

# Conditional quantiles estimation of the incubation period of COVID-19

**Xiaohui Liu**

Jiangxi University of Finance and Economics

**Lei Wang** (✉ [leiw@mails.ccnu.edu.cn](mailto:leiw@mails.ccnu.edu.cn))

Central China Normal University

**Xiansi Ma**

Jiangxi University of Finance and Economics

**Jiwen Wang**

Jiangxi University of Finance and Economics

---

## Research Article

**Keywords:** COVID-19; Incubation period; Conditional quantiles; Biased sampling

**Posted Date:** April 22nd, 2020

**DOI:** <https://doi.org/10.21203/rs.3.rs-23529/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

## RESEARCH

# Conditional quantiles estimation of the incubation period of COVID-19

Xiao-Hui Liu<sup>1,2</sup>, Lei Wang<sup>3\*</sup>, Xian-Si Ma<sup>1,2</sup> and Jie-Wen Wang<sup>1,2</sup>

\*Correspondence:

[leiw@mails.cnu.edu.cn](mailto:leiw@mails.cnu.edu.cn)

<sup>3</sup>National Engineering Laboratory for Educational Big Data, Central China Normal University, NO.152 Luoyu Road, 430079 Wuhan, Hubei Province, China

Full list of author information is available at the end of the article

## Abstract

**Background:** In December 2019, some cases of pneumonia with unknown etiology were identified in Wuhan, Hubei province in China. The World Health Organization (WHO) has named this disease as COVID-19, standing for “2019 coronavirus disease”, and announced the disease have become a public health incident on December 31, 2019. This study aimed to investigate the conditional distribution of the incubation period of COVID-19 on the age of infected cases, and estimate its corresponding conditional quantiles from information on 2172 confirmed cases from 29 provinces outside Hubei in China.

**Methods:** We collected data including the infection dates, onset dates, and ages of the confirmed cases from the websites of the centres of disease control, or the daily public reports through February 16th, 2020. A new maximum likelihood method was developed to account for the biased sampling, or right truncation, issue of the data as the epidemic is still ongoing. The estimators can be shown to be consistent asymptotically under mild conditions.

**Results:** Based on the collected data, we found that the conditional quantiles of the incubation period distribution of COVID-19 varies over ages. In detail, the high conditional quantiles of people in the middle age group are shorter than those of others. We estimated that the 0.95-th quantile related to people in the age group 23~55 is less than 15 days.

**Conclusions:** Observing that the conditional quantiles vary over ages, we may take more precise measures for people of different ages. For example, we may consider carrying out an age-dependent quarantine duration, rather than a uniform 14-days quarantine, in practice. Remarkably, we may need to extend the current quarantine duration for people aged 0 ~ 22 and over 55 because the related 0.95-th quantiles are much greater than 14 days.

**Keywords:** COVID-19; Incubation period; Conditional quantiles; Biased sampling

## Background

In December 2019, some cases of pneumonia with unknown etiology were identified in Wuhan, Hubei province in China. After investigation by the National Coronavirus Research Group, this pneumonia was identified as caused by a new coronavirus (2019-nCoV). The World Health Organization (WHO) has named this disease as COVID-19, standing for “2019 coronavirus disease” [1].

It turns out that the novel coronavirus, similar to SARS-COV, is the seventh member of the Nidovirales family of coronaviruses [2], but COVID-19 has a shorter serial interval than that of SARS [3] and higher transmissibility than MERS in the Middle East countries [4], nevertheless. It is highly infectious [5] and even contagious

during incubation period [6]. It can cause severe symptoms or even death [7]. The novel coronavirus not only threatens cities in China [8], but also seems to have exploded worldwide. Hence, it is important to take necessary measures to prevent and control it as quick as possible.

In prevention and control efforts, it is well known that the incubation period distribution plays important roles. Knowledge of this distribution can help mathematically model the size of the epidemic [8], predict the time at which the disease will outbreak, and determine the efficacy of medical intervention [9], etc. The pioneering work on deriving the incubation period distribution was conducted by Philip Sartwell in 1950 [10]. After that, the lognormal distribution was widely used to model the incubation period distribution for infectious diseases. Many authors studied the incubation period distributions of various other diseases. Some other distributions, e.g., Gamma distribution and Weibull distribution, were also suggested to fit the observed incubation periods; see for example [9, 11, 12, 13].

In the literature, Li *et al.* [14] first studied the incubation period distribution of COVID-19 based on the early 10 observations in Hubei province in China. Relying on their estimation, Li *et al.* [14] suggested a 14-day medical observation period or quarantine for exposed persons. Guan *et al.*[15] reported the median incubation period, i.e., 3.0 days (range 0 to 24.0 days), of 1099 patients from 552 hospitals in 31 provinces/provincial municipalities through January 29th, 2020. Recently, Backer *et al.*[16] updated this distribution based on the reported travel histories and symptom onset dates of 88 travellers from Wuhan with confirmed 2019-nCoV infection in the early outbreak phase. Backer *et al.*[16] estimated that the 97.5 percentile of the incubation period distribution of COVID-19 is 11.1 days. Linton *et al.*[17] further took into account of the biased sampling issue, and obtained that the estimated 95%-th quantile is greater than 14 days.

However, no existing literature above investigates the distributed characteriser of the incubation period of COVID-19 over people of different ages. Based on 2172 confirmed cases collected outside Hubei provinces in China, a simple ANOVA indicates that the age of confirmed cases has a significant effect on the incubation period of COVID-19. This motivates us to estimate the conditional incubation period distribution on ages. Note that the collected data subject to biased sampling because COVID-19 is still ongoing throughout February 16th, 2020 in China. The current study differs itself from [18], [19], [20] and [21], which investigated the relationship between the age and the incubation period of AIDS, but did not touch the biased sampling issue.

In this study, we developed the conditional quantiles model of the incubation period of COVID-19 on the age of infected cases, and provided the estimating method in detail. The main results were calculated based on the collected data, and the conclusion was presented accordingly.

## Data and methods

In this section, we provide a summary on the collected data, and introduce the estimating method according to the major characters of the collected data.

The data set is taken from the websites of the centres of disease control, or the daily public reports on COVID-19 in 29 provinces outside Hubei province through

February 16th, 2020. It consists of 2172 confirmed cases, including four indexes, i.e., gender, age, onset time and infection time. The incubation period here is calculated by using the formula “*Incubation Period = Onset date – Infection date + 1*”. Note that the default count unit is supposed to be ‘day’ throughout this paper. Among these 2172 cases, there are many cases from Zhejiang, Henan and Anhui than from the other provinces because of the large population of confirmed cases in these provinces. Figure 1 reports the scatter plot of the incubation period of COVID-19 v.s. the age of confirmed cases.

We conduct a preliminary one-way ANOVA study on the incubation period of COVID-19 over four age groups, i.e., 0~17, 18~40, 41~65 and over 65, and find that the age of confirmed cases has a significant effect on the incubation period. Hence, we further investigate the incubation period distribution of COVID-19 conditional on age as follows.

Note that the Weibull distribution fits well the data set, and the mean incubation period varies over people of different ages. We propose to model the relationship between the true incubation period, say  $T$ , and the age, i.e.,  $X$ , through the conditional distribution  $G(t|\lambda(X), \eta)$  on  $X$ . The related density function is specified as follows.

$$g(t|\lambda(X), \eta) = \frac{\eta}{\lambda(X)} \left( \frac{t}{\lambda(X)} \right)^{\eta-1} \exp \left( - \left( \frac{t}{\lambda(X)} \right)^{\eta} \right) I(t \geq 0),$$

where  $I(\cdot)$  denotes the indicator function, and  $\eta > 0$  and  $\lambda(X) := \lambda_3(X) = \mathbf{X}^{\top} \boldsymbol{\beta} > 0$ , where  $\mathbf{X} = (1, X, X^2, X^3)^{\top}$ ,  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\top}$ .

The reasons of using this kind of conditional distribution form are as follows: (i) The conditional mean of  $T$  takes the form  $E(T|X) = \lambda(X)\Gamma(1+1/\eta)$ , which implies that the age  $X$  has a obvious effect on  $E(T|X)$  through  $\lambda(X)$ ; (ii)  $\lambda_3(x)$  is flexible enough to characterize the trend of the change of  $E(T|X)$  over  $X$ . Note that  $\lambda_3(x)$  includes  $\beta_0$ ,  $\beta_0 + \beta_1x$ , and  $\beta_0 + \beta_1x + \beta_2x^2$  as special cases. Here  $\Gamma(\cdot)$  denotes the gamma function.

Furthermore, information from the empirical result shown in Figure 2 indicates that one may model the distribution of the age  $X$  by normal distribution. Write its density as  $\phi(x; \mu, \sigma^2)$ . Then, a natural idea is through maximzing the likelihood function

$$L_T(\boldsymbol{\beta}, \eta, \mu, \sigma^2) = \prod_{j=1}^m \{g(T_j|\lambda(X_j), \eta)\phi(X_j; \mu, \sigma^2)\}^{1/m} \quad (1)$$

to estimate the conditional distribution based on  $\{T_j, X_j\}_{j=1}^m$ .

Unfortunately, the incubation period of some infected cases can not be fully observed when the COVID-19 is still ongoing. The observed incubation period of COVID-19, say  $Y$ , subjects to biased sampling. That is,  $Y$  observed at some fixed time  $t_*$  is not the same as the true incubation period  $T$ . This is because for some case infected at time  $t_S$ , we only can observe such incubation period  $Y$  with  $Y = T$  at time  $t_*$  if  $T \in (0, \Delta]$ , where  $\Delta = t_* - t_S$ . This implies that the distribution of  $Y$  is in fact the conditional distribution depending on the random event  $\{T \in (0, \Delta]\}$ .

That is, we have

$$\begin{aligned}
 F(y|X) &= P(Y \leq y|X) \\
 &= P(Y \leq y|T \in (0, \Delta], X) \\
 &= \frac{P(T \leq y, 0 < T \leq \Delta|X)}{P(T \in (0, \Delta]|X)} \\
 &= \begin{cases} 0, & \text{if } y \leq 0 \\ \frac{G(y|\lambda(X), \eta)}{G(\Delta|\lambda(X), \eta)}, & \text{if } 0 < y < \Delta \\ 1, & \text{if } y \geq \Delta. \end{cases}
 \end{aligned} \tag{2}$$

Denote the collected samples as  $\{Y_i, \Delta_i, X_i\}_{i=1}^n$ , where  $Y_i$ 's denote the observed incubation periods,  $X_i$ 's the ages, and  $\Delta_i = t_* - t_{S,i}$  the difference between the infected time of the  $i$ -case and the observing time  $t_*$ , i.e. February 16th, 2020.

Since the number of infected cases do not grow exponentially throughout February 16th, 2020, it is unreasonable to use the likelihood function developed in [17] again in this paper. Fortunately, note that there are cases infected in almost every day throughout the data collecting time. Hence, it is reasonable to assume that  $\Delta_i$ 's are non-random. Furthermore, note that  $Y_i, X_i$  are independent of the number of infected cases in each day  $t_{S,i}$ . Then, after obtaining  $F(y|X)$  in (2), we propose to use the following likelihood function:

$$\begin{aligned}
 &\ell_Y(\boldsymbol{\beta}, \eta, \mu, \sigma^2) \\
 &= \prod_{i=1}^n \left\{ [g(Y_i|\lambda(X_i), \eta)I(Y_i \in (0, \Delta_i])]^{\frac{G(\Delta_i)}{n}} \times \right. \\
 &\quad \left. \exp\left(\frac{1}{n} \int_{\Delta_i}^{+\infty} g(s|\lambda(X_i), \eta) \log(g(s|\lambda(X_i), \eta)) ds\right) \times \phi(X_j; \mu, \sigma^2)^{\frac{1}{n}} \right\}.
 \end{aligned} \tag{3}$$

Note that  $E(\log(L_T(\boldsymbol{\beta}, \eta, \mu, \sigma^2))) = E(\log(\ell_Y(\boldsymbol{\beta}, \eta, \mu, \sigma^2)))$  when all samples involved are independent, which is mild because all observations in this paper are collected nationwide. It is easy to check that the maximum likelihood estimator based on (3) is asymptotically the same as that based on (1), which is consistent and satisfies the asymptotic normality under some general conditions.

However,  $\ell_Y(\boldsymbol{\beta}, \eta, \mu, \sigma^2)$  contains some integral values

$$\left\{ \int_{\Delta_i}^{+\infty} g(s|\lambda(X_i), \eta) \log(g(s|\lambda(X_i), \eta)) ds \right\}_{i=1}^n,$$

which is computationally difficult. Fortunately, it holds that

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \int_{\Delta_i}^{+\infty} g(s|\lambda(X_i), \eta) \log(g(s|\lambda(X_i), \eta)) ds \\
= & \frac{1}{n} \sum_{i=1}^n (\log(\eta) - \eta \log(\lambda(X_i))) (1 - G(\Delta_i|\lambda(X_i), \eta)) \\
& - \frac{1}{n} \sum_{i=1}^n \left\{ \left[ \left( \frac{\Delta_i}{\lambda(X_i)} \right)^\eta + 1 \right] \exp \left( - \left( \frac{\Delta_i}{\lambda(X_i)} \right)^\eta \right) \right\} \\
& + (\eta - 1) \frac{1}{n} \sum_{i=1}^n \int_{\Delta_i}^{+\infty} \log(s) dG(s|\lambda(X_i), \eta).
\end{aligned}$$

Noting that the estimation of  $\mu, \sigma^2$  is trivial based on  $\{X_i\}_{i=1}^n$ , we focus on how to estimate  $\beta, \eta$  in the sequel. To handle the integral values, we propose to use the EM algorithm as follows. Here suppose that we have obtained some initial estimators  $\hat{\beta}^{(0)}, \hat{\eta}^{(0)}$ , which may be easily computed by pretending  $Y_i$ 's having no bias.

**E-step.** Set  $\beta^{(k)} = \hat{\beta}^{(0)}, \eta^{(k)} = \hat{\eta}^{(0)}$ . Generate  $m = 1000$  independent random numbers  $\{u_j\}_{j=1}^m$ , which is uniformly distributed over  $[p_{\min}, 1)$ , where  $p_{\min} = \min\{p_i\}$  with  $p_i = G(\Delta_i|\mathbf{X}_i^\top \beta^{(k)}, \eta^{(k)})$ . Put

$$\begin{aligned}
\tilde{L}_Y^k(\beta, \eta) & := \frac{1}{n} \sum_{i=1}^n G(\Delta_i|\mathbf{X}_i^\top \beta, \eta) \log \{g(Y_i|\mathbf{X}_i^\top \beta, \eta)\} I(Y_i \in (0, \Delta_i]) \\
& + \frac{1}{n} \sum_{i=1}^n (\log(\eta) - \eta \log(\mathbf{X}_i^\top \beta)) (1 - G(\Delta_i|\mathbf{X}_i^\top \beta, \eta)) \\
& - \frac{1}{n} \sum_{i=1}^n \left\{ \left[ \left( \frac{\Delta_i}{\mathbf{X}_i^\top \beta} \right)^\eta + 1 \right] \exp \left( - \left( \frac{\Delta_i}{\mathbf{X}_i^\top \beta} \right)^\eta \right) \right\} \\
& + (\eta - 1) \frac{1}{n} \sum_{i=1}^n \left\{ (1 - p_i) \frac{1}{m_i} \sum_{u_j \geq p_i} \log \left( G^{-1}(u_j|\mathbf{X}_i^\top \beta^{(k)}, \eta^{(k)}) \right) \right\},
\end{aligned}$$

where  $\mathbf{X}_i^\top = (1, X_i, X_i^2, X_i^3)$ ,  $G^{-1}(\cdot)$  denotes the inverse function of  $G(\cdot)$ , and  $m_i$  is the cardinal number of the set  $\{j : u_j \geq p_i, j = 1, 2, \dots, m\}$ .

**M-step.** Maximize  $\tilde{L}_Y^k(\beta, \eta)$  with respect to  $(\beta, \eta)$  and obtain the new estimators, namely,  $\hat{\beta}^{(k+1)}, \hat{\eta}^{(k+1)}$ . Check whether or not

$$\sum_{i=0}^3 |\hat{\beta}_i^{(k+1)} - \hat{\beta}_i^{(k)}| + |\hat{\eta}^{(k+1)} - \hat{\eta}^{(k)}| < \frac{1}{n^\delta}, \quad \text{for some } \delta > 0.5.$$

If it is true, return  $\hat{\beta} = (\hat{\beta}_0^{(k+1)}, \hat{\beta}_1^{(k+1)}, \hat{\beta}_2^{(k+1)}, \hat{\beta}_3^{(k+1)})^\top$  and  $\hat{\eta} = \hat{\eta}^{(k+1)}$ ; otherwise, repeat the **E-step** and **M-step** until convergence achieves.

We have coded this algorithm by **R** program relying on the optimization function *constrOptim()*. The implementation runs quite fast. Usually, convergence can achieve by several iterations.

## Results

Based on the information of 2172 confirmed cases, we compute the estimated parameters by using the implementation of EM algorithm mentioned above; see Table 1. It is worth mentioning that  $\hat{\beta}_3 = -1.1 \times 10^{-6}$  is very small, which implies that the cubic form  $\lambda_3(x) = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3$  is flexible enough to characterize the trend of the conditional mean  $E(T|X)$  on  $X$ . We do not need to assume a higher order polynomials for  $\lambda(x)$ .

Table 1: The results of parameter estimation.

| Density                  | Parameter        | Estimate  |
|--------------------------|------------------|---|
| $\phi(x; \mu, \sigma^2)$ | $\hat{\mu}$      | 43.15   |
|                          | $\hat{\sigma}^2$ | 230.49  |
| $g(y \lambda(x), \eta)$  | $\hat{\eta}$     | 1.81  |
|                          | $\hat{\beta}$    | (10.24, -0.13, $2.0 \times 10^{-3}$ , $-1.1 \times 10^{-6}$ ) |

Using results in Table 1, we obtain the conditional 0.05, 0.25, 0.5, 0.75, 0.9 and 0.95-th quantiles of the incubation period distribution of COVID-19 on ages; see Figure 4 for details.

Figure 4 indicates that quantiles corresponding to people of the middle ages seem to be less than those of the others. Especially, the estimated 0.95-th conditional quantile of the children and the elderly is greater obviously than that of the middle-aged. To be more detailed, we specify the 0.95-th conditional quantiles of the incubation period distribution of COVID-19 on different ages in Table 2. Table 2 indicates that the 0.95-th conditional quantiles of people in the age group 23 ~ 55 lie between 14 and 15 days, shorter than those of the other groups. We also list the numbers of cases in each groups and the corresponding proportions. It turns out the infected cases of ages 23 ~ 55 account for more than 70% of the total collected cases. Further, note that we collected 136 cases whose incubation periods are greater than 14 days. We provide the distribution of these cases over all age groups. It is shown that the proportion of the age group 23 ~ 55 is the smallest. Its value is only 5.11%. On the other hand, the proportions of other age groups are much higher than 5%.

Furthermore, in order to further verify the results above, we divide the observed incubation periods into three groups by age: 0 to 25 years old, 26 to 60 years old and over 60 years old, according to *World Population Prospects: the 2019 Revision*<sup>[1]</sup>. Then we fit the Weibull distribution in each group by setting  $\lambda(x) = \beta_0$ .

Figure 5 shows that people under age of 0 to 25 years old or over 60 years old has higher probability that would emerge longer incubation period than people under age of 26 to 60 years old. Moreover, the right figure is the fitted Weibull distribution function of incubation period of COVID-19 in three age groups. It is obvious that the 0.95 quantile of people under 26 and over 60 is greater than that of people aged from 26 to 60. This roughly coincides with the results reported in Figure 4, and hence indicates that the conditional distribution considered above is able to characterize the relationship of age and real incubation period of COVID-19.

[1]<https://population.un.org/wpp/>

Table 2: Summary on the age groups divided according to the estimated 0.95-th quantile of the incubation period distribution of COVID-19, where  $\tilde{n}_j$  denotes the number of infected cases in each age group, and  $\tilde{m}_j$  the number of those cases with incubation period over 14 days in each age group.

| Item                               | Incubation period interval |                 |                 |                 |                 |          |          |         |
|------------------------------------|----------------------------|-----------------|-----------------|-----------------|-----------------|----------|----------|---------|
|                                    | 14-15                      | 15-16           | 16-17           | 17-18           | 18-19           | 19-20    | 20-21    | 21-22   |
| Age group(s)                       | 23~55                      | 15~22<br>56~63  | 9~14<br>64~69   | 4~8<br>70~74    | 0~3<br>75~79    | 80~83    | 84~86    | ≥ 88    |
| Number $\tilde{n}_j$ in each group | 1545                       | 93<br>232       | 24<br>156       | 24<br>45        | 18<br>22        | 6        | 6        | 1       |
| (total number $n_i$ )              | (1545)                     | (325)           | (180)           | (69)            | (40)            | (6)      | (6)      | (1)     |
| Proportion $\tilde{n}_j/n$         | 71.13%                     | 4.28%<br>10.68% | 1.1%<br>7.18%   | 1.1%<br>2.07%   | 0.83%<br>1.01%  | 0.28%    | 0.28%    | 0.05%   |
| (total proportion $n_i/n$ )        | (71.13%)                   | (14.96%)        | (8.28%)         | (3.17%)         | (1.84%)         | (0.28%)  | (0.28%)  | (0.05%) |
| Number $\tilde{m}_j$ in each group | 79                         | 3<br>19         | 3<br>16         | 1<br>8          | 3<br>2          | 1        | 1        | 0       |
| (total number $m_i$ )              | (79)                       | (22)            | (19)            | (9)             | (5)             | (1)      | (1)      | (0)     |
| Proportion $\tilde{m}_j/n_i$       | 5.11%                      | 3.23%<br>8.19%  | 12.5%<br>10.26% | 4.17%<br>17.78% | 16.67%<br>9.09% | 16.67%   | 16.67%   | 0%      |
| (total proportion $m_i/n_i$ )      | (5.11%)                    | (11.42%)        | (22.76%)        | (21.95%)        | (25.76%)        | (16.67%) | (16.67%) | (0%)    |

## Discussion

Our estimation of the conditional quantiles indicates that the incubation period of COVID-19 varies depending on the age of the infected cases. Precisely, the incubation period of the young and the old tends to be longer than that of the middle-aged people.

It seems that we can find some supports from the immune theory in medical science. Note that the human immunity refers to the sensitivity of the immune system in response to infection. During the incubation period, since the host’s immune system has not yet been activated, and the body has not begun to show symptoms, the virus can use this period to make a lot of replication. In many situations, to the infection, the more responsive the host’s immune system is, the shorter the incubation period tends to be. By further noting that the human immunity is weak at the beginning, improves with age, and will decline in the old period [22, 23], it hence may be not surprised to see that the incubation period of the young and the old cases are longer than that of the middle-aged cases.

Currently, the quarantine duration is fixed to 14 days. It does not take into account of any other facts, e.g., age. Hence, our results may be helpful for disease control and prevention efforts, because it enables us to take some more precise measures. For example, personalized quarantine duration can be taken for individuals of different ages. Especially, people of ages between 23~55 play important roles in real life and are a significant part of the labor force. Besides, they account for the largest proportion of the population. A relative short quarantine duration for them not only can reduce the burden of the medical staff but also is conducive to social economic development. In contrary, the conditional quantiles on ages 0~22 and over 55 are much greater than 14 days. We may need to extend the quarantine duration for people of these ages. Such extension may help the prevention but have limit impacts on the social economic development.

It is worth mentioning there exist some other ways in statistics to characterize the conditional quantile of the incubation period over the age. That is, first model

the relationship between  $T$  and  $X$  by the following linear model:

$$T = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon,$$

and then use the technique of quantile regression to estimate the unknown parameters  $\beta_i, i = 0, 1, 2, 3$ . However, note that the true incubation period  $T$  can not be fully observed, and the observed incubation period  $Y$  is randomly smaller than  $T$ . The estimated quantiles may suffer from some problem, e.g., underestimation.

In fact, we also report in Figure 6 the result of the ordinary quantile regression mentioned above. Figure 6 shows that the regression quantiles follow a similar to the conditional quantiles reported in Figure 4. Nevertheless, we also note that the 0.25-th quantile and the 0.05-th quantile intersect with each other when the age is greater than 80. It seems difficult to have a reasonable explain about this phenomenon. This strange result may be caused by the biased sampling issue. Hence, we did not take the regression quantiles to analyze the current data, although it provides some similar results as the conditional quantiles.

## Conclusion

Based on the collected data, our model showed that the incubation period of COVID-19 varies depending on the age of the infected cases. Specifically, the incubation period of the young and the old tends to be longer than that of the middle-aged people. These findings enable us to take some more precise measures rather than fixed ones, and thus may be helpful for disease control and prevention efforts. For example, personalized quarantine duration, namely shorter for middle-aged people and longer for the young and the old, can be taken for individuals of different ages, instead of fixed 14 days. People aged 23~55 are a major part of the labor force and account for the largest proportion of the population, therefore, such methods may help the prevention but have limit impacts on the social economic development.

## List of abbreviations

WHO: World Health Organization  
COVID-19: Coronavirus disease 2019  
2019-nCoV: 2019-novel coronavirus  
SARS-CoV: Severe acute respiratory syndrome coronavirus  
SARS: Severe acute respiratory syndrome  
MERS: Middle East respiratory syndrome

### Ethics approval and consent to participate

This study was based on publicly available data and did not require ethical approval.

### Consent for publication

Not applicable.

### Availability of data and materials

The data used during the current study are collected from the websites of the centres of disease control, or the daily public reports on COVID-19 in 29 provinces outside Hubei province through February 16th, 2020, and can be checked according to the related links of the websites. See the dataset in additional file 1.

### Competing interests

The authors declare that they have no competing interests.

### Author's contributions

Xiaohui Liu proposed the main idea, analysed the data and wrote the first draft. Lei Wang, Xiansi Ma and Jiewen Wang contributed to the analysis. All authors interpreted the results and co-wrote the manuscript.

### Acknowledgements

Xiaohui Liu's research was supported by NSF of China (Grant No.11601197, 11461029), China Postdoctoral Science Foundation funded project (2016M600511, 2017T100475), NSF of Jiangxi Province (No.20171ACB21030).

### Author details

<sup>1</sup>School of Statistics, Jiangxi University of Finance and Economics, No.169, East Shuanggang Road, 330013 Nanchang, Jiangxi Province, China. <sup>2</sup>Research Center of Applied Statistics, Jiangxi University of Finance and Economics, No.169, East Shuanggang Road, 330013 Nanchang, Jiangxi Province, China. <sup>3</sup>National Engineering Laboratory for Educational Big Data, Central China Normal University, NO.152 Luoyu Road, 430079 Wuhan, Hubei Province, China.

### References

- World Health Organization (WHO). Novel Coronavirus (2019-nCoV) Situation Report 22. [https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200211-sitrep-22-ncov.pdf?sfvrsn=fb6d49b1\\_2](https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200211-sitrep-22-ncov.pdf?sfvrsn=fb6d49b1_2)
- Riou, J., Althaus, C.L.: Pattern of early human-to-human transmission of wuhan 2019 novel coronavirus (2019-ncov), december 2019 to january 2020. *Eurosurveillance* **25**(4), 2000058 (2020)
- Nishiura, H., Linton, N.M., Akhmetzhanov, A.R.: Serial interval of novel coronavirus (covid-19) infections. *International Journal of Infectious Diseases* (2020)
- Chen, T.-M., Rui, J., Wang, Q.-P., Zhao, Z.-Y., Cui, J.-A., Yin, L.: A mathematical model for simulating the phase-based transmissibility of a novel coronavirus. *Infectious Diseases of Poverty* **9**(1), 1–8 (2020)
- Liu, Y., Gayle, A.A., Wildersmith, A., Rocklöv, J.: The reproductive number of covid-19 is higher compared to sars coronavirus. *Journal of Travel Medicine* **27**(2) (2020)
- Li, P., Fu, J., Li, K., Chen, Y., Wang, H., Liu, L., Liu, J., Zhang, Y., Liu, S., Tang, A., et al.: Transmission of covid-19 in the terminal stage of incubation period: a familial cluster. *International Journal of Infectious Diseases* (2020)
- Lai, C.-C., Shih, T.-P., Ko, W.-C., Tang, H.-J., Hsueh, P.-R.: Severe acute respiratory syndrome coronavirus 2 (sars-cov-2) and corona virus disease-2019 (covid-19): the epidemic and the challenges. *International journal of antimicrobial agents*, 105924 (2020)
- Wu, J.T., Leung, K., Leung, G.M.: Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study. *The Lancet* **395**(10225), 689–697 (2020)
- Wilkening, D.A.: Modeling the incubation period of inhalational anthrax. *Medical Decision Making* **28**(4), 593–605 (2008)
- Sartwell, P.E., et al.: The distribution of incubation periods of infectious diseases. *American Journal of Hygiene* **51**(3), 310–18 (1950)
- Armenian, H.K., Lilienfeld, A.M.: The distribution of incubation periods of neoplastic diseases. *American Journal of Epidemiology* **99**(2), 92–100 (1974)
- Chan, M., Johansson, M.A.: The incubation periods of dengue viruses. *PLOS ONE* **7**(11) (2012)
- Goulet, V., King, L.A., Vaillant, V., De Valk, H.: What is the incubation period for listeriosis. *BMC Infectious Diseases* **13**(1), 11–11 (2013)
- Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S., Lau, E.H., Wong, J.Y., et al.: Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine* (2020)
- Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D.S.C., et al.: Clinical characteristics of 2019 novel coronavirus infection in china. *medRxiv* (2020)
- Backer, J.A., Klinkenberg, D., Wallinga, J.: Incubation period of 2019 novel coronavirus (2019-ncov) infections among travellers from wuhan, china, 20–28 january 2020. *Eurosurveillance* **25**(5) (2020)
- Linton, N.M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A.R., Jung, S., Yuan, B., Kinoshita, R., Nishiura, H.: Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine* **9**(2), 538 (2020)
- Mariotto, A.B., Mariotti, S., Pezzotti, P., Rezza, G., Verdecchia, A.: Estimation of the acquired immunodeficiency syndrome incubation period in intravenous drug users: A comparison with male homosexuals. *American Journal of Epidemiology* **135**(4), 428–437 (1992)
- Medley, G.F., Anderson, R.M., Cox, D.R., Billard, L.: Incubation period of aids in patients infected via blood transfusion. *Nature* **328**(6132), 719–721 (1987)
- Medley, G.F., Billard, L., Cox, D.R., Anderson, R.M.: The distribution of the incubation period for the acquired immunodeficiency syndrome (aids). *Proceedings of The Royal Society B: Biological Sciences* **233**(1272), 367–377 (1988)
- Munoz, A., Xu, J.: Models for the incubation of aids and variations according to age and period. *Statistics in Medicine* **15**(22), 2459–2473 (1996)
- Qin, L., Jing, X., Qiu, Z., Cao, W., Jiao, Y., Routy, J.-P., Li, T.: Aging of immune system: immune signature from peripheral blood lymphocyte subsets in 1068 healthy adults. *Aging (Albany NY)* **8**(5), 848 (2016)
- Cao, X.: *Medical Immunology*. Beijing: People's Medical Publishing House, Beijing, China (2013)

Figure 1: **Scatter plot of the incubation period of COVID-19 versus the ages of confirmed cases.** Note that many data points are overlapped.

Figure 2: **Histogram of the ages of confirmed cases.** The line stands for the fitted density function of normal distribution.

Figure 3: **Scatter plot of the number of infected cases versus the infected time.** The line stands for the fitted line through the least squares method.

Figure 4: **The estimated conditional quantiles of the incubation period distribution of COVID-19.** 15.05 is the estimated 0.95-th quantile of the unconditional incubation periods.

Figure 5: **The fitted density and distribution functions of the incubation period for three age groups.** The age groups are 0-25, 26-60 and over 60. The left indicates fitted density functions. The right indicates fitted distribution functions.

Figure 6: **The estimated regression quantiles.** The quantiles are 0.05, 0.25, 0.5, 0.75, 0.9, 0.95 from the bottom to top.

## Figures

### Additional Files

#### Additional file 1 — Data

The data was included in the file "Additional file 1.xls".

##### 1. Variables description

Age: age of the confirmed case

InDate: infection date

OutDate: onset date

InPeriod: incubation Period

##### 2. Values of variables

Age: the original age of the confirmed case

InDate and OutDate: for convenience, the values of InDate and OutDate were represented as the number of days departing from December 31, 2019, on which WHO announced the disease COVID-19 have become a public health incident. For example, the date January 1, 2020 takes the value 1.

Incubation Period: Onset date - Infection date + 1

#### Additional file 2 — Code

The code was included in the file "Additional file 2.R" (use software RStudio or R to open it).

##### 1. Data file path

To ensure that the program runs successfully, the data file "Additional file 1.xls" should be placed in the current working directory of R, which can be shown by running the code "getwd()" in R.

An alternative way is adding the file path in Line 18 of the code.

##### 2. R packages needed

R packages including "readxl", "splines", "ggplot2", "quantreg" and "SparseM" should be successfully installed before running the program. All the packages can be installed by running "install.packages()", such as "install.packages("readxl")" for package "readxl".

# Figures

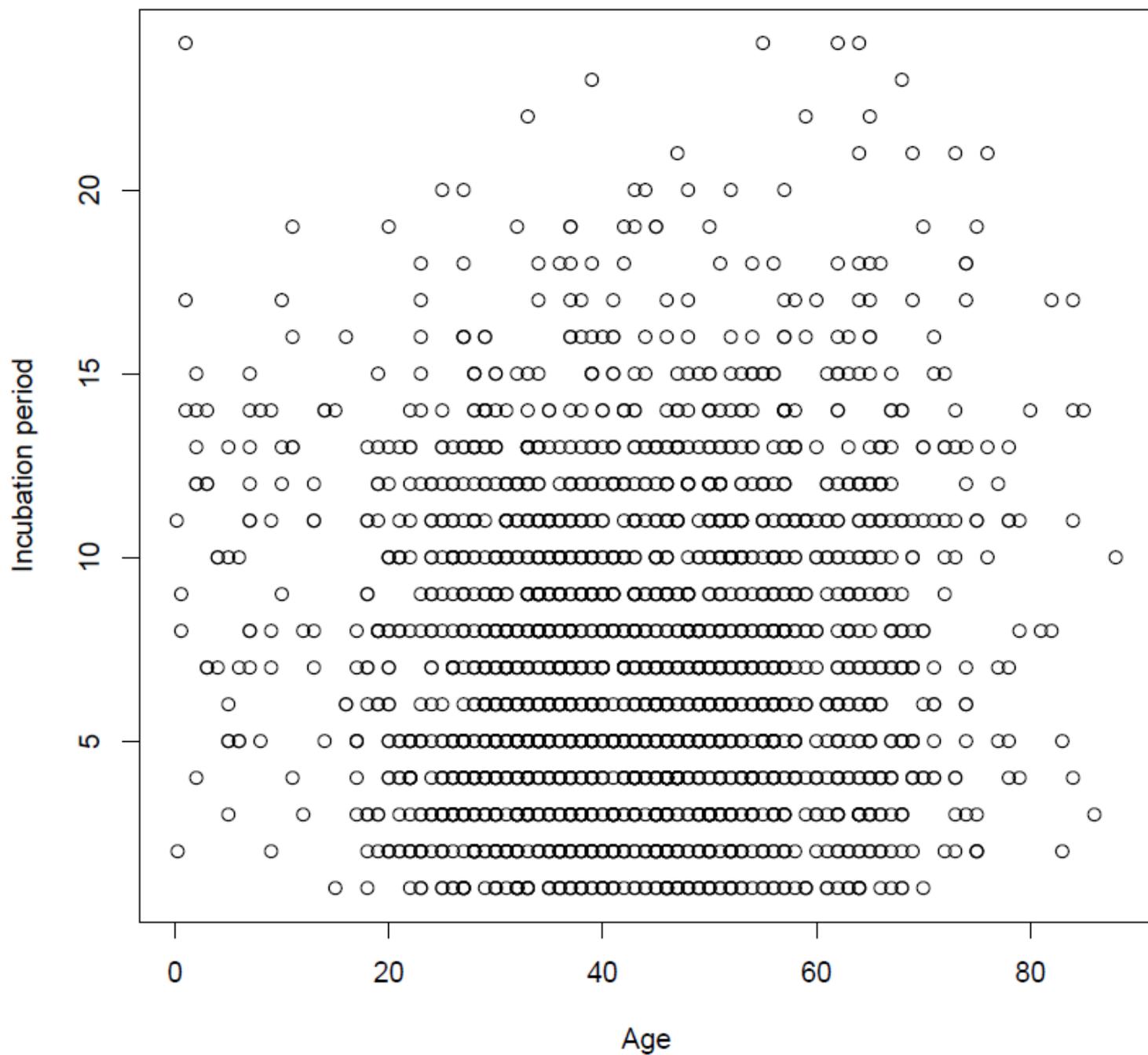
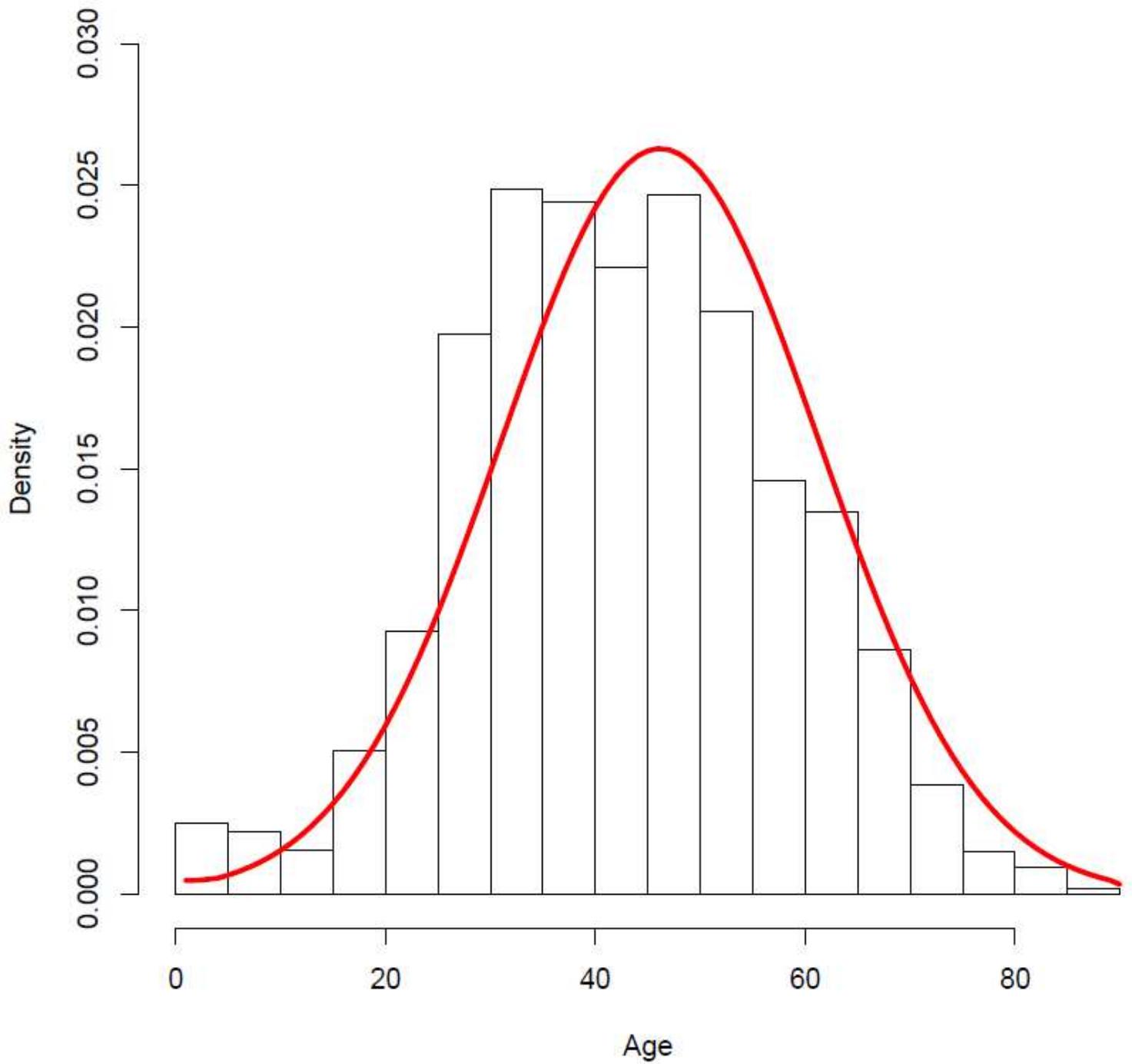


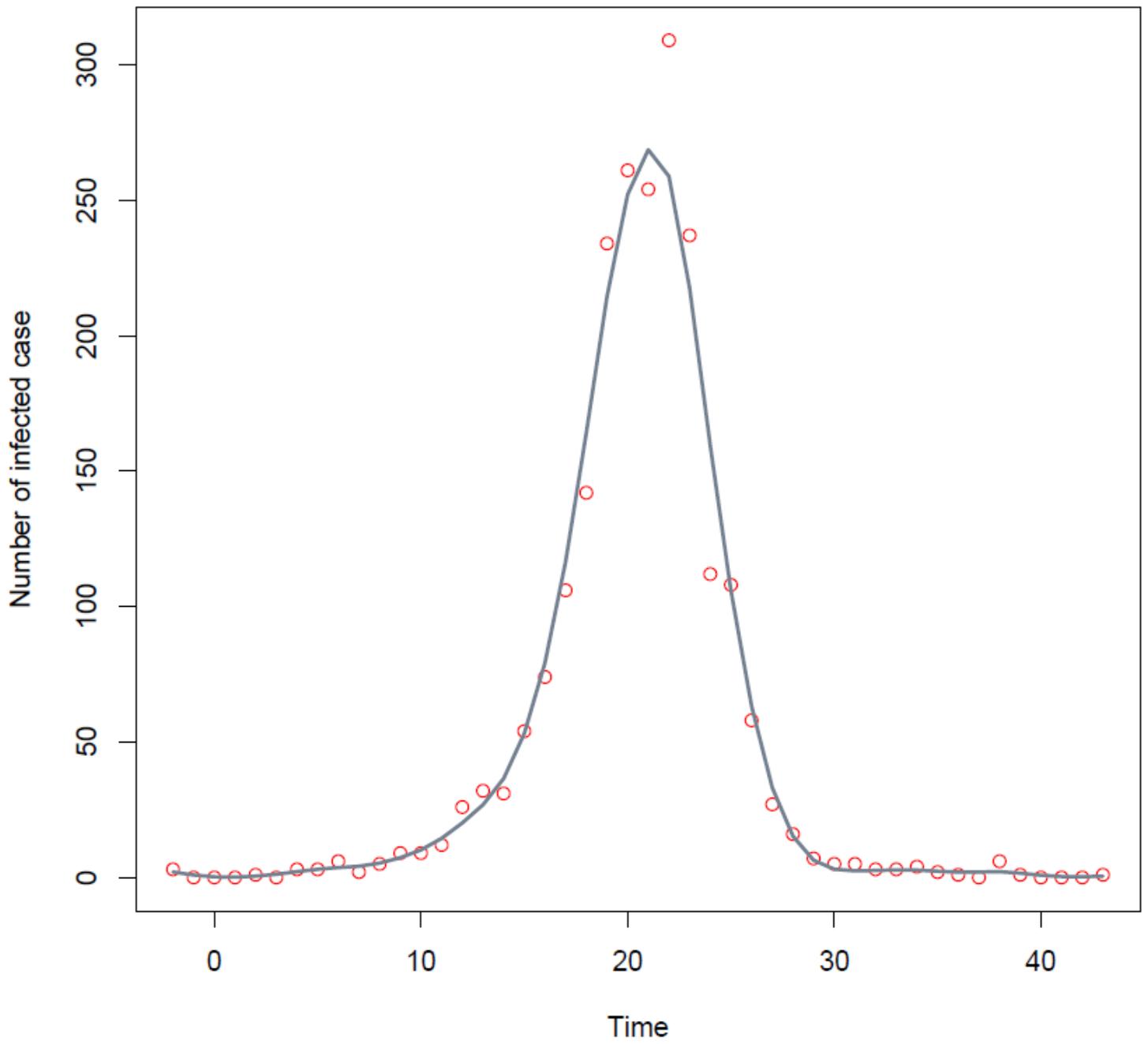
Figure 1

Scatter plot of the incubation period of COVID-19 versus the ages of confirmed cases. Note that many data points are overlapped.



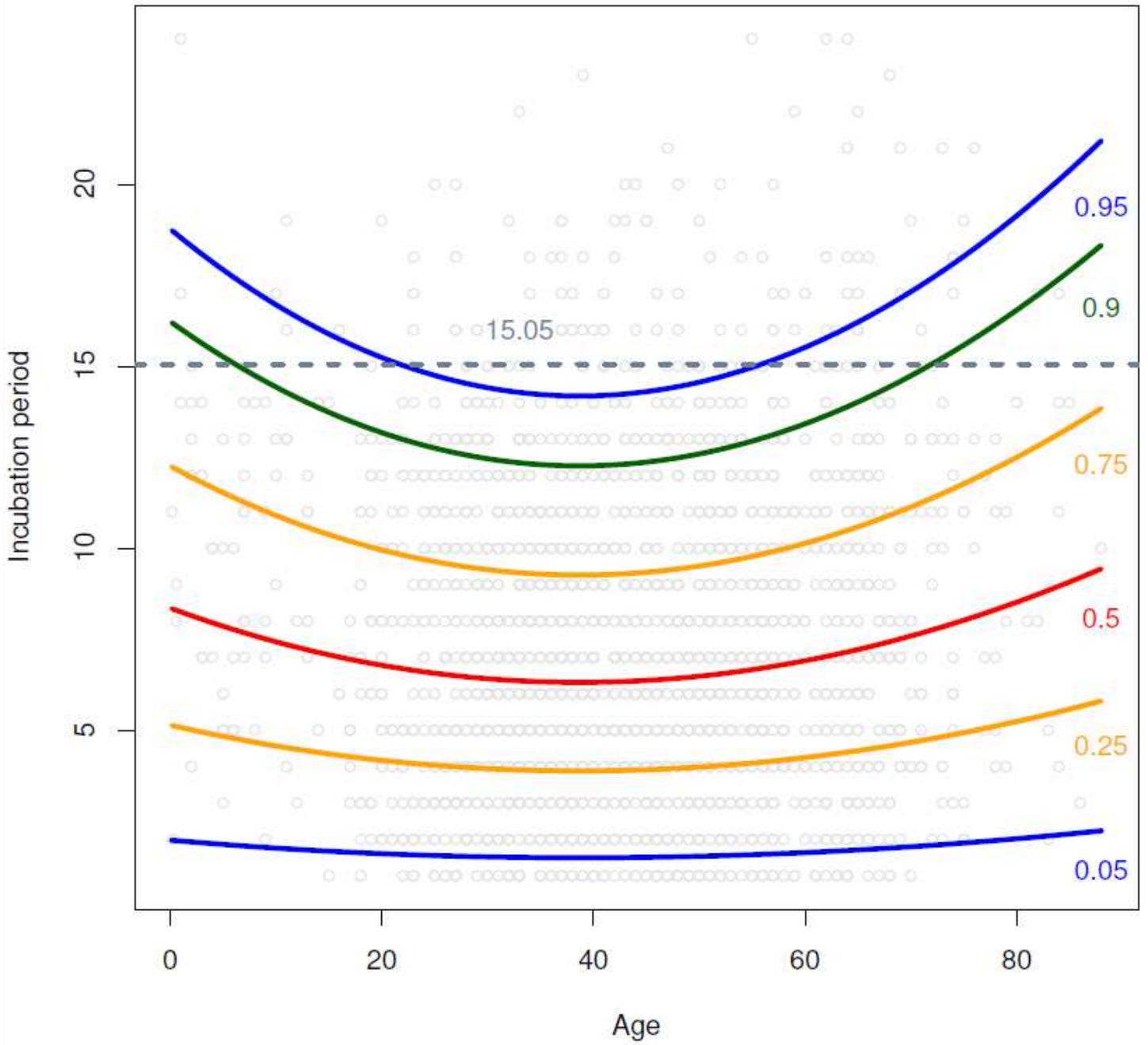
**Figure 2**

Histogram of the ages of confirmed cases. The line stands for the fitted density function of normal distribution.



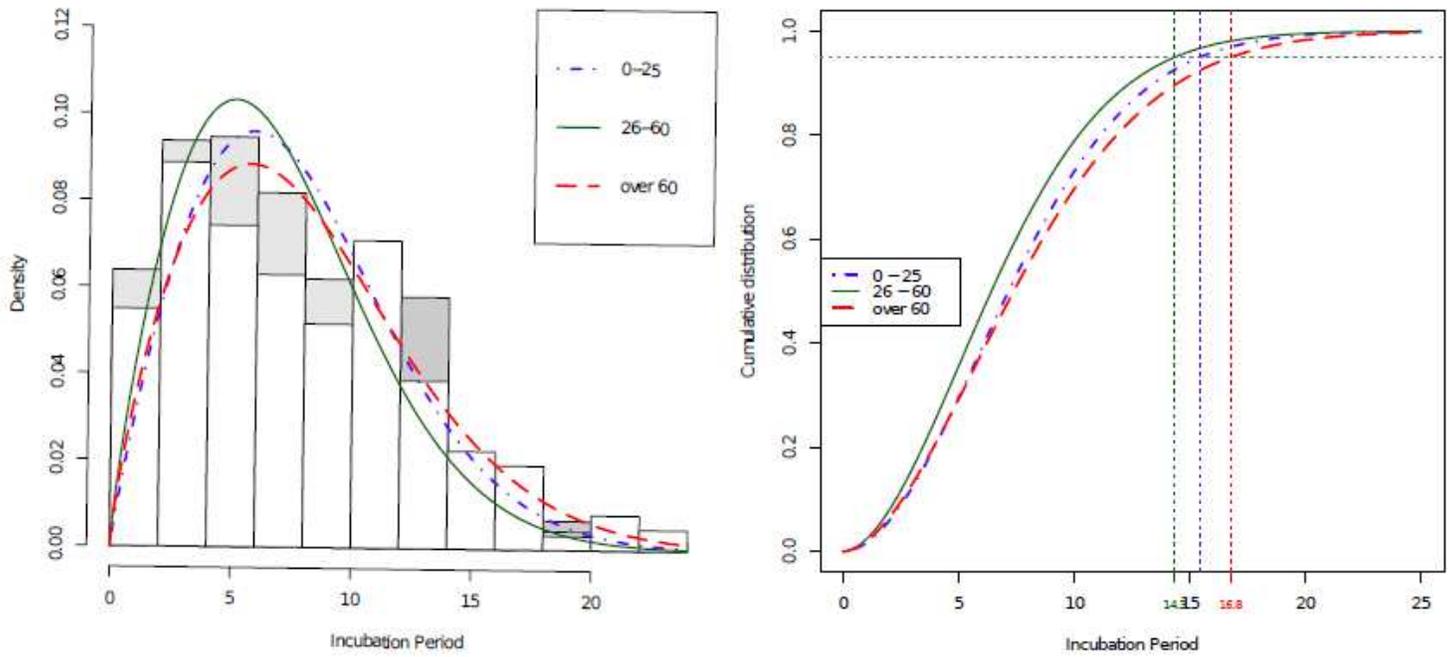
**Figure 3**

Scatter plot of the number of infected cases versus the infected time. The line stands for the fitted line through the least squares method.



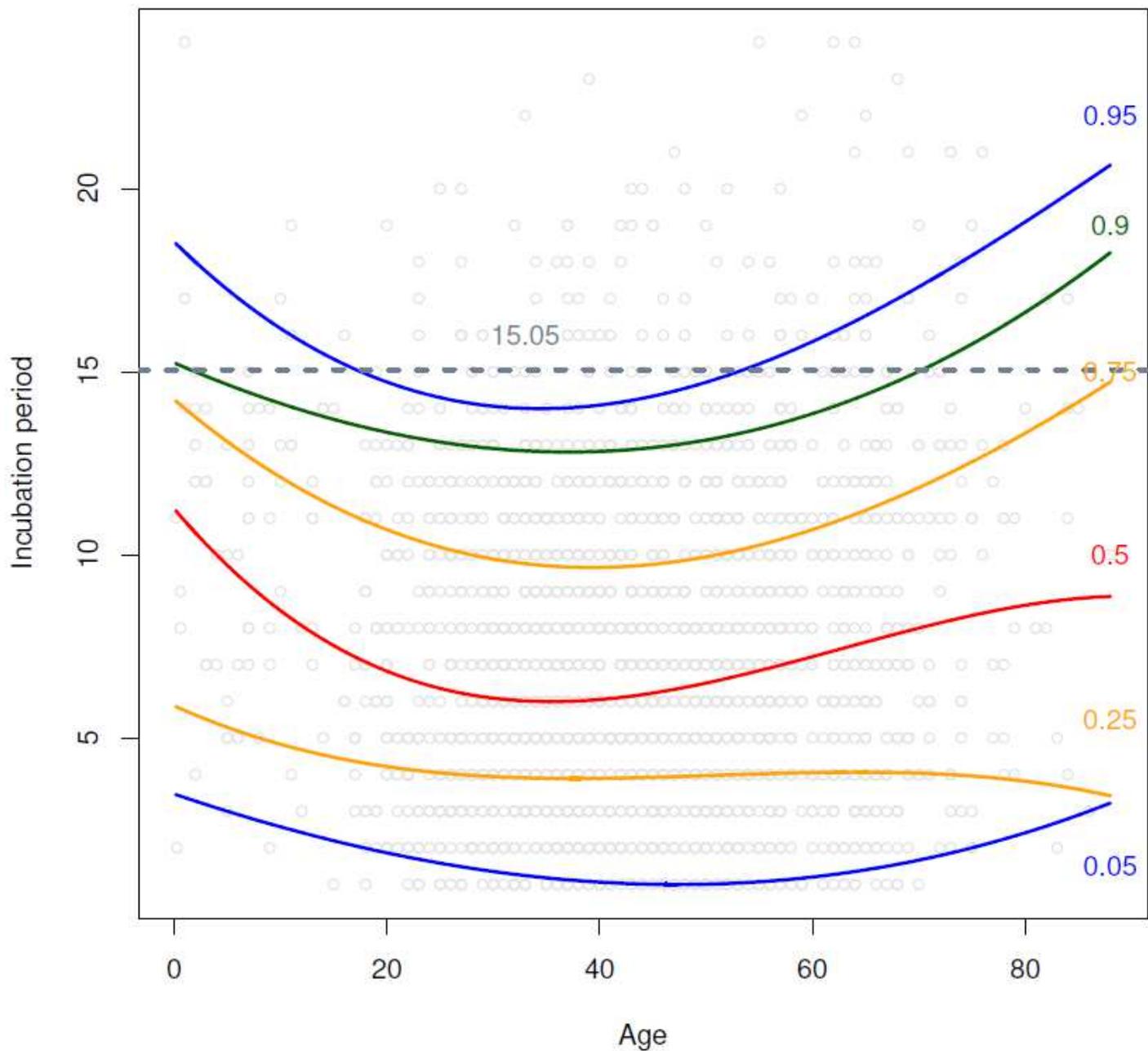
**Figure 4**

The estimated conditional quantiles of the incubation period distribution of COVID-19. 15.05 is the estimated 0.95-th quantile of the unconditional incubation periods.



**Figure 5**

The fitted density and distribution functions of the incubation period for three age groups. The age groups are 0-25, 26-60 and over 60. The left indicates fitted density functions. The right indicates fitted distribution functions.



**Figure 6**

The estimated regression quantiles. The quantiles are 0.05, 0.25, 0.5, 0.75, 0.9, 0.95 from the bottom to top.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1.xlsx](#)
- [Additionalfile2.R](#)