '0') & (newdf.recovered != '1')]
```

```
[ ]  # cleaning death dataframe

     death = death.drop(columns = ['exposure_start', 'exposure_end', 'recovered',
                                   'hosp_visit_date']) # unnecessary cols
     death = death[death.country != 'Iran'] # rows with too many missing vals
     death = death.drop(labels = 389, axis = 0)
     death = death.reset_index()

     # changing to datetime
     death['reporting date'] = death['reporting date'].apply(pd.to_datetime)
     death['symptom_onset'] = death['symptom_onset'].apply(pd.to_datetime)
     death['death'] = death['death'].apply(pd.to_datetime)

     # days until death
     start = []
     date1 = death['death'] - death['symptom_onset']
     date2 = death['death'] - death['reporting date']
     for i in range(len(date1)):
       if pd.isna(date1[i]):
         start.append(date2[i])
       else:
         start.append(date1[i])

     death['days'] = start
     death.days = (death.days / np.timedelta64(1, 'D')).astype(int)

     # dropping odd values
     death = death[death.days > 0]
```

```
[ ]  # cleaning recovered dataframe

     recovered = recovered.drop(columns = ['death', 'exposure_start',
                                           'exposure_end', 'hosp_visit_date'])
     recovered = recovered.dropna(subset = ['gender', 'age'])

     # changing to datetime
     recovered['reporting date'] = recovered['reporting date'].apply(pd.to_datetime)
     recovered['symptom_onset'] = recovered['symptom_onset'].apply(pd.to_datetime)
     recovered['recovered'] = recovered['recovered'].apply(pd.to_datetime)
     recovered = recovered.reset_index()
     # days until recovery
     start = []
     date1 = recovered['recovered'] - recovered['symptom_onset']
     date2 = recovered['recovered'] - recovered['reporting date']

     for i in range(len(date1)):
       if pd.isna(date1[i]):
         start.append(date2[i])
       else:
         start.append(date1[i])

     recovered['days'] = start
     recovered.days = (recovered.days / np.timedelta64(1, 'D')).astype(int)

     # dropping odd values
     recovered = recovered[recovered.days > 0]
```

```
[ ]  recovered = recovered.drop(columns = ['index', 'reporting date',
                                           'symptom_onset', 'recovered'])
     recovered['status'] = 1
```

```
[ ]  death = death.drop(columns = ['index', 'reporting date',
                                   'symptom_onset', 'death'])
     death['status'] = 0
```

```
[ ]  alldata = [death, recovered]
     data = pd.concat(alldata, ignore_index = True)
```

```
[ ]  age_group = []
     for i in data.age:
       if i <= 20:
         age_group.append('adolescent')
       elif i > 20 and i <= 35:
         age_group.append('young adult')
       elif i > 35 and i <= 60:
         age_group.append('middle aged')
       else:
         age_group.append('senior')
     age_group

     data['age_group'] = age_group
     data.age_group.value_counts()
```

```
middle aged    71
young adult    35
senior         25
adolescent      9
Name: age_group, dtype: int64
```

6.2 SAS Code

```
* project code;

proc import datafile = "C:\Users\casti\OneDrive\Desktop\Sch
out = work.covid
dbms = CSV; run;

/* days it takes to recover -- all data */
proc lifetest data = covid plots = (survival);
    time days*status(0);
run;

/* days it takes to die -- all data */
proc lifetest data = covid plots = survival(nocensor test);
    time days*status(1);
run;

/* recovery based on age group */
proc lifetest data = covid plots = (survival);
    time days*status(0);
        strata age_group;
run;

/* recovery based on gender */
proc lifetest data = covid plots = (survival);
    time days*status(0);
        strata gender;
run;

/* cox-proportional hazards model */
data covid1; set covid;
    M = (gender = 'male');
drop gender VAR1 id symptom country age_group;
run;


    proc phreg data = covid1 outest = betas;
        model days*status(1) = age M;
            baseline out = outdata survival = sbar;
     run;

    proc print data = outdata;
     run;
```

6.3 SAS Output

Table 1: Recovery time survival curves for all patients.

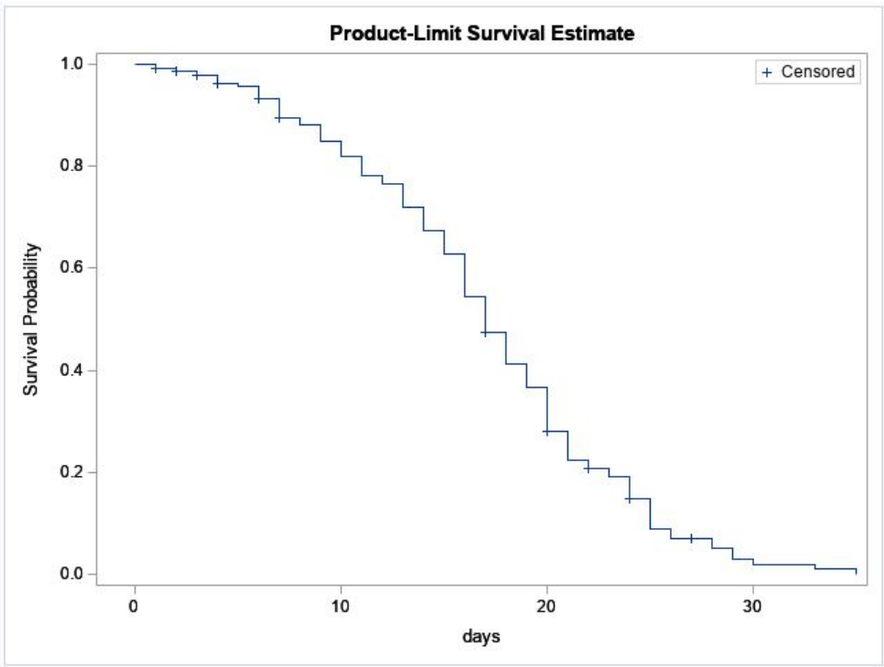

Table 2: Recovery time survival curves by age

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	6.2164	3	0.1015
Wilcoxon	5.8391	3	0.1197
-2Log(LR)	2.5165	3	0.4723

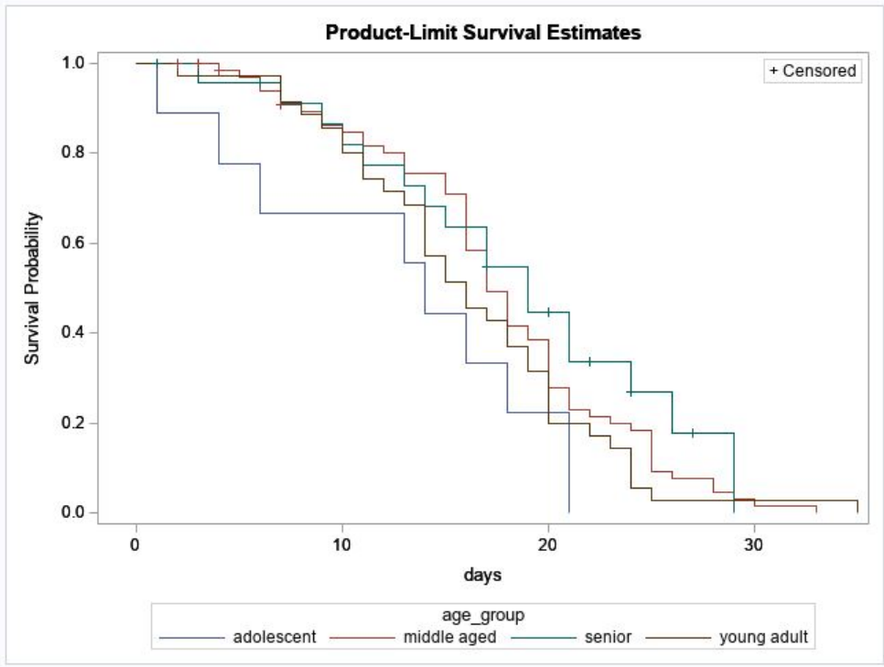

Table 3: Recovery time survival curves by Sex

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	1.9156	1	0.1663
Wilcoxon	2.3329	1	0.1267
-2Log(LR)	0.7163	1	0.3974

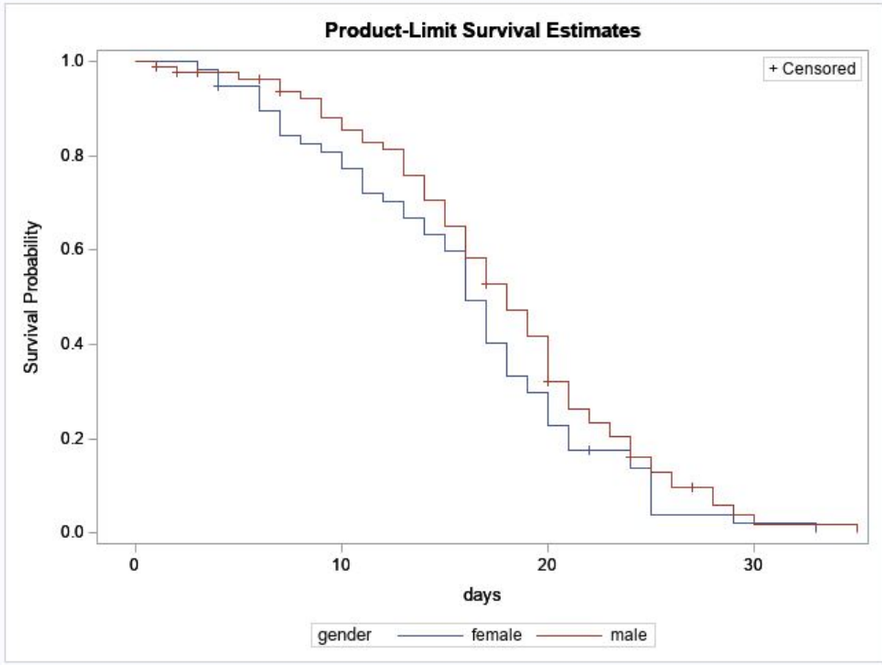

Table 4: Cox Proportional Hazard Model

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
age	1	0.07252	0.01685	18.5301	<.0001	1.075
M	1	1.46380	0.76823	3.6307	0.0567	4.322

The SAS System

Obs	age	M	days	sbar
1	44.432142857	0.58571	0	1.00000
2	44.432142857	0.58571	1	0.99469
3	44.432142857	0.58571	2	0.98906
4	44.432142857	0.58571	3	0.98335
5	44.432142857	0.58571	4	0.98041
6	44.432142857	0.58571	6	0.97743
7	44.432142857	0.58571	7	0.97430
8	44.432142857	0.58571	17	0.97016
9	44.432142857	0.58571	20	0.96414
10	44.432142857	0.58571	22	0.95568
11	44.432142857	0.58571	24	0.94622
12	44.432142857	0.58571	27	0.91907

6.4 R Code

```r
covid <- read.csv("C:/Users/casti/OneDrive/Desktop/School_Stuff/Masters_Stuff/STAT !
attach(covid)
library(survival)

# KM Survival Curves by Age Group
age_group <- factor(age_group,
                    levels = c("adolescent", "young adult",
                               "middle aged", "senior"))

days.surv <- survfit(Surv(days, status==1) ~ age_group,
                     conf.type="none", se.fit=FALSE)
summary(days.surv)

plot(days.surv, mark.time = TRUE,
     pch = 1, col = c('darkorange', 'black', 'darkgreen', 'red'),
     main = 'Kaplan-Meier Survival Curve',
     xlab = 'Days',
     ylab = 'Survival Distribution Function')


legend('topright', legend = c('Adolescent', 'Young Adult', 'Middle Age', 'Senior'),
       text.col = c('darkorange', 'black', 'darkgreen', 'red'))

survdiff(Surv(days, status==1)~ age_group)

# KM Survival Curves by Sex
gender <- factor(gender, levels = c('female', 'male'))

days.surv <- survfit(Surv(days, status==1) ~ gender,
                     conf.type="none", se.fit=FALSE)
summary(days.surv)

plot(days.surv, mark.time = TRUE,
     pch = 1, col = c('darkorchid3', 'deepskyblue4'),
     main = 'Kaplan-Meier Survival Curve',
     xlab = 'Days',
     ylab = 'Survival Distribution Function')


legend('topright', legend = c('Female', 'Male'),
       text.col =c('darkorchid3', 'deepskyblue4'))

survdiff(Surv(days, status==1) ~ gender)

# cox-ph model
covid.ph <- coxph(Surv(days, status == 1) ~ age + gender)
summary(covid.ph)
```

6.5 R Output

Table 5: Recovery time survival curves for all patients

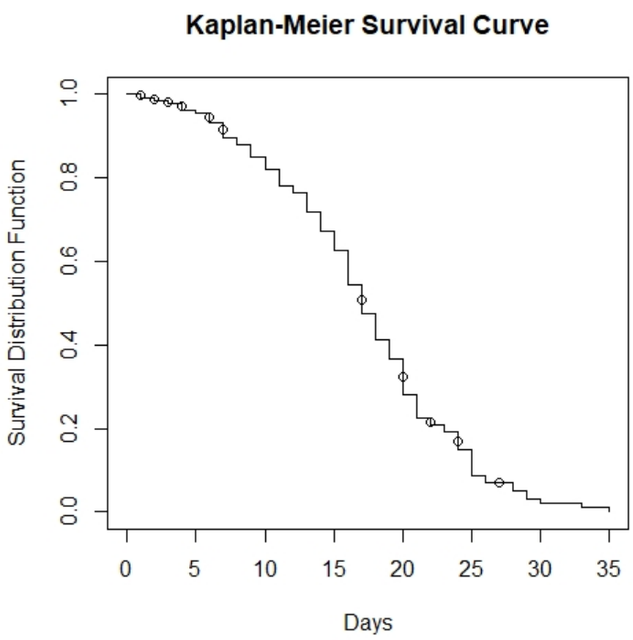

Table 6: Recovery time survival curves by age

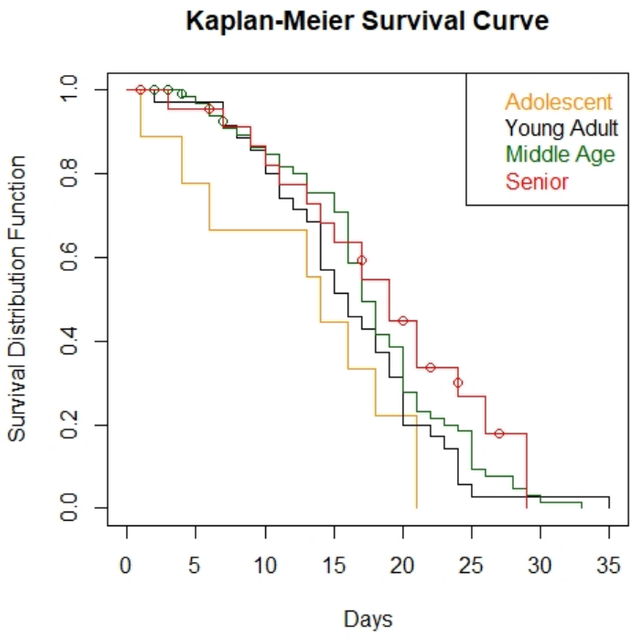

Table 7: Log-Rank Test for Age Groups

```
> survdiff(Surv(days, status==1)~ age_group)
Call:
survdiff(formula = Surv(days, status == 1) ~ age_group)

                        N Observed Expected (O-E)^2/E (O-E)^2/V
age_group=adolescent    9        9     5.05     3.088     3.562
age_group=young adult  35       35    30.12     0.792     1.204
age_group=middle aged  71       65    67.99     0.132     0.327
age_group=senior       25       17    22.84     1.495     2.083

 Chisq= 6.2  on 3 degrees of freedom, p= 0.1
```

Table 8: Recovery time survival curves by sex

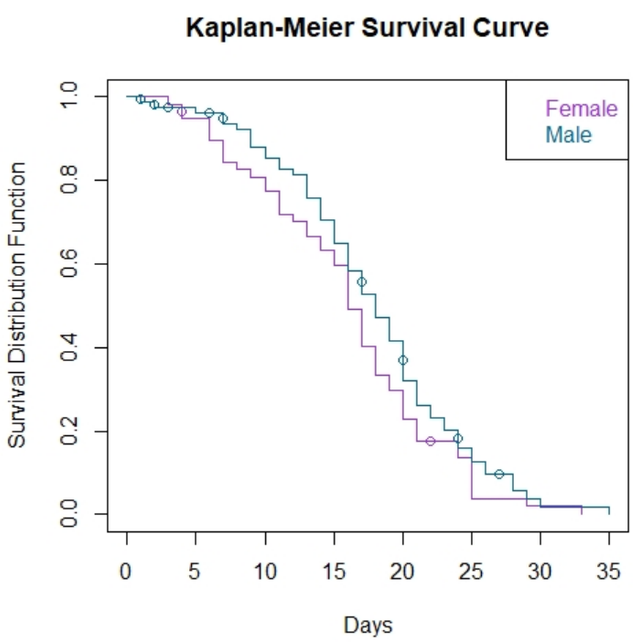

Table 9: Log-Rank test for Sex

```
> survdiff(Surv(days, status==1) ~ gender)
Call:
survdiff(formula = Surv(days, status == 1) ~ gender)

                N Observed Expected (O-E)^2/E (O-E)^2/V
gender=female 58       56     48.9     1.031      1.92
gender=male   82       70     77.1     0.654      1.92

 Chisq= 1.9  on 1 degrees of freedom, p= 0.2
```

Table 10: Cox-Proportional Hazard Model output

```
Call:
coxph(formula = Surv(days, status == 1) ~ age + gender)

  n= 140, number of events= 126

                coef exp(coef)  se(coef)      z Pr(>|z|)
age        -0.017065  0.983080  0.005511 -3.096  0.00196 **
gendermale -0.319375  0.726603  0.181921 -1.756  0.07916 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
age           0.9831      1.017    0.9725    0.9938
gendermale    0.7266      1.376    0.5087    1.0379

Concordance= 0.579  (se = 0.033 )
Likelihood ratio test= 11.81  on 2 df,   p=0.003
Wald test            = 11.46  on 2 df,   p=0.003
Score (logrank) test = 11.53  on 2 df,   p=0.003
```

References

[1] "Men and COVID-19: A Biophysical Approach" Griffith, Derek. Sharma, Garima. Centers for Disease Control and Prevention, June 16, 2020. https://www.cdc.gov/pcd/issues/2020/20_0247.htm#:~:text=In%20the%20United%20States%2C%20as,in%20reporting%20sex%2Ddisaggregated%20data.

[2] "Ending Isolation and Precautions for People with COVID-19" Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, January 12, 2022. https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html