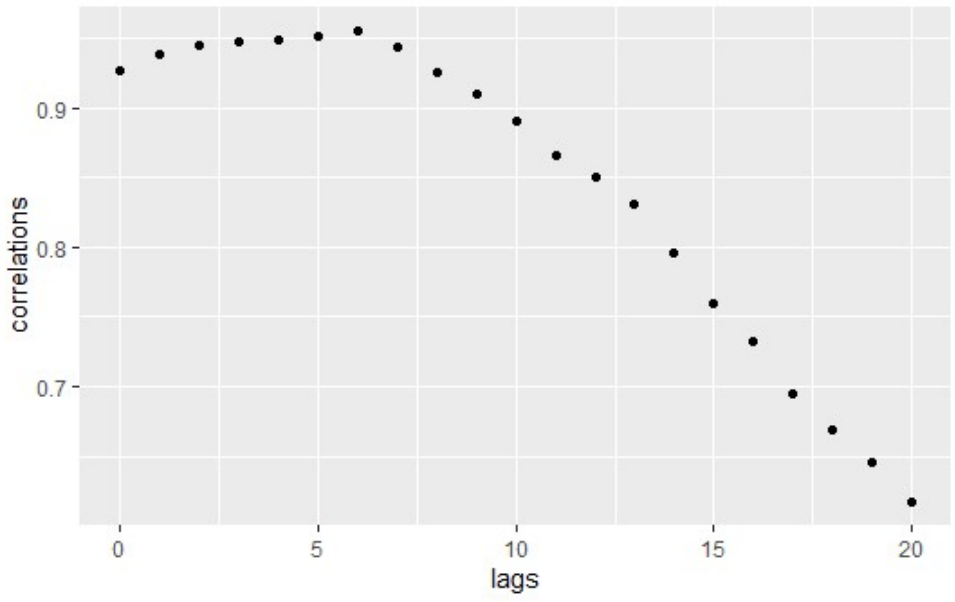
What is the daily correlation of **Confirmed** versus **Death** Cases in Covid-19. In other words, the people who have passed away, on average, how many days ago they have been reported (i.e. “Confirmed”) as Covid-19 new cases.

To answer this question, we can take the correlation between the Daily Confirmed vs Daily Deaths and trying different lag values of the confirmed cases, since the assumption is that it will take some days for someone to pass away since has been diagnosed with Covid-19.

The problem with the data is that are affected by the number of tests and also during some days like weekends they do not report all the cases. This implies that our analysis is not valid, but we will try to see what get. We will analyze **Italy**.

Italy: Correlation Between Confirmed Cases and Deaths

```
df<-coronavirus%>%filter(country=='Italy', date>='2020-02-15')%>%select(date,
country, type, cases)%>%
  group_by(date, country, type) %>%pivot_wider(names_from =type,
values_from=cases) %>%ungroup()

correlations<-c()
lags<-c(0:20)

for (k in lags) {

  tmp<-df%>%mutate(lagk=lag(confirmed,k))%>%select(death,lagk)%>%na.omit()

  correlations<-c(correlations,cor(tmp$death, tmp$lagk))
}

data.frame(lags, correlations)

data.frame(lags, correlations)%>%ggplot(aes(x=lags,
y=correlations))+geom_point()
```



---

lags correlations

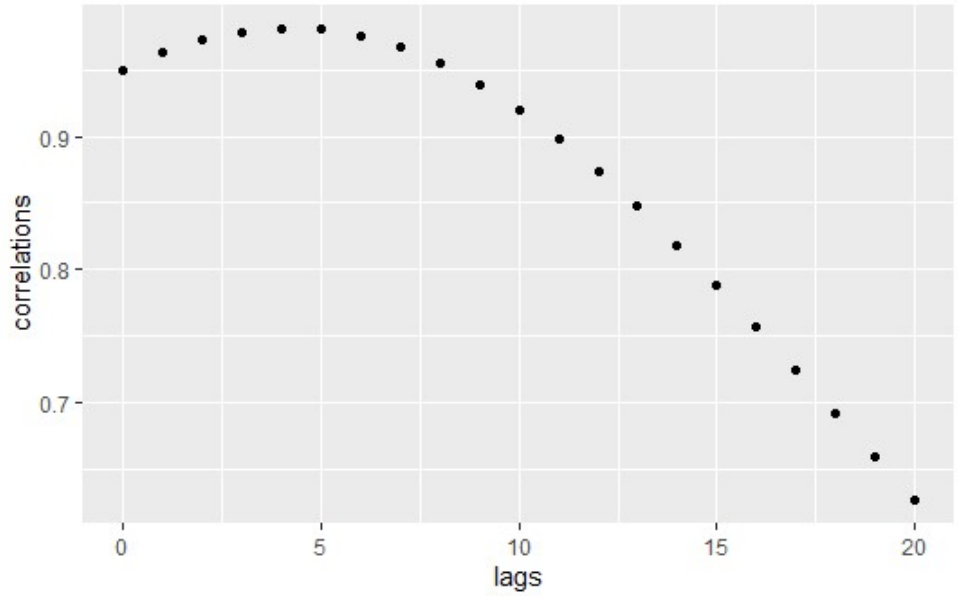
0	92.64%
1	93.78%
2	94.44%
3	94.79%
4	94.92%
5	95.16%
6	**95.53%**
7	94.35%
8	92.58%
9	91.00%
10	89.00%
11	86.64%
12	85.07%
13	83.09%
14	79.59%
15	76.00%
16	73.26%
17	69.52%
18	66.85%
19	64.60%
20	61.73%

As we see, the argmax correlation is at k=6, which implies (if the data were accurate), that from the people who have passed away, most of them diagnosed with Covid-19 **6 days ago**.

Italy: Correlation Between Confirmed Cases and Deaths SMA 5

Let's do the same analysis, but this time by taking into consideration the Simple Moving Average of 5 days.

```
df<-coronavirus%>%filter(country=='Italy', date>='2020-02-15')%>%select(date,
country, type, cases)%>%
  group_by(date, country, type) %>%pivot_wider(names_from =type,
values_from=cases) %>%ungroup()%>%
  mutate(confirmed = stats::filter(confirmed, rep(1 / 5, 5), sides = 1), death =
stats::filter(death, rep(1 / 5, 5), sides = 1))%>%na.omit()

correlations<-c()
lags<-c(0:20)

for (k in lags) {

  tmp<-df%>%mutate(lagk=lag(confirmed,k))%>%select(death,lagk)%>%na.omit()

  correlations<-c(correlations,cor(tmp$death, tmp$lagk))
}

data.frame(lags, correlations)
data.frame(lags, correlations)%>%ggplot(aes(x=lags,
y=correlations))+geom_point()
```



lags	correlations
0	95.00%
1	96.36%
2	97.32%
3	97.90%
4	**98.13%**
5	98.04%
6	97.62%
7	96.77%
8	95.50%
9	93.91%
10	92.00%
11	89.80%
12	87.41%
13	84.77%
14	81.85%
15	78.77%
16	75.63%
17	72.39%
18	69.15%
19	65.94%
20	62.65%

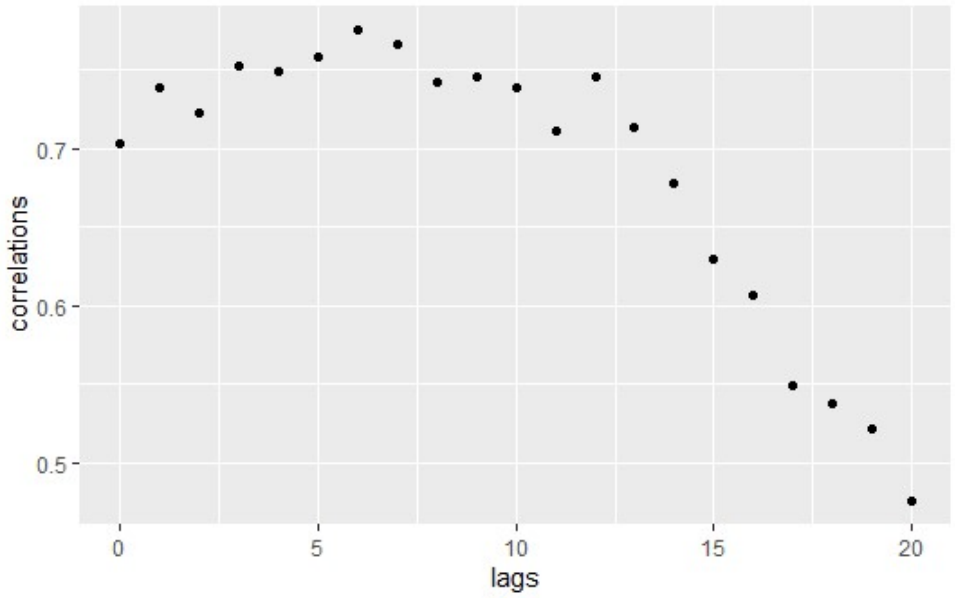
When we consider the SMA of 5 days the maximum correlation **is at day 4**.

Belgium: Correlation Between Confirmed Cases and Deaths

Let's do the same analysis for Belgium.

```
df<-coronavirus%>%filter(country=='Belgium', date>='2020-02-15')%>%select(date,
country, type, cases)%>%
  group_by(date, country, type) %>%pivot_wider(names_from =type,
values_from=cases) %>%ungroup()
```

```
correlations<-c()
lags<-c(0:20)

for (k in lags) {

  tmp<-df%>%mutate(lagk=lag(confirmed,k))%>%select(death,lagk)%>%na.omit()

  correlations<-c(correlations,cor(tmp$death, tmp$lagk))
}

data.frame(lags, correlations)


data.frame(lags, correlations)%>%ggplot(aes(x=lags,
y=correlations))+geom_point()
```



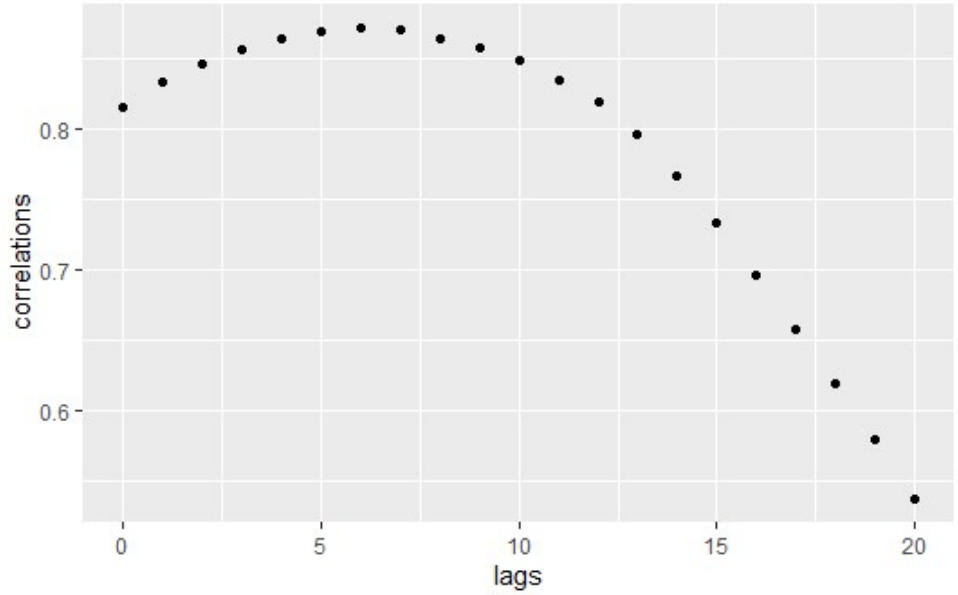
lags	correlations
0	0.703768
1	0.738962
2	0.722847
3	0.752669
4	0.749367
5	0.75888
6	**0.775802**
7	0.766534
8	0.741903
9	0.745851
10	0.739051
11	0.711148
12	0.745839
13	0.714
14	0.677464
15	0.629853
16	0.606283
17	0.549728

18	0.538276
19	0.522196
20	0.47582

Again, in Belgium, the highest correlation between Confirmed cases and Deaths, occurs after **6 days** that people have been reported as new cases.

Finally, let's run the same analysis by taking into consideration the SMA 5.

```
df<-coronavirus%>%filter(country=='Belgium', date>='2020-02-15')%>%select(date,
country, type, cases)%>%
  group_by(date, country, type) %>%pivot_wider(names_from =type,
values_from=cases) %>%ungroup()%>%
  mutate(confirmed = stats::filter(confirmed, rep(1 / 5, 5), sides = 1), death =
stats::filter(death, rep(1 / 5, 5), sides = 1))%>%na.omit()

correlations<-c()
lags<-c(0:20)

for (k in lags) {

  tmp<-df%>%mutate(lagk=lag(confirmed,k))%>%select(death,lagk)%>%na.omit()

  correlations<-c(correlations,cor(tmp$death, tmp$lagk))
}

data.frame(lags, correlations)
data.frame(lags, correlations)%>%ggplot(aes(x=lags,
y=correlations))+geom_point()
```



---

lags	correlations
0	81.53%
1	83.34%
2	84.61%
3	85.66%

4	86.43%
5	86.96%
6	**87.18%**
7	86.98%
8	86.45%
9	85.77%
10	84.80%
11	83.42%
12	81.88%
13	79.58%
14	76.65%
15	73.26%
16	69.58%
17	65.72%
18	61.88%
19	57.94%
20	53.72%

Again, the maximum correlation observed on **the 6th day**.

Discussion

I would like to stress out that this analysis is not valid because we lack much of the information about the way of collecting and reporting the data. However, it is clear that there is a lag between the Confirmed cases and Deaths but we cannot specify the number accurately.