Introduction

Study Design

Covid-19 has been lasted for more than 1 years, which has been impacting our society in many perspectives. As time goes by, the case rate has been rising up quite fast. There're a lot of factors associated with it, like population, unemployment rate, elderly people rate and mask use rate. In this project, I'm interested in exploring the correlation between these factors and the covid case rate in each county in the US, in terms of event and time using survival methods.

The study designations are as follows:

subject: each county in the US

event: reach a 10% the case rate

start: the 1st case confirmed date in each county

end: either the event happen date or the last day from the dataset (02-01-2021) for each county

time: date difference between end and start for each county

status: either event happened (1) or still censoring (0)

covariates: interest factors like population, unemployment, elderly people rate,education, mask use

Variables

case rate: the confirmed covid cases comparing with the population in each county

fips: unique code for each county in the US

time: number of days since the 1st case confirmed date for each county

status: either event happened or not accomplishing with the time

65+ rate: the rate of people aged 65 or older

bachelor+ rate: the rate of people with bachelor or higher degrees

unemployment rate: the percentage of people unemployed

mask rate: the rate of people frequently or always wearing masks in public places

population: population in each county by thousand

Methodology

Exploratory Analysis

- Summary Statistics: 5 statistics of case rate at 02-01-2021
- Histograms and Boxplots: histogram boxplots visualizing the distributions of covariates
- Pairwise Scatterplot: comparisons between covariates
- Multicollinearity Checking: VIFs for the covariates

Statistical Analysis

Survival Analysis

- Cox Proportional Hazard Model: protective effects in terms of the factors considered
- Model Diagnosis: proportional hazard assumption, outlier checking, nonlinearity checking

Results

Exploratory Analysis

summary statistics

The case rate in each county at the last date (02-01-2021) in the dataset is distributed like the following histogram and boxplot in figure1, also with the 5 statistics showing in the table1:

figure1:

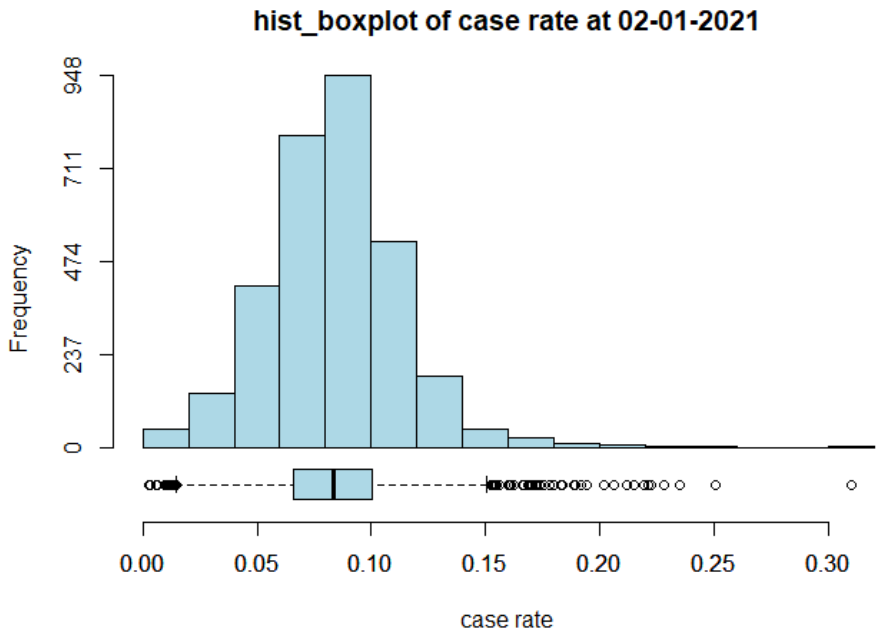


table1:

	min	1st quarter	median	3rd quarter	max
case rate	0.002509	0.065901	0.0832588	0.100236	0.310003

From figure1 and table1, we can see that at 02-01-2021, the US county level case rate were nearly symmetrically distributed, with slight right skewness of extreme large case rate values. While most of the counties were with lower than 10% the case rate, more specific, there're about 75% the case rate less than 10%, while 25% have already reached 10% the case rate before. A frequency table showing the event censoring status at 02-01-2021 is as table2:

table2:

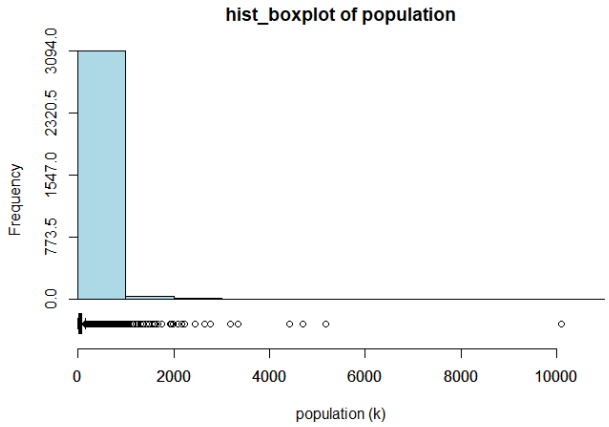
event status 02-01-2021	happened	censoring
frequency (rate)	797 (25.42%)	2338 (74.58%)

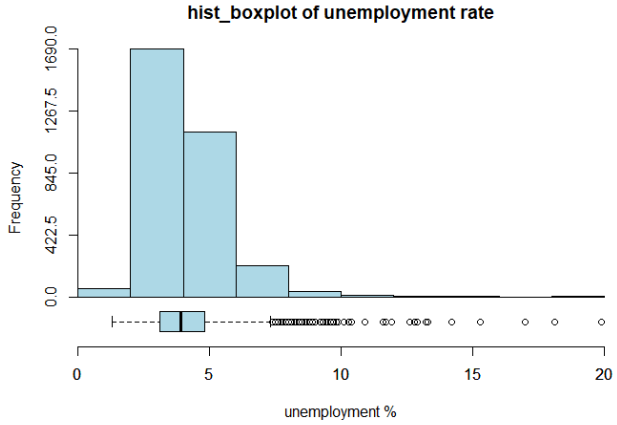
From table2, we can see that there're 797 (about 25.42%) the counties already reached the 10% case rate, while 2338 (about 74.58%) didn't reach the event yet.

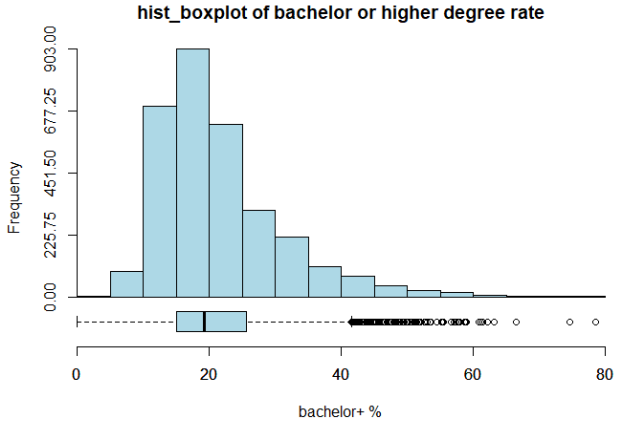
histograms & boxplots

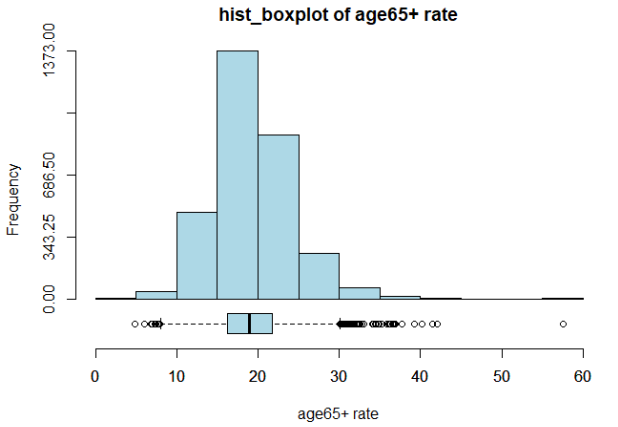
The histograms and boxplots showed the overall distributions of the covariates I was interested in terms of their impacts to the covid case rate in each county, and the graphs are displayed in figure2 as followed:

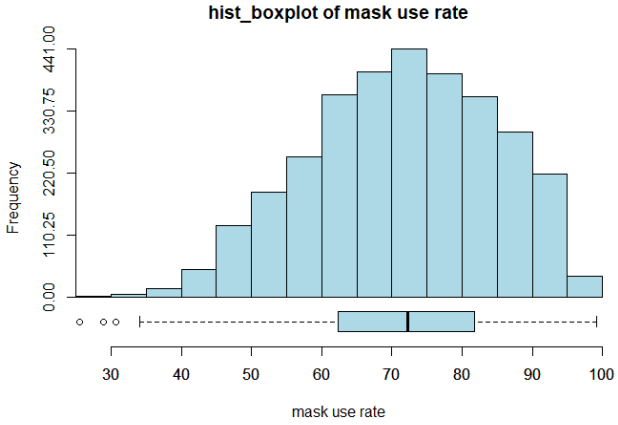
figure2:


hist_boxplot of population
Frequency
0.0
773.5
1547.0
2320.5
3094.0
0
2000
4000
6000
8000
10000
population (k)


hist_boxplot of unemployment rate
Frequency
0.0
422.5
845.0
1267.5
1690.0
0
5
10
15
20
unemployment %


hist_boxplot of bachelor or higher degree rate
Frequency
0.00
225.75
451.50
677.25
903.00
0
20
40
60
80
bachelor+ %

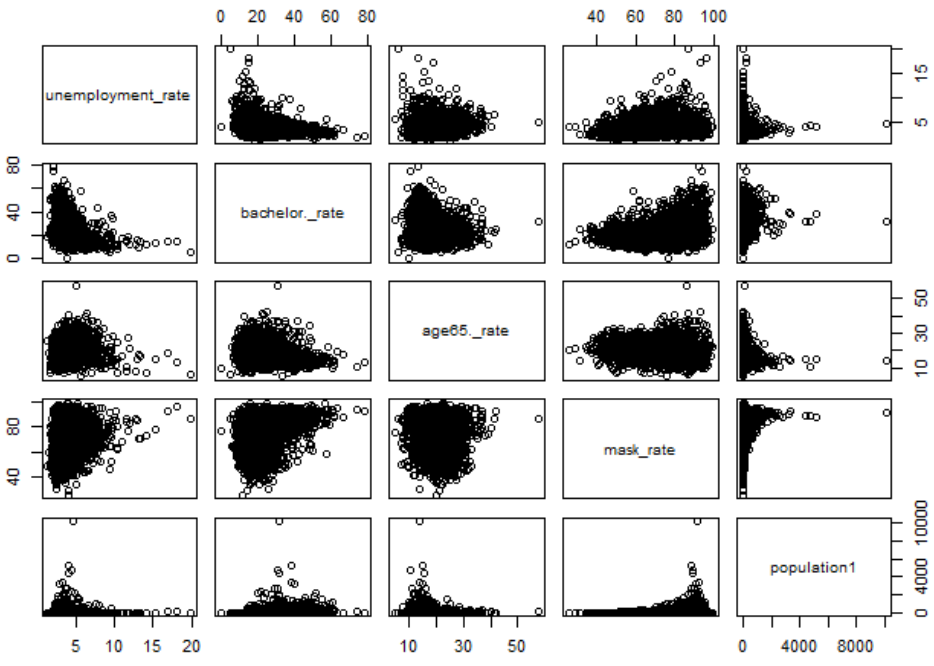




From figure2, we can see that the distributions of population, unemployment rate and bachelor or higher degree rate among US counties are not so symmetric, they are all right skewed. The rate of elderly people aged 65 or higher and the frequently/always using of masks rate among US counties can be nearly symmetrically distributed.

pairwise scatterplot

figure3:



From figure3, we can see that the population can correlated closely with other covariates, it is likely that as population increases, other factors seem to be less variated. Besides, the unemployment rate and bachelor or higher degree rate are likely to be correlated, the higher bachelor or higher degree rate may related to a less variate in unemployment rate.

multicollinearity checking

The VIF values are able to check the multicollinearity from a model. Here in this project, the factors I considered are all continues values, therefore VIF value for each factor was computed and displayed in the following table3:

table3:

	unemployment rate	bachelor+ rate	age65+ rate	mask rate	population (k)
VIFs	1.2486	1.3764	1.0685	1.2022	1.1412

From table3, we can see that these VIFs are all less than $\sqrt{5}$, which means the multicollinearity is not a problem.

Statistical Analysis

models

The model I used was the Cox Proportional Hazard Model, where the format for the model type is like this way:

$$h(t) = h_0(t)exp(b_1x_1 + \ldots + b_px_p)$$

Which corresponds to the proportional log(hazard) effects from risk factors.

Considering that the mask use rate data was based on the survey during the February of 2020, which may updates afterwards. Though I'm still so interested in this factor, there still exist some bias, which may impact the availability of the model. Therefore, I explored 2 models, with one included all risk factors I talked before, and the other one exclude the mask use rate factor. The models are as follows:

Model1:

$$Coxph(Surv(time, status)) = UnemploymentRate + BachelorRate + Age65Rate + MaskRate + Population$$

Model2:

$$Coxph(Surv(time, status)) = UnemploymentRate + BachelorRate + Age65Rate + Population$$

For model1, the estimated model was with the following coefficients, see table4:

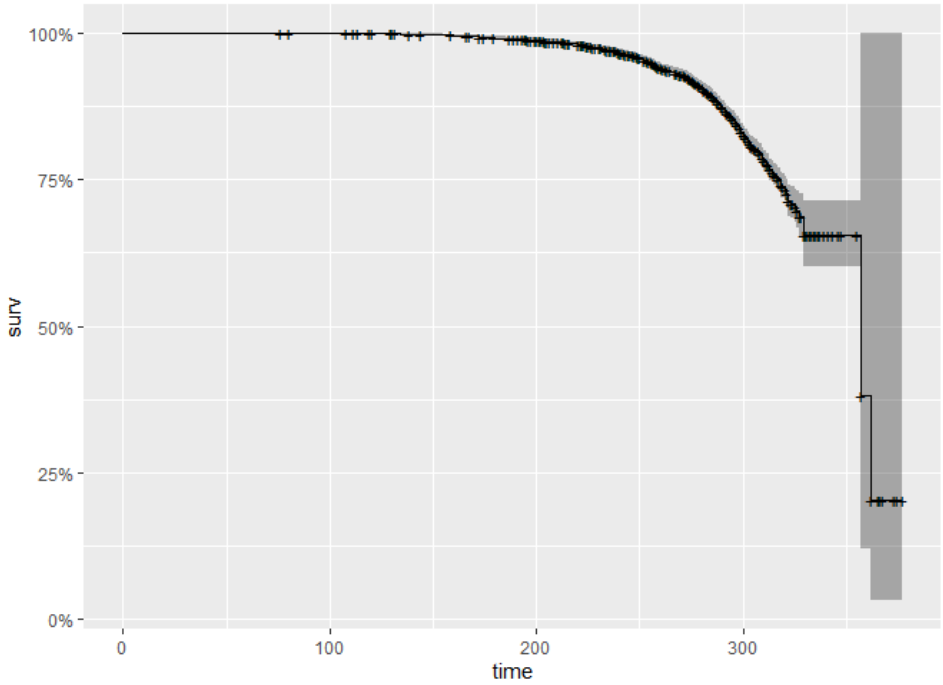
table4:

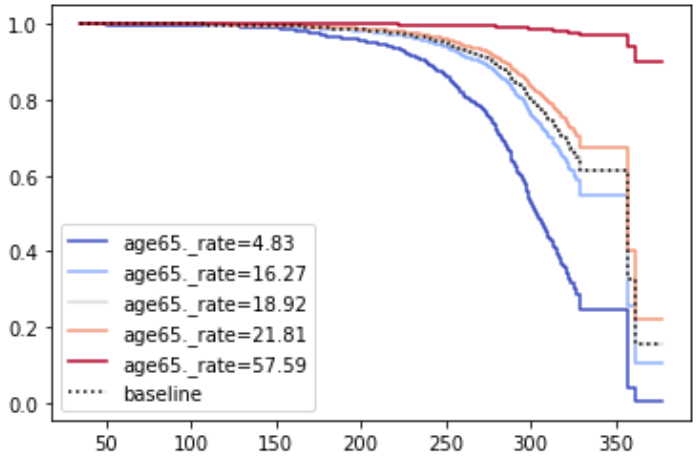
	coef	p-value
unemployment rate	-0.1389	0.00002 ***
bachelor rate	-0.0490	$<2 * 10^{-16}$ ***
age65 rate	-0.0747	$4.23 * 10^{-14}$***
mask rate	-0.0440	$<2 * 10^{-16}$***
population (k)	-0.0001	0.413

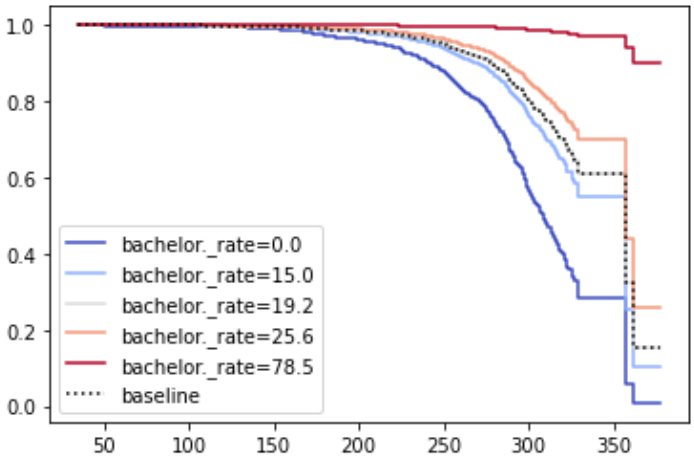
From table4, we can see that the unemployment rate, bachelor rate, age65 rate and mask rate are significantly correlated to the hazard, while population effect is not so significant. Besides, the coefficients are all negative, which means these the value of these risk factors increases, the hazards will decrease, therefore, these factors can all provide protective effects in terms of reaching the 10% case rate, and related to longer lasting time.

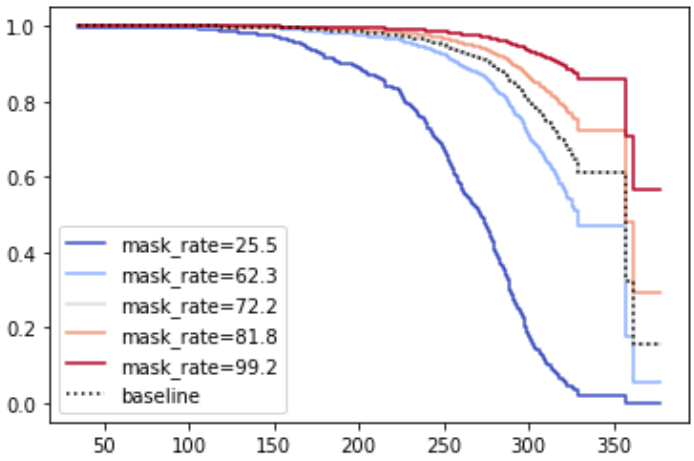
The plot representing the status of event along time is like the figure4, where the first plot was the overall status along time, while the rest plots showing status along time with partial effects from significant factors in terms of their 5 numbers of quartiles.

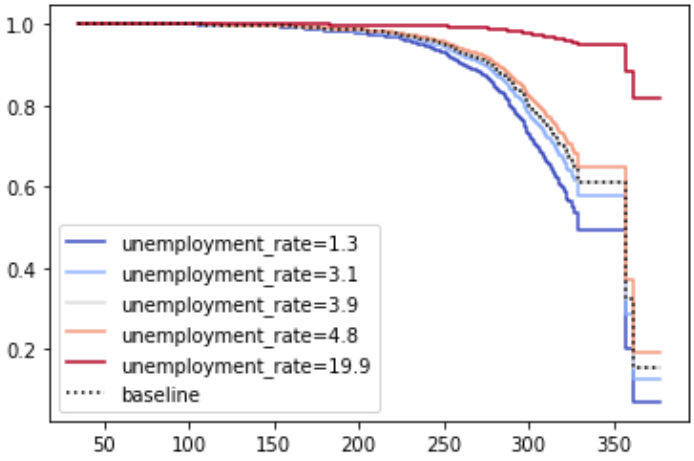
figure4:











Similarly, if excluding the effects from mask uses, the model2 will be like table5:

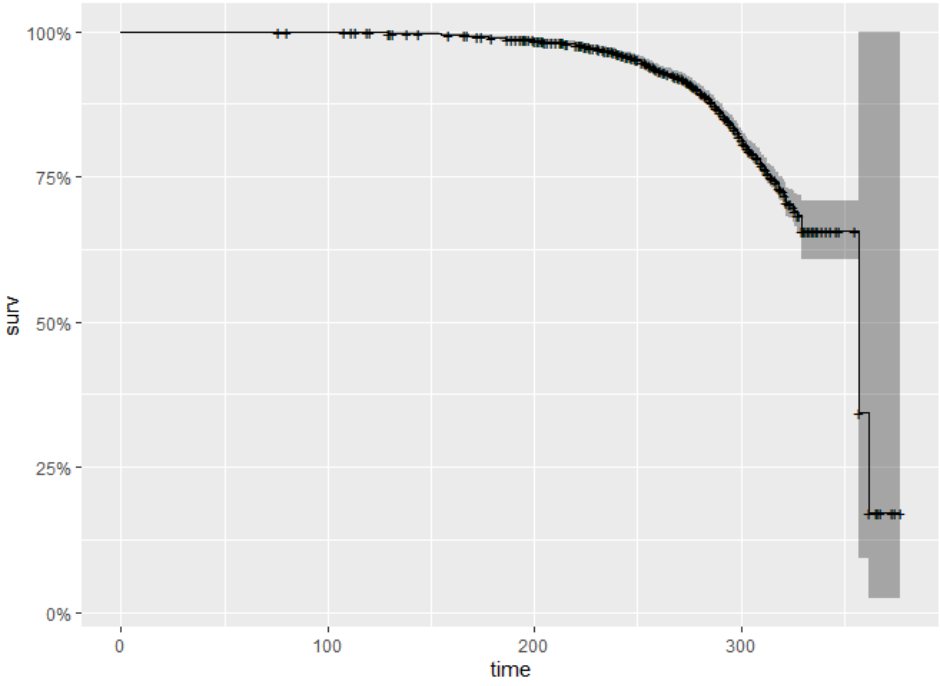
table5:

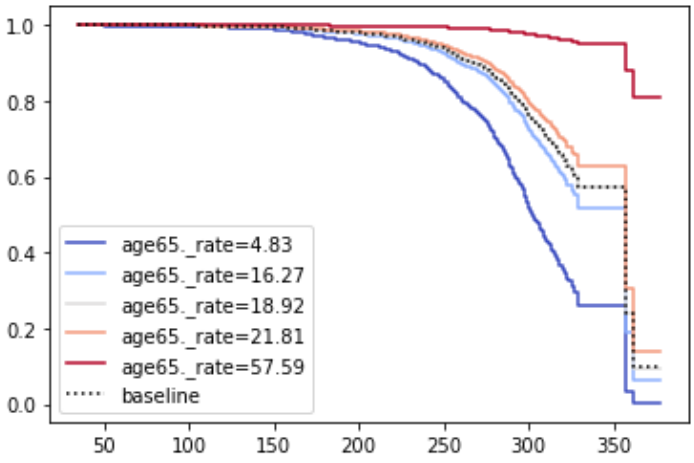
	coef	**p-value**
unemployment rate	-0.2609	$2.61 * 10^{-15}$ ***
bachelor rate	-0.0704	$< 2 * 10^{-16}$ ***
age65 rate	-0.0623	$8.4 * 10^{-11}$ ***
population	-0.0005	0.0251 *

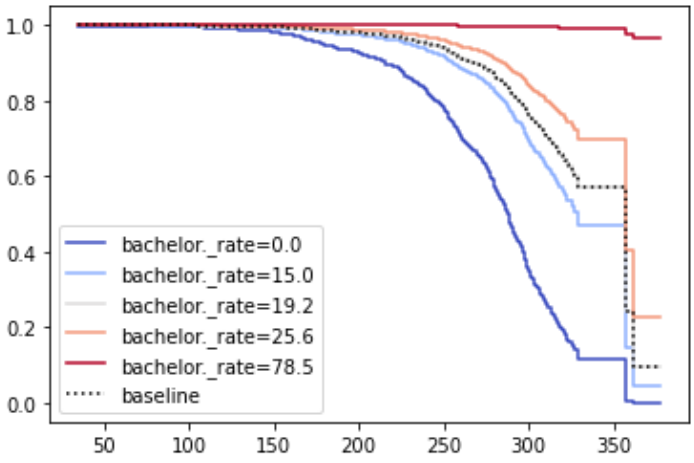
From table5, we can see that all these factors (unemployment rate, bachelor rate, age65 rate and population) are significant in terms of having proportional hazard to the time till the event of a 10% the case rate happen. Besides, the estimated coefficients are all negative, which means these factors can all provide protective effects to have longer time till appearing the event of 10% case rate.

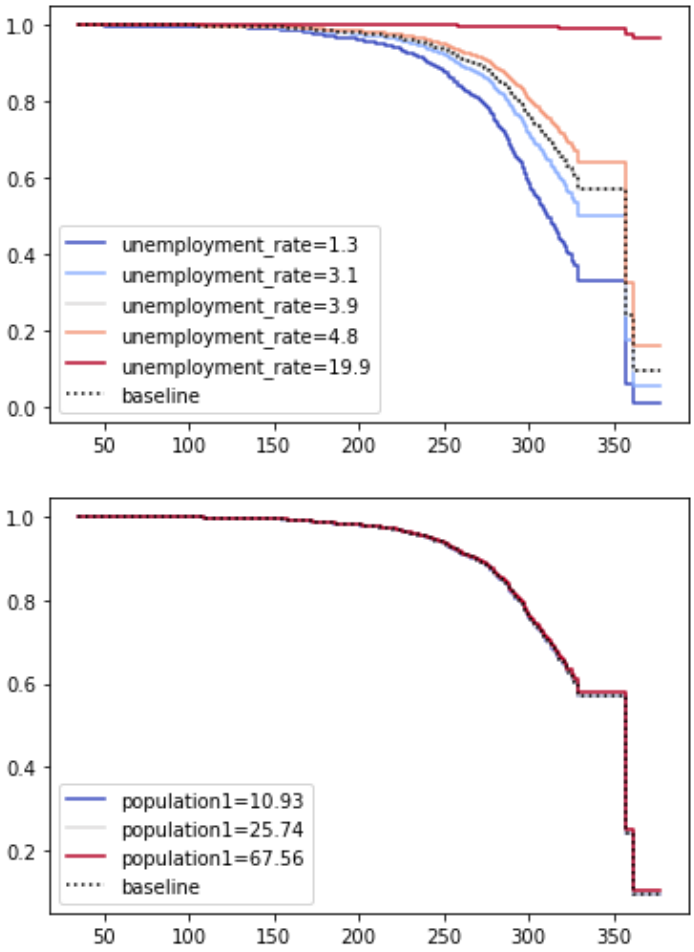
Also, the plot showing the status of event along time is like the figure5:

figure5:









From figure5, we can see that, though population factor was significant here, the impact from it still very weak. For other factor, different percentile of the rates may impact the status of event along time strongly.

model comparison

These two cox proportional hazard models were pretty similar, and the only difference is one with mask rate regressor, while one without mask rate. To see whether the factor mask rate impact the fitness of the models. Some model comparison was made here. Likelihood ratio test can actually test the performance of these two models, where the null hypothesis was that the unrestricted model was better. Here in table6 is the outcome of this test:

table6:

model	log likelihood	chisq	df	p-value
model2	-5995.4			
model1	-5901.2	188.44	1	$< 2.2 * 10^{-16}$

Here, model1 has more factors considered, thus it is restricted model.

$$LR = 2 * (l(\hat{\theta}) - l(\theta_0)) = 2 * (-5901.2 + 5995.4) = 188.44$$

The difference of degrees of freedom is 1, so we test the statistic with χ_1^2 distribution, which eventually get a p-value<0.05. Therefore, the null hypothesis been rejected and the restricted model seems to be better. Here in this condition, that is the model1 (with mask_rate) is better.

model diagnosis

proportional hazard assumption:

Cox proportional hazard models need to meet the assumption of proportional correlations between log(hazard) and risk factors, which can be check from the chi-square test of Schoenfield residuals. Here in figure6 and table7, it shows the results of the test outcome:

figure6:

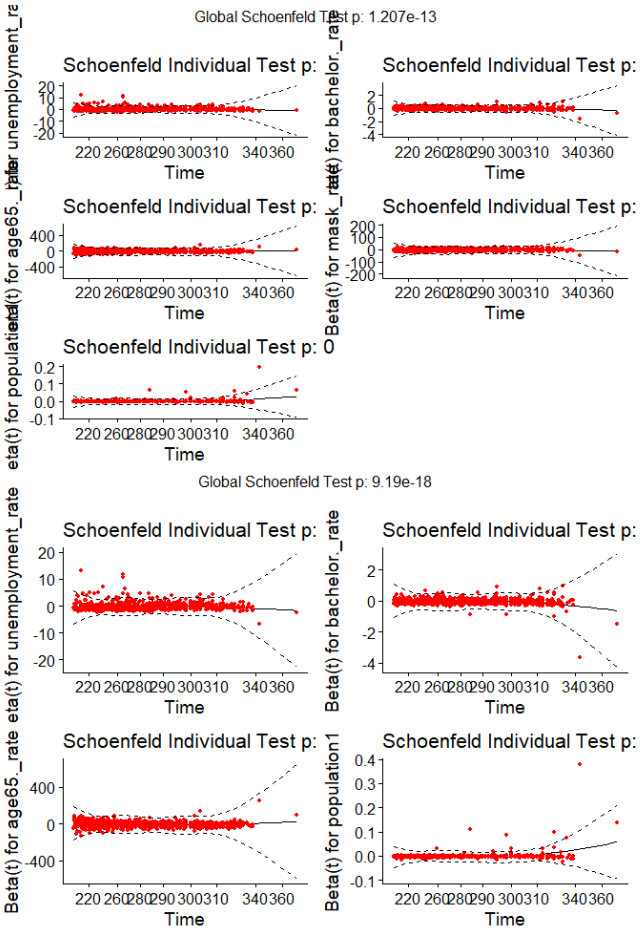


table7:

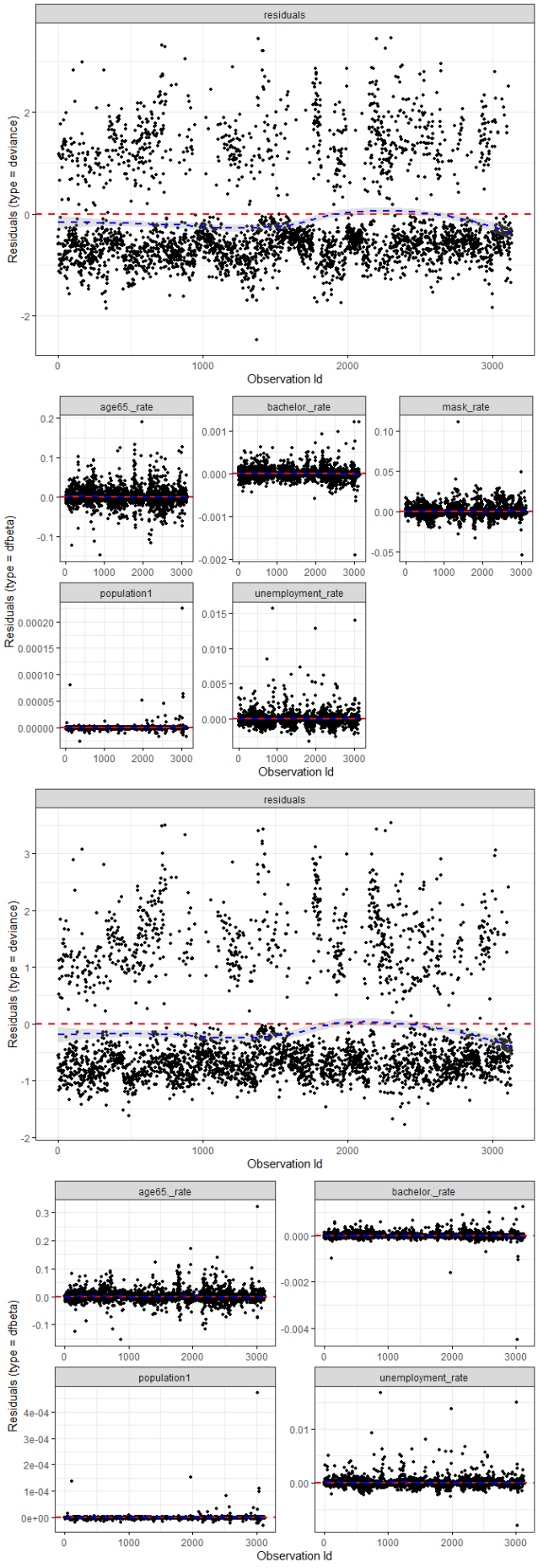
	p-value (model1)	p-value (model2)
unemployment rate	0.0363	0.014
bachelor rate	0.1572	0.08
age65 rate	0.0002	$3.8 * 10^{-5}$
mask rate	$3 * 10^{-8}$	N/A
population	$6.4 * 10^{-7}$	$6.4 * 10^{-14}$
Global	$1.2 * 10^{-13}$	$< 2 * 10^{-16}$

From figure6 and table7, we can see that the model are overall meeting the assumption of proportional hazard for both model1 and model2. However, when focusing on the individual factors, we can find that for both models, the unemployment rate and bachelor rate are not so significantly meet the proportional hazard assumption, while the rest factors meet the assumption.

outlier checking:

The outlier checking is also part of the model diagnosis. For cox model, the standard normal distributed deviance residuals can be used to check it. In figure7, it shows the deviance residuals overall and in terms of each factor.

figure7:

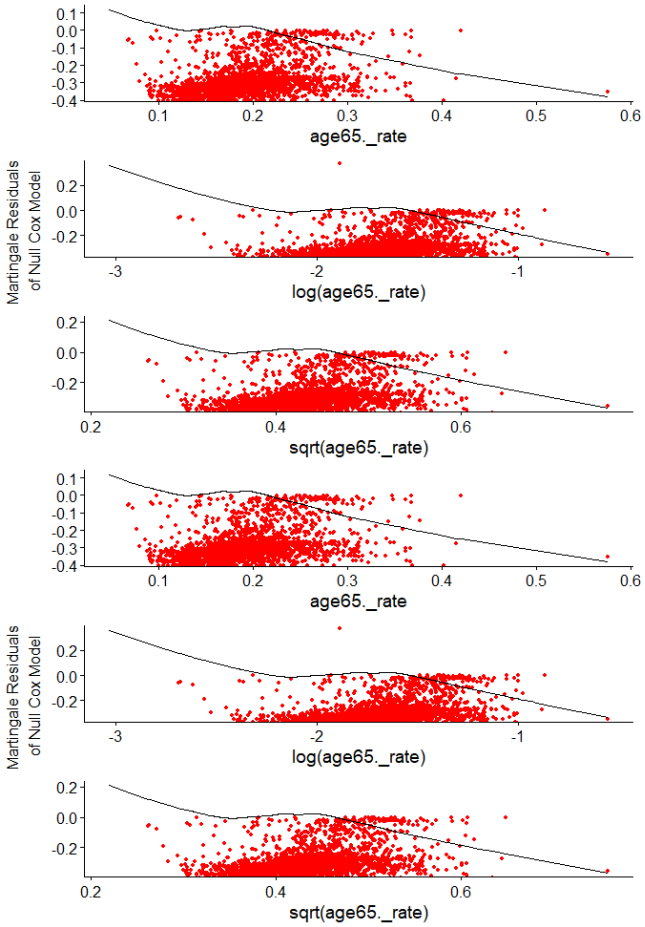


From figure7, we can see that either the overall residuals or the individual factor residuals are with quite small variance for both models, which means no outlier problems.

nonlinearity checking:

As we see from the figure7, in terms of individual factor residuals, it seems that the age65 rate factor are with the highest variation, which to a certain extent could leading some concerns on nonlinearity. Where I tried several transformations and visualized the martingale residuals to check the nonlinearity effects from age65 rate. Here shows the results in figue8:

figure8:



From figure8, we can see that there do exist slightly nonlinearity issues in terms of age65 rate for both models, even some transformations like log or sqrt been made, the problems still not been solved. While the nonlinearity problems are not so serious, so overall it will be fine.

Conclusions

In terms of the time to reach the event that having 10% the case rate for a county in the US, the factors like unemployment rate, bachelor or higher degree rate, age 65 or older rate can significantly impact the happening of the event regardless considering the mask use rate or not. If considering the effects from mask use, the population impact becomes not significant, otherwise, the population impact is significant, which may because of the joint effects between population and mask wearing. From model diagnosis, in both models, the proportion hazard assumptions on unemployment rate and bachelor degree rate not been met, which may cause some bias in predicting, though the model overall met the proportional hazard assumption requirement. The outlier problems and nonlinearity problems are fine for both models.

References

https://stats.idre.ucla.edu/sas/seminars/sas-survival/

https://rviews.rstudio.com/2017/09/25/survival-analysis-with-r/

http://www.sthda.com/english/wiki/cox-model-assumptions

https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Survival/BS704_Survival6.html

Appendix

```
library(reshape2)
library(survival)
library(survminer)
library(lubridate)
library(dplyr)
library(ggplot2)
library(plotly)
library(ggfortify)
library(packHV)

df = read.csv('us_counties_covid-19.csv')
df1 <- read.csv('us_counties_covariates.csv')
df2 <- read.csv('mask-use-by-county.csv')

df$fips <- as.factor(df$fips)
df$date <- as.Date(df$date)

sum(is.na(df$fips))
df <- na.omit(df)

## case by fips & date
df_temp1 <- df%>%group_by(fips, date)%>%summarise(cases=cases, deaths=deaths)
df_mask <- df2
df_mask$mask_rate <- df_mask$FREQUENTLY+df_mask$ALWAYS
df_mask_temp <- df_mask[,c('COUNTYFP', 'mask_rate')]
df_pop_temp <- df1[,c('FIPS', 'POP_ESTIMATE_2018')]

df_case_pop <- merge(df, df_pop_temp, by.x = 'fips', by.y = 'FIPS')
# sum(is.na(df_case_pop$fips))  # check for missing

#### --------------- revised event censoring data -------------- ####

df_event_censoring <- df_case_pop%>%group_by(fips)%>%
  summarise(t=min(date[cases>POP_ESTIMATE_2018*0.1]),t0=min(date),t1=max(date))
# sum(!is.finite(df_event_censoring$t))
df_event_censoring$time <-
ifelse(is.finite(df_event_censoring$t),df_event_censoring$t-
df_event_censoring$t0, df_event_censoring$t1-df_event_censoring$t0 )
df_event_censoring$status <- ifelse(is.finite(df_event_censoring$t), 1, 0)

keys_covarites <- c('FIPS', 'POP_ESTIMATE_2018', 'Total_age65plus',
                    'Unemployment_rate_2018',
'Percent.of.adults.with.a.bachelor.s.degree.or.higher.2014.18')
df_cov_temp <- df1[,keys_covarites]
names(df_cov_temp)[2:5] <- c('population', 'age65+', 'unemployment_rate',
'bachelor+_rate')
# head(df_cov_temp)
df_cov_temp$`age65+_rate` <- df_cov_temp$`age65+`/df_cov_temp$population

```

```r
df_cov_temp1 <- merge(df_cov_temp, df_mask_temp, by.x='FIPS', by.y = 'COUNTYFP')
# head(df_cov_temp1)

df_event_cov_temp <- merge(df_event_censoring, df_cov_temp1, by.x = 'fips', by.y = 'FIPS')
# head(df_event_cov_temp)
df_event_cov_temp <- df_event_cov_temp[,-c(2,3,4)]
df_event_cov_temp$population1 <- df_event_cov_temp$population/1000
df_event_cov <- df_event_cov_temp[,-c(4,5)]
head(df_event_cov)
write.csv(df_event_cov, 'covid_clean_revised.csv', row.names = F)

#### ------------- EDA ---------------- ####
### reload data
covid_clean <- read.csv('covid_clean_revised.csv')
head(covid_clean)
# pairwise
pairs(covid_clean[,-c(1,2,3)])
## multicollinearity
car::vif(coxph(Surv(time, status)~.-fips, data=covid_clean))

## Data summary
## case rate at 02-01-2021
df_case_rate <- df_case_pop[df_case_pop$date==as.Date('2021-02-01'),c('fips', 'cases', 'POP_ESTIMATE_2018')]
df_case_rate$rate <- df_case_rate$cases / df_case_rate$POP_ESTIMATE_2018
head(df_case_rate)
packHV::hist_boxplot(df_case_rate$rate, col='lightblue',
             main = 'hist_boxplot of case rate at 02-01-2021',
             xlab = 'case rate')
summary(df_case_rate$rate) # the case rate 5 numbers of statistics up to Feb 1st

hist_boxplot(df_case_rate$rate, col='lightblue',
                     main = 'Histogram & Boxplot of Case Rate in 02/01/2021',
                     xlab = 'case rate')
summary(df_case_rate$rate) # the case rate 5 numbers of statistics up to Feb 1st

hist_boxplot(covid_clean$population1, col='lightblue',
             main='hist_boxplot of population',
             xlab = 'population (k)')

hist_boxplot(df_cov_temp1$unemployment_rate, col='lightblue',
             main='hist_boxplot of unemployment rate',
             xlab = 'unemployment %')

hist_boxplot(df_cov_temp1$`bachelor+_rate`, col='lightblue',
             main='hist_boxplot of bachelor or higher degree rate',
             xlab = 'bachelor+ %')

hist_boxplot(df_cov_temp1$`age65+_rate`, col='lightblue',
             main='hist_boxplot of age65+ rate',
             xlab = 'age65+ rate')

hist_boxplot(df_cov_temp1$mask_rate, col='lightblue',
             main='hist_boxplot of mask use rate',
             xlab = 'mask use rate')

#### ------------- Statistical Analysis -------------- ####
## Cox model
## with mask_rate
cox_model1 <- coxph(Surv(time, status)~.-fips, data = covid_clean)
```

```r
summary(cox_model1)
cox_fit1 <- survfit(cox_model1)
autoplot(cox_fit1)

## without mask_rate
cox_model2 <- coxph(Surv(time, status)~.-fips-mask_rate, data=covid_clean)
summary(cox_model2)
cox_fit2 <- survfit(cox_model2)
autoplot(cox_fit2)

#### ------------- Model Diagnosis ------------- ####
## cox_model1
#test for the proportional-hazards (PH) assumption
cox_ph1 <- cox.zph(cox_model1)
cox_ph1
ggcoxzph(cox_ph1)

## outliers & influential obs
# influential observations
ggcoxdiagnostics(cox_model1, type = "dfbeta",
                 linear.predictions = FALSE, ggtheme = theme_bw())

# outliers, need symmetric, kinda positively skewed, lots of too long
ggcoxdiagnostics(cox_model1, type = "deviance",
                 linear.predictions = FALSE, ggtheme = theme_bw())

# nonlinearity, martingale, (-infin, 1), from dfbeta residuals, check for the largest residual covariate, here's age65_rate
ggcoxdiagnostics(cox_model1)
ggcoxfunctional(Surv(time, status)~`age65._rate`+log(`age65._rate`)+sqrt(`age65._rate`), data = covid_clean)
# slightly nonlinearity

## cox_model2
#test for the proportional-hazards (PH) assumption
cox_ph2 <- cox.zph(cox_model2)
cox_ph2
ggcoxzph(cox_ph2)

# influential observations, age65_rate largest residual deviance
ggcoxdiagnostics(cox_model2, type = "dfbeta",
                 linear.predictions = FALSE, ggtheme = theme_bw())

# outliers, need  symmetric, kinda positively skewed
ggcoxdiagnostics(cox_model2, type = "deviance",
                 linear.predictions = FALSE, ggtheme = theme_bw())

# nonlinearity, same as model1
ggcoxfunctional(Surv(time, status)~`age65._rate`+log(`age65._rate`)+sqrt(`age65._rate`), data = covid_clean)

lmtest::lrtest(cox_model2, cox_model1)
```