Data Analysis of COVID-19 Outbreak in South Korea

1. Background

The COVID-19 has been spreading all over the world with more than 980,000 confirmed cases, claiming the lives of more than 50,000 by April 2, 2020. This project aims to use survival analysis methodology to assess the age different in the case fatality rate by analyzing the public dataset pulled from Korea Centers for Disease Control & Prevention. From the raw dataset, we have 2243 confirmed COVID-19 cases in total. Several key variables of interest include the patient ID, gender, age, confirmed date, released date, deceased date, and the state.

To systematically analyze the data, we first describe how we clean the data in Section 2, second conduct the descriptive analysis in Section 3, and conduct the survival analysis in Section 4. Finally, a short summary of key findings intended for the mass media will be presented in Section 5. The R code will also be attached for the reproducibility at the end of this report.

2. Data Cleaning

As indicated by the client, the outcome of interest is the time from being confirmed to death for survival analysis. In addition, considering the censoring date, if a patient is released from the isolation, we set the released date as censoring date. If a patient is still in isolation, we set the censoring date as March 21, 2020, since this is the time point that the dataset was last updated. Next, we clean the data by dealing with the missing observations for each variable.

a. Age

Since our goal is to study the age difference in the fatality rate, the first step is to look at the age variable. After converting the birth year to age, we have 148 confirmed cases that do not have information about either birth year or age. As a result, we have to delete those cases and the number of confirmed cases is 2095.

b. Confirmed date

First, we observe that there are 10 cases without confirmed date. Among these 10 cases, there are only 3 cases that have information about whom they are infected by. Since the first two cases do not have state information either, we have to delete these since it is not accurate or workable to impute both confirmed date and state simultaneously. For the third case, he was recorded as the isolation. The patient who spread the virus to the third case was confirmed in 02/18/2020. Thus, we decide to impute the third case's confirmed with 02/18/2020. For the remaining 7 cases, since we don't have any other information to impute the confirmed date, we decide to delete those and the number of cases remains 2086.

c. State

After finishing the previous step on confirmed date, we observe that there are 89 confirmed cases without the information about the state. To impute the state for these cases, we first calculate the probability of death for each age and gender based on the data with state information, as shown in Table 1.

Table 1 Probability of death for each age and gender

	0s	10s	20s	30s	40s	50s	60s	70s	80s	90s	100s
Female	0	0	0	0	0	0.012	0.008	0.028	0.054	0	0
Male	0	0	0	0.008	0.009	0.020	0.055	0.130	0.138	0.25	N/A

Second, we choose the threshold 0.05 for the imputation. Specifically, among 89 cases, for female, if patients' age is less than 80, we think they have a low probability to die and thus we consider the state as isolated conservatively. If patients are 80 years old or older, since we don't have any information to infer the deceased date, we have to delete those cases. Similarly, for male, if patients' age is less than 60, we think they have a low probability to die and thus we consider the state as isolated. Otherwise, we delete them. As a result, among 89 cases, we impute 82 cases and delete 7 cases and the number of cases is 2079.

d. Deceased date

First, we observe there are two cases that the deceased date is before the confirmed date, which is not possible in practice. Thus, we have to delete these two cases. After the deletion, we have 29 deceased cases. Among them, there is only one case which does not have deceased date. To impute it, we first calculate the mean time from the confirmed date to deceased date, which is 3.57 days. Thus, we impute the missing deceased date as 02/26/2020, which is 4 days after his confirmed date. The remaining number of confirmed cases is 2077.

e. Released date

To facilitate calculating the censoring time, for the missing data on released date, if the state is isolated, we will impute with 03/21/2020 as we have mentioned at the beginning of section 2. If the state is released and the released date is missing (82 cases), we impute it as follows. First, we calculate the mean time from confirmed date to released date for each age and gender based on the data with released information, as shown in Table 2.

Table 2 Mean time (days) from confirmed date to released date for each age and gender

	0s	10s	20s	30s	40s	50s	60s	70s	80s	90s	100s
Female	14	14	13	12	15	13	15	17	N/A	N/A	N/A
Male	17	14	13	15	14	12	15	14	17	N/A	N/A

Second, we may use mean time to impute those case without released date based on the age and gender. However, since we need to calculate the time from confirmed date to released date later for survival analysis, we don’t have to impute the released date here and can directly adopt these mean times as the time to censoring. This step is easy and convenient from the coding perspective.

f. Gender

We have only 2 cases that do not have gender information. Since we don't have any other information to infer, we decide to delete these two cases and the remaining number of cases is 2075.

3. Descriptive Analysis

a. Age distribution by gender and overall

We first tabulate the age distribution by gender and overall in Table 3. Based on Table 3, we can see that for each age group, the percentage of female cases is similar to that of male cases. Overall, the age group with ≤40s accounts for more than 50% of the total confirmed cases and as the age goes greater, the percentage of confirmed cases tends to decrease.

Table 3 Age distribution by gender and overall

Age group	Female	Male	Overall
≤ 40s	636 (54%)	556 (61.9%)	1192 (57.4%)
50s	259 (22%)	153 (17.1%)	412 (19.9%)
60s	137 (11.6%)	109 (12.2%)	246 (11.9%)
70s	76 (6.5%)	46 (5.1%)	122 (5.9%)
≥ 80s	70 (5.9%)	33 (3.7%)	103 (4.9%)

b. CFR by gender and age

We also calculate the CFR (proportion of deaths among the confirmed cases) by age and gender, which is shown in Table 4. From Table 4, we can see that for each age group, the CFR by male is always higher than that by female, which indicates that the male has a higher fatality rate than female. In addition, for each gender, the older the people, the higher the CFR. This shows that the old people are more likely to die than the young people. We also plot the age-specific CFR in Figure 1, from which we can see the similar results that the old people have a higher CFR than the young people. Compared with the report by the China CDC, the results here are comparable since they have a similar trend that the fatality rate increases as people ages.

Table 4 CFR by gender and age

	≤40s	50s	60s	70s	≥80s
Female	0.000	0.008	0.007	0.026	0.043
Male	0.004	0.020	0.046	0.130	0.152

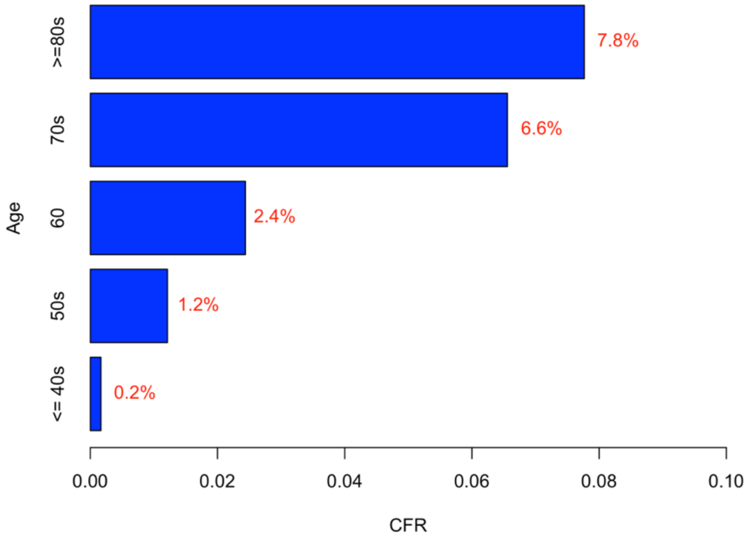


Figure 1 Age-specific distribution of CFR

4. Survival Analysis

a. Kaplan-Meier curves

The gender-specific and age-specific Kaplan-Meier curves for the case survival probabilities are plotted in Figure 2, respectively. By comparison, we can see that the female appears to enjoy higher survival probabilities than the male. Meanwhile, similar to previous observations, the young people is more likely to survive than the old people.

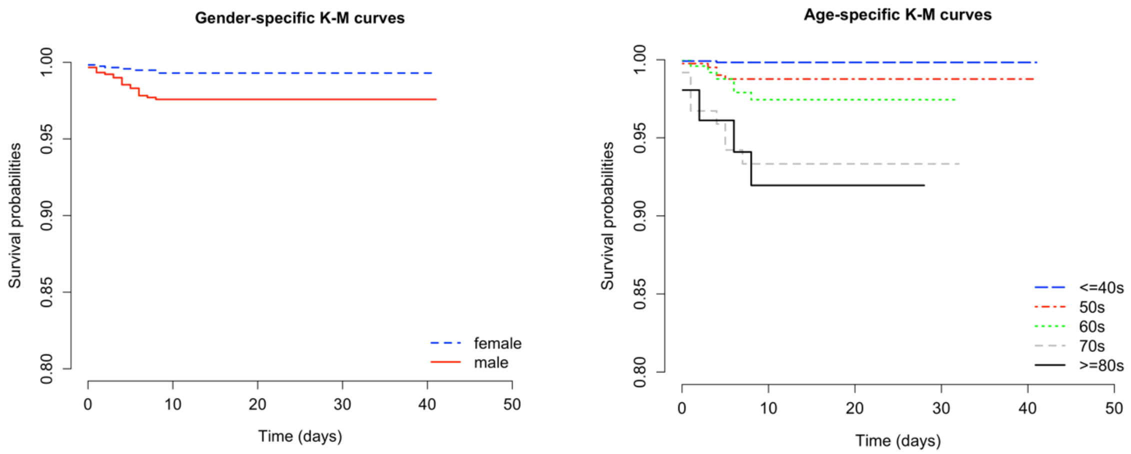
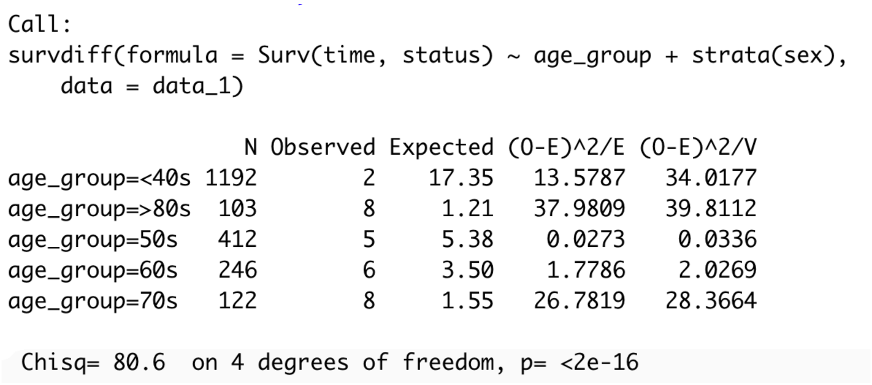


Figure 2: Gender-specific (left) and age-specific (right) Kaplan-Meier curves

To conduct a formal test on the effect of the age difference, we use the log-rank test stratified by the gender. The output of the log-rank test is shown in Figure 3. The resulting χ_4^2 test statistic is 80.6 with a p-value less than 2×10^{-16}, which suggest a statistically significant difference among age groups by controlling for the gender. In addition, according to the observed and expected number of events by age group, we can see that the aging effect is in the direction of increasing the death.



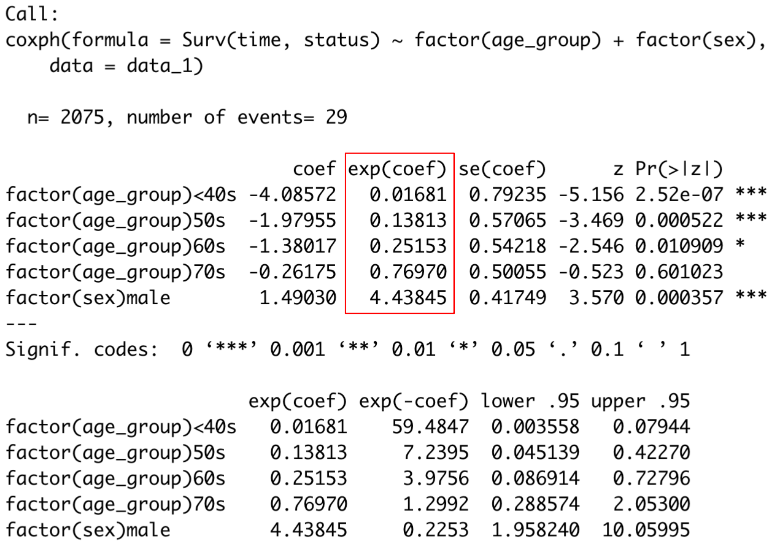
```
Call:
survdiff(formula = Surv(time, status) ~ age_group + strata(sex),
    data = data_1)

                   N Observed Expected (O-E)^2/E (O-E)^2/V
age_group=<40s 1192        2    17.35   13.5787   34.0177
age_group=>80s  103        8     1.21   37.9809   39.8112
age_group=50s   412        5     5.38    0.0273    0.0336
age_group=60s   246        6     3.50    1.7786    2.0269
age_group=70s   122        8     1.55   26.7819   28.3664

 Chisq= 80.6  on 4 degrees of freedom, p= <2e-16
```

Figure 3. The output of log-rank test stratified by the gender

b. Cox proportional hazard (PH) model

The form of the Cox-PH model is as follows: $\lambda(t|Z) = \lambda_0(t)e^{\beta^T Z}$, where $\beta = (\beta_1, \beta_2)^T$ is a 2-dimensional vector of regression coefficients, Z denotes two covariates (namely age group and gender) and $\lambda_0(t)$ is a baseline hazard function. We fit the Cox-PH model with age groups (with $\geq$80s as reference groups) and gender as covariates and obtain the summary table below.

Table 5 The summary table of Cox-PH model



```
Call:
coxph(formula = Surv(time, status) ~ factor(age_group) + factor(sex),
    data = data_1)

  n= 2075, number of events= 29

                            coef exp(coef) se(coef)      z Pr(>|z|)
factor(age_group)<40s -4.08572   0.01681  0.79235 -5.156 2.52e-07 ***
factor(age_group)50s  -1.97955   0.13813  0.57065 -3.469 0.000522 ***
factor(age_group)60s  -1.38017   0.25153  0.54218 -2.546 0.010909 *
factor(age_group)70s  -0.26175   0.76970  0.50055 -0.523 0.601023
factor(sex)male        1.49030   4.43845  0.41749  3.570 0.000357 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                      exp(coef) exp(-coef) lower .95 upper .95
factor(age_group)<40s   0.01681    59.4847  0.003558   0.07944
factor(age_group)50s    0.13813     7.2395  0.045139   0.42270
factor(age_group)60s    0.25153     3.9756  0.086914   0.72796
factor(age_group)70s    0.76970     1.2992  0.288574   2.05300
factor(sex)male         4.43845     0.2253  1.958240  10.05995
```

In particular, the “coef” column in Table 5 contains the estimated regression parameter $\hat{\beta}$ and the “exp(coef)” column represents the estimated hazard ratios exp($\hat{\beta}$). For instance, we can see that the age group of ≤40s is 0.017 times as likely to die as the age group of ≥80s, adjusting for the gender. We observe that the exp(coef) for age groups (≤40s, 50s, 60s, 70s) are less than 1, which indicates these age groups are less likely to die compared with the reference group, controlling the gender. Also, the men are 4.44 times as likely to die as the women, adjusting for the age group. In addition, to test the effect of age, we conduct the Wald test based on $\left(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4\right)^T$ jointly. Specifically, the null hypothesis is $\hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = \hat{\beta}_4 = 0$ and the resulting χ_4^2 test statistic is 36.87 with a *p*-value of 1.9×10^{-7}, suggesting a significant effect of the age on the survival.

We further conduct two different residual analysis to assess our chosen model and to see if certain assumptions are violated. The first is Cox-Snell residual analysis. For Cox-Snell residuals, under correctly specified model, the points would cluster tightly around the straight line, which is similar to Q-Q plot. In this study, we calculate the Cox-Snell residuals using relationship with the martingale residuals and plot the log-transformed Nelsen-Aalen estimates in Figure 4. We can see that the points are mostly clustered around the straight line (except for some at earlier times), indicating that overall the model fits the data reasonably well.

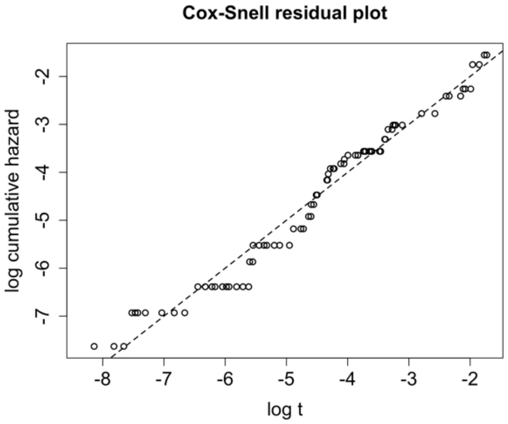
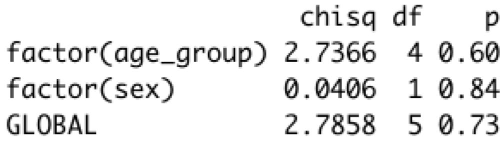


Figure 4 Nelsen-Aalen estimates of Cox-Snell residuals

The second is Schoenfeld residual analysis, which provides information about the proportionality on the covariates. The residual plots are presented in Figure 5 and the test results are in Table 6. From both the graphical and numerical results with *p*-values larger than 0.05, it seems that there is no substantial non-proportionality in the covariates. The global proportionality test is also non-significant. In summary, the residual analysis validates the assumptions of our Cox-PH model and there is no obvious violation on model assumptions.

Table 6 Test results of Schoenfeld residual analysis



```
                    chisq df    p
factor(age_group) 2.7366  4 0.60
factor(sex)       0.0406  1 0.84
GLOBAL            2.7858  5 0.73
```

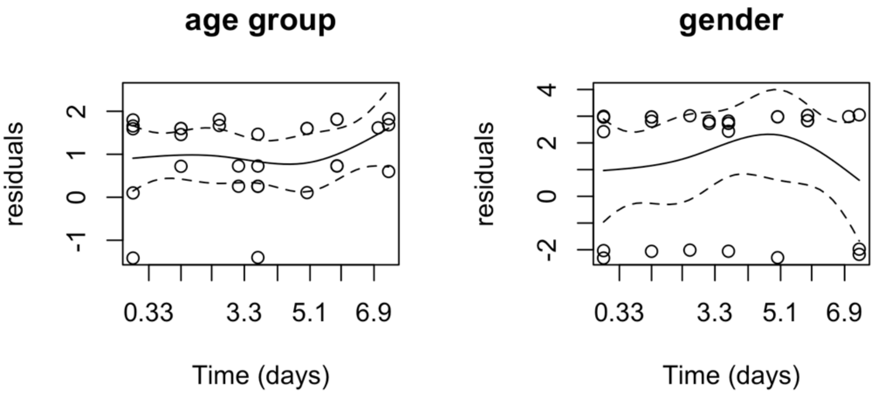


Figure 5 Scaled Schoenfeld residuals for each covariate

5. Summary

Based on the public dataset from Korea CDC, we carried out a comprehensive survival analysis to assess the age difference in the case fatality rate. From our analysis, we obtain several important insights. First, the men are much more likely to die from COVID-19 than the women. In detail, the fatality rate is more than 0.023 for men but less than 0.007 for women among 2075 confirmed cases. Second, the fatality rate for the elder people who are infected by the coronavirus is much higher than that for other young groups. COVID-19 kills an estimated 7.8% of patients 80 and older, compared to 1.2% of those in their 50s and 0.2% of those in their 40s or younger. Therefore, based on several statistical tests, we conclude that the age difference has a significant effect on the fatality rate. However, the age difference may be not only a contributor to COVID-19 death but also chronic illnesses and declining immunities, which needs further study.