Journal of Data Science 19 (2), 253–268
April 2021

DOI: 10.6339/21-JDS992
Statistical Data Science

Assessment of Effects of Age and Gender on the Incubation Period of COVID-19 with a Mixture Regression Model

Siming Zheng[1], Jing Qin[2], and Yong Zhou[3,4,*]

[1]*Academy of Mathematics and Systems Science, University of Chinese Academy of Sciences, Beijing, China*

[2]*National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, U.S.A.*

[3]*Key Laboratory of Advanced Theory and Application in Statistics and Data Science, MOE*

[4]*Academy of Statistics and Interdisciplinary Sciences, Faculty of Economics and Management, East China Normal University, Shanghai, China*

Abstract

Following the outbreak of COVID-19, various containment measures have been taken, including the use of quarantine. At present, the quarantine period is the same for everyone, since it is implicitly assumed that the incubation period distribution of COVID-19 is the same regardless of age or gender. For testing the effects of age and gender on the incubation period of COVID-19, a novel two-component mixture regression model is proposed. An expectation-maximization (EM) algorithm is adopted to obtain estimates of the parameters of interest, and the simulation results show that the proposed method outperforms the simple regression method and has robustness. The proposed method is applied to a Zhejiang COVID-19 dataset, and it is found that age and gender statistically have no effect on the incubation period of COVID-19, which indicates that the quarantine measure currently in operation is reasonable.

Keywords *EM algorithm; incubation period; length-biased data; mixture model*

1 Introduction

On March 11, 2020, the World Health Organization (WHO) declared a pandemic of COVID-19, which is also called SARS-CoV-2. The outbreak of COVID-19 poses a serious challenge to global public health and economy. By April 11, 2020, the pandemic had caused more than $1,700,000$ confirmed cases of infection and over $100,000$ fatalities. To cope with this crisis, several containment measures, including isolation of infected individuals, travel restrictions, quarantine, etc, have been implemented in many countries to suppress virus transmission via human-to-human contact. It is worth mentioning that in addition to taking these measures, China has also established an efficient close-contact tracing system. Through these efforts, the outbreak of COVID-19 in China is now well under control, which shows that such measures can effectively block the transmission chain of COVID-19. One effective way of finding potentially infected individuals is to keep those who may have been exposed to infectious pathogens in quarantine for some time. As is well known, one of the key factors determining the optimal quarantine time for suspected cases is a good understanding of the incubation period.

To date, there has been some excellent work on the incubation period of COVID-19. Based on the first 425 laboratory-confirmed cases reported on January 22, 2020 in China, but with

*Corresponding author. Email: yzhou@amss.ac.cn..

Received April 1, 2020; Accepted May 1, 2020

only 10 of them having exactly recalled dates of getting infected, Li et al. (2020) fitted a log-normal distribution and found a mean incubation period of 5.2 days. Similarly, by analyzing 291 patients who recalled their dates of exposure to infectious pathogens, Guan et al. (2020) found a median incubation period of 4.0 days. However, these studies involved individuals' recall bias. To fix this problem, Lauer et al. (2020) collected time data of four events, including possible exposure to COVID-19 and symptom onset. For example, the exact exposure date was obtained if available; otherwise, upper and lower bounds were obtained to form a possible interval of exposure. A parametric accelerated failure time model was adopted and gave an estimate of the median incubation period of COVID-19 that was also 5.2 days. Analogously, Backer et al. (2020) estimated the distribution of the incubation period using the censored intervals for the incubation periods of some confirmed cases, with these intervals having been obtained from the relevant dates of travel history and symptom onset. However, both of these works suffered from two sampling bias problems. One is that the short follow-up time meant that shorter incubation periods would be observed more frequently. The other was that as the observations were of time lags between two specific timings (e.g., between the date of departure from an epidemic focus and the date of symptom onset), patients with longer incubation periods were more easily observed. Linton et al. (2020) adopted a similar approach to Backer et al. (2020), but corrected the shorter incubation period bias. To handle the longer incubation period bias, Qin et al. (2020) used renewal process theory and proposed a length-biased Weibull distribution to fit the specific time lag data of 1211 confirmed cases leaving Wuhan between January 19 and 23, 2020. They estimated the median of the incubation period to be 8.13 days, much longer than the results mentioned above.

The first motivation for proposing a mixture regression model is that although the assumptions in Qin et al. (2020) are quite reasonable, some cases may become infected on the day of departure, and thus their observed time lags between departure from Wuhan and symptom onset are also complete incubation periods, which are the time lags between infection and symptom onset. This was also noted by Qin et al. (2020), but they only considered it in the context of sensitivity analysis, whereas our focus here is clearly different from theirs. Therefore, a mixture model would be more appropriate, since it provides a flexible tool to model data arising from a heterogeneous population. Traditional mixture models involve no regression, and much excellent work has been done on these models. For example, based on the Hessian of the multivariate normal mixture model, Boldea and Magnus (2009) gave estimates of all parameters that appeared to be superior to previous estimates. Later, Qin and Priebe (2013) obtained robust estimates by maximizing a novel L_q likelihood through the expectation-maximization (EM) algorithm. Chen (2017) also presented detailed and correct consistency results from a maximum likelihood estimator (MLE) with traditional mixture models and streamlined some previously obtained results. More detailed and significant reviews can be found in the book by McLachlan and Peel (2000).

As for the second motivation for our proposed model, we note that, except for the mixture problem, all of the incubation period studies mentioned above considered only the incubation period in the whole population. However, for physiological reasons, the incubation period distributions of infected individuals may differ depending on age and gender. These effects may have a negative influence on current quarantine measures. Consequently, it is meaningful to assess the effect of age and gender on the incubation period of COVID-19 through a regression model. The mixture model approaches mentioned above are clearly not appropriate for a regression problem. For the regression model, Jiang and Tanner (1999) considered a hierarchical mixtures-of-experts model in which exponential family regression was mixed, and they obtained their estimates by a maximum likelihood method. Khalili and Chen (2007) defined a family of parametric conditional

density functions and created a finite mixture of regression models (FRM). Through the use of a weighted penalized log-likelihood function, they implemented variable selection procedure for the FRM. However, these two models are not appropriate for our regression problem.

In this paper, we propose a novel two-component mixture regression model (1). Parameter estimates are obtained through the maximum likelihood method and the EM algorithm is adopted. For a more detailed literature review of the EM algorithm, readers can refer to the book by Liang et al. (2010). As is well known, one drawback of the EM algorithm is the dependence of its solution on the initial values that are used, and this is often a consequence of a local maxima problem with the objective functions. Dimitris and Evdokia (2003) studied the effect of initial values on the EM algorithm for a finite normal mixture model with each normal component having common variance and for a finite Poisson mixture model. They compared several methods for choosing initial values for the EM algorithm in these models, including, for example, a random starting point and starting at some moment estimates. Although their new initial value estimator produced better results, it cannot be easily extended to our setting. In this paper, owing to its simplicity, we adopt the classic random initialization method to partially handle the initial value dependence of the EM algorithm, and we conduct a sensitivity analysis for our settings with the aim of showing that this initialization method is a reasonable one.

The rest of the paper is organized as follows. In Section 2, a novel mixture regression model (1) is proposed and the reason for formulating such a model is discussed. The estimator of the parameters of interest is the MLE based on the conditional likelihood of observed data, and the estimates are calculated using the EM algorithm. In Section 3, several simulation studies are conducted to test the performance of the proposed method. Furthermore, under various possible fixed ranges of uninteresting parameters, a sensitivity analysis is implemented. The results show that the estimates of the parameters of interest are robust to these various settings. Some simulations are also conducted to test whether likelihood ratio tests (l.r.t.'s) work for our proposed model. The results reveal that the application of l.r.t.'s is appropriate. In Section 4, the proposed model and method are applied to a Zhejiang COVID-19 dataset. The sensitivity analysis shows the estimates of the parameters of interest and maximum likelihood are robust and makes the results more reliable. Finally, based on l.r.t.'s for the regression coefficients of age and gender in the model (1), we find that it is statistically not rejected that age and gender have no effect on the incubation period of COVID-19.

2 Model and Estimation Method

In this section, we propose a mixture model for the data analysis. Let V denote the time lag of a confirmed case between departure from Wuhan and onset of symptoms. His/her covariate column vector $X \in \mathcal{R}^p$ contains different risk factors, such as age and gender. We define the conditional density of V given $X = \boldsymbol{x}$ as $h(v|\boldsymbol{x}, \lambda, \alpha, \boldsymbol{\theta}, \boldsymbol{\beta})$:

$$h(v|\boldsymbol{x}, \lambda, \alpha, \boldsymbol{\theta}, \boldsymbol{\beta}) = \pi(\boldsymbol{x}, \boldsymbol{\theta}) f_{\lambda,\alpha}(v \exp(\boldsymbol{x}'\boldsymbol{\beta})) \exp(\boldsymbol{x}'\boldsymbol{\beta}) + (1 - \pi(\boldsymbol{x}, \boldsymbol{\theta})) g_{\lambda,\alpha}(v \exp(\boldsymbol{x}'\boldsymbol{\beta})) \exp(\boldsymbol{x}'\boldsymbol{\beta}), \quad (1)$$

where

$$g_{\lambda,\alpha}(v) = \frac{\bar{F}_{\lambda,\alpha}(v)}{\int \bar{F}_{\lambda,\alpha}(t)\, dt}, \quad \bar{F}_{\lambda,\alpha}(t) = 1 - F_{\lambda,\alpha}(t),$$

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\exp(\theta_0 + \boldsymbol{x}'\boldsymbol{\theta}_1)}{1 + \exp(\theta_0 + \boldsymbol{x}'\boldsymbol{\theta}_1)}, \quad \boldsymbol{\theta} = (\theta_0, \boldsymbol{\theta}_1')'.$$

Here, $F_{\lambda,\alpha}(\cdot)$ is a cumulative distribution function (c.d.f.) and $f_{\lambda,\alpha}(\cdot)$ is the corresponding probability density function (p.d.f.), which is assumed to be the Weibull distribution density, i.e., $f_{\lambda,\alpha}(y) = \alpha\lambda(\lambda y)^{\alpha-1}\exp\{-(\lambda y)^{\alpha}\}$, $y \geqslant 0, \lambda > 0, \alpha > 0$. Then $g_{\lambda,\alpha}(y) = \alpha\lambda\exp\{-(\lambda y)^{\alpha}\}/\Gamma(1/\alpha)$. In the following, we suppose that there are n independent and identically distributed (i.i.d.) realizations $\{(v_i, \boldsymbol{x}_i)\}_{i=1}^{n}$ of (V, X) and that the true values of $\lambda, \alpha, \boldsymbol{\beta}, \boldsymbol{\theta}$ are $\lambda_0, \alpha_0, \boldsymbol{\beta}_0, \boldsymbol{\theta}_0$, respectively.

The motivation for formulating (1) is as follows. Assume that an infected case with covariates X has incubation period T satisfying

$$\log(T) = X'\boldsymbol{\gamma} + \epsilon_f.$$

The error term ϵ_f has density function $f(\cdot)$. Thus, the conditional density of T given $X = \boldsymbol{x}$ is $f(t\exp(\boldsymbol{x}'\boldsymbol{\beta}))\exp(\boldsymbol{x}'\boldsymbol{\beta})(\boldsymbol{\beta} = -\boldsymbol{\gamma})$. When $\boldsymbol{\gamma} = 0$, i.e., X has no effect on the incubation period T, the density of T will be $f(t)$ and typically is assumed to be the Weibull distribution density. This is the reason for our choice of $f_{\lambda,\alpha}$ for f in the density (1).

For a confirmed infected case, let V denote his/her time lag between departure from Wuhan and onset of symptoms, which can be considered as the forward time in a renewal process, and let A be the time lag between infection and departure from Wuhan, which can be considered as the backward time and unobservable. As pointed out in Qin et al. (2020), V is a length-biased version of the incubation period T, since it is easier to observe V if $T = A + V$ is longer. Now, for the above infected case who has covariates $X = \boldsymbol{x}$ and incubation period density function $f(t\exp(\boldsymbol{x}'\boldsymbol{\beta}))\exp(\boldsymbol{x}'\boldsymbol{\beta})$, given $X = \boldsymbol{x}$, by renewal process theory, the joint density of (A, V) is

$$\frac{f((a+v)\exp(\boldsymbol{x}'\boldsymbol{\beta}))\exp(\boldsymbol{x}'\boldsymbol{\beta})}{\mu(\boldsymbol{x})}, \qquad \mu(\boldsymbol{x}) = \int tf(t\exp(\boldsymbol{x}'\boldsymbol{\beta}))\exp(\boldsymbol{x}'\boldsymbol{\beta})\,dt.$$

Marginally, given $X = \boldsymbol{x}$, A and V have the same density, i.e.,

$$g(v\exp(\boldsymbol{x}'\boldsymbol{\beta}))\exp(\boldsymbol{x}'\boldsymbol{\beta}), \qquad g(v) = \frac{\bar{F}(v)}{\int \bar{F}(t)\,dt}, \qquad \bar{F}(t) = \int_t^{\infty} f(w)\,dw.$$

This is equivalent to $\log(V) = X'\boldsymbol{\gamma} + \epsilon_g$, and the error term ϵ_g has density function $g(\cdot)$.

In short, given $X = \boldsymbol{x}$, if $A = 0$ and V actually is the complete incubation period T, then it has density $f(v\exp(\boldsymbol{x}'\boldsymbol{\beta}))\exp(\boldsymbol{x}'\boldsymbol{\beta})$. Otherwise, its density is $g(v\exp(\boldsymbol{x}'\boldsymbol{\beta}))\exp(\boldsymbol{x}'\boldsymbol{\beta})$. Combined with the mix proportion $\pi(\boldsymbol{x}, \boldsymbol{\theta})$, which is typically assumed to have logistic regression model form, and our choice for f, we obtain the proposed density (1).

Unlike the classic two-component mixture model, the two density components in (1) share the same parameters $\lambda, \alpha, \boldsymbol{\beta}$. Thus, estimation of $\lambda, \alpha, \boldsymbol{\beta}$ is always possible. As is well known, the unidentifiability problem of parameters in a Gaussian mixture model is caused by the unconstrained mix proportion (Boldea and Magnus, 2009), but this does not happen with our model. The identification of the parameter of interest $\boldsymbol{\beta}$ in our mixture regression model can be summarized as the following theorem.

Theorem 1. *If (V, X) has conditional density (1), then if $\alpha \neq 1$ holds in the proposed mixture density (1), all parameters in the density are identifiable. However, the parameter of interest $\boldsymbol{\beta}$ is always identifiable.*

Proof: See the Supplementary Material.

In practice, we obtain the MLE $(\hat{\lambda}_{ML}, \hat{\alpha}_{ML}, \hat{\boldsymbol{\beta}}_{ML}, \hat{\boldsymbol{\theta}}_{ML})$ of $(\lambda, \alpha, \boldsymbol{\beta}, \boldsymbol{\theta})$ through maximizing the conditional likelihood of the observed data:

$$\mathcal{L}(\lambda, \alpha, \boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{i=1}^{n} h(v_i | \boldsymbol{x}_i, \lambda, \alpha, \boldsymbol{\theta}, \boldsymbol{\beta}). \tag{2}$$

Under some regularity conditions, when the underlying model is (1), the consistency of the MLE is obvious. Interestingly, in some cases, the MLE of $\boldsymbol{\beta}$ is robust to misspecified model setting.

Theorem 2. *If $\pi(\boldsymbol{x}, \boldsymbol{\theta})$ in (1) is known to be independent of $\boldsymbol{x}$, which means $\boldsymbol{\theta}_1 \equiv \boldsymbol{0}$ and the mix proportion is constant, then, even when the underlying incubation period distribution $F_{\lambda,\alpha}$ actually is not the Weibull c.d.f., $\hat{\boldsymbol{\beta}}_{ML}$ obtained by our model still is a consistent estimate of $\boldsymbol{\beta}_0$.*

Proof: See the Supplementary Material.

Next, we discuss how to calculate the MLE. Owing to $\pi(\boldsymbol{x}, \boldsymbol{\theta})$, direct maximization of the conditional likelihood of the observed data is hard to implement. Thus, the EM algorithm is adopted in this paper. We give the detailed computation procedure in the Supplementary Material.

3 Simulation Studies

In this section, we conduct several simulation studies to test the performance of the proposed method, its sensitivity, and the inference method used in the following section. Monte Carlo samples of size n are independently generated B times and the estimate Est is averaged over estimates of all replications, and SE is the standard error of the B estimates.

3.1 Estimation Performance

Here, we first test the estimation performance using two examples.

Example 1. The data generating model is (1) with

$$\boldsymbol{\beta} = (\beta_1, \beta_2)' = \begin{bmatrix} 0.8 \\ 0.6 \end{bmatrix}, \qquad \lambda = 0.2, \qquad \alpha = 2,$$

$$X = [X_1, X_2]' \sim N\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \right), \qquad \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

Since the conditional observed likelihood (2) may have various local maxima, a random initialization method is adopted here. We set the initial values as $\lambda_I, \alpha_I, \boldsymbol{\beta}_I = [\beta_{I1}, \beta_{I2}]', \boldsymbol{\theta}_I = [\theta_{I0}, \theta_{I1}, \theta_{I2}]'$. For a test of this method, we take two settings:

1. Fixed initial value setting: $\lambda_I = 0.1, \alpha_I = 3.5, \beta_{Ii} = 0, i = 1, 2, \theta_{Ik} = 0, k =, 0, 1, 2$. This setting is denoted by FS.
2. Random initial value setting: $\lambda_I = U(0, 1), \alpha_I = U(1, 10), \beta_{Ii} = U(-5, 5), i = 1, 2, \theta_{Ik} = U(-5, 5), k =, 0, 1, 2$. In this random setting, the estimate is chosen as the one with the maximum likelihood among 10 and 100 estimates, which are obtained by starting at 10 and 100 random initial points, respectively. The setting starting at 10 random initial points is denoted by $Rnd10$ and that starting at 100 random initial points by $Rnd100$.

Table 1: Results for Example 1.

Settings		FS		$Rnd10$		$Rnd100$		SR	
Quantity	True value	Est	SE	Est	SE	Est	SE	Est	SE
λ	0.2	0.1944	0.0075	0.1978	0.0079	0.1993	0.0078	–	–
α	2.0	2.0677	0.0701	2.0372	0.0716	2.0236	0.0660	–	–
β_1	0.8	**0.8081**	0.0253	**0.8045**	0.0249	**0.8025**	0.0243	0.7131	0.0232
β_2	0.6	**0.5986**	0.0181	**0.5962**	0.0173	**0.5917**	0.0172	0.5205	0.0221
θ_0	0	−0.1900	0.3814	−0.0686	0.3975	−0.0126	0.4011	–	–
θ_1	1.0	1.0128	0.3846	1.0731	0.4203	1.0982	0.4057	–	–
θ_2	1.0	0.8406	0.2931	0.9086	0.3312	0.9308	0.3193	–	–

For the purpose of comparison, a competing model is chosen as a regression model fitted by regressing the logarithm of the observed time lag over covariates. From the discussion about the motivation for (1), the estimate of β_i $(i = 1, 2)$ is the negative regression coefficient estimate of this simple regression. We denote this estimation method by SR. The simulation results with $n = 1500$, $B = 200$ are summarized in Table 1.

From Table 1, we see that when we implement the proposed EM estimation method, the regression coefficients $\boldsymbol{\beta}, \boldsymbol{\theta}$ are estimated very well, no matter the initial setting with which the EM computation procedure starts. However, compared with the results of FS, the random initialization method produces results with smaller bias and SE. Comparing the results of $Rnd10$ and $Rnd100$, we note that the proposed EM computation procedure is robust to a random choice of initial values and thus is reasonable. From comparison with the results of $Rnd10$ and $Rnd100$, it is found that estimates obtained by SR will produce large bias, and this may lead to unreliable statistical inference. This shows the greater practical power of our model and method.

Example 2. The data generating model is the same as that in Example 1, but with $\lambda = \alpha = 1$. This actually means $h(v|\boldsymbol{x}, \lambda, \alpha, \boldsymbol{\theta}, \boldsymbol{\beta}) = w(v \exp(\boldsymbol{x}'\boldsymbol{\beta})) \exp(\boldsymbol{x}'\boldsymbol{\beta})$, $w(v) = e^{-v}$. We again take two initialization settings:

1. Fixed initial value setting: $\lambda_I = 0.1, \alpha_I = 2.5, \beta_{Ii} = 0, i = 1, 2, \theta_{Ik} = 0, k =, 0, 1, 2$. This setting is denoted by FS.
2. Random initial value setting: $\lambda_I = U(0, 1), \alpha_I = U(1, 10), \beta_{Ii} = U(-3, 3), i = 1, 2, \theta_{Ik} = U(-3, 3), k =, 0, 1, 2$. In this setting, the estimate is chosen as the one with the maximum likelihood among 10 estimates, which are obtained by starting at 10 random initial points. This setting is denoted by $Rnd10$.

The competing model is the same as in Example 1. The simulation results with $n = 1500$, $B = 200$ are summarized in Table 2.

In Example 2, since $V|\boldsymbol{x} \sim w(v \exp(\boldsymbol{x}'\boldsymbol{\beta})) \exp(\boldsymbol{x}'\boldsymbol{\beta})$, it is obvious that $\boldsymbol{\theta}$ has no effect on the generated data, and Theorem 1 shows that $\boldsymbol{\theta}$ is unidentifiable. However, $\boldsymbol{\beta}$ is still identifiable, and so we can estimate it. We see that the EM implementation with fixed initial value setting and the 10 random initial values setting produces unbiased estimates of $\boldsymbol{\beta}$. However, interestingly, SR also produces an unbiased result. The reason behind this is as follows.

In this special case, $\log V = \boldsymbol{x}'\boldsymbol{\gamma} + \epsilon$, $\boldsymbol{\gamma} = -\boldsymbol{\beta}$, and $e^{\epsilon} \sim \exp(1)$. This can be rewritten as $\log V = c + \boldsymbol{x}'\boldsymbol{\gamma} + \epsilon'$, $\epsilon' = \epsilon - E\epsilon$, and $c = E\epsilon$. Thus, the SR method can produce unbiased

Table 2: Results for Example 2.

Settings		*FS*		*Rnd*10		*SR*	
Quantity	True value	Est	SE	Est	SE	Est	SE
λ	1.0	1.0040	0.0820	1.0034	0.0910	–	–
α	1.0	1.0006	0.0445	1.0039	0.0554	–	–
β_1	0.8	**0.8001**	0.0356	**0.7993**	0.0360	0.8026	0.0494
β_2	0.6	**0.6013**	0.0357	**0.6008**	0.0368	0.6026	0.0439
θ_0	0	−2.4434	0.6487	NA	NA	–	–
θ_1	1.0	−1.3047	1.2544	NA	NA	–	–
θ_2	1.0	−1.0822	2.2313	NA	NA	–	–

NA means that the absolute value of the quantity is large; e.g., for θ_2, Est and SE are respectively −27.2430 and 26.9600.

results. However, compared with the results obtained by our method, the SE produced by the *SR* method is much larger.

3.2 Sensitivity Analysis

In practice, $\boldsymbol{\theta}$ is not usually the parameter of interest, and too large a value of θ_i may lead to a computational barrier. In the absence of a constraint, convergence may then be slow. To avoid this problem, we can constrain each component of $\boldsymbol{\theta}$ to a given range when implementing optimization. Thus, it is necessary to test the sensitivity of the estimates of $\lambda, \alpha, \boldsymbol{\beta}$ to such constrained optimization. For this purpose, we first explore the case with no covariate. Let $f_{\lambda,\alpha}(y)$ and $g_{\lambda,\alpha}(y)$ be defined as before. The mixture density function $m(v; \lambda, \alpha, p) = pf_{\lambda,\alpha}(v) + (1-p)g_{\lambda,\alpha}(v)$ for different mix proportions p $(0 \leqslant p \leqslant 1)$ is shown in Figures 1 and 2 for $\lambda = 0.2, \alpha = 2$ and $\lambda = 0.1, \alpha = 0.6$, respectively. From these figures, it can be seen that in contrast to the multimodality of the traditional mixture model, the mixture of the Weibull distribution and its length-biased version is unimodal when $\alpha > 1$ and has no peak when $\alpha \leqslant 1$. This makes it impossible to determine the distribution from which the sample arises. Fortunately, this is not our goal, which is simply estimation of the parameters of interest λ, α. Take $\lambda = 0.2, \alpha = 2, p = 0.5$ as an example. Through fixing the specific value and range of p, we get the MLE of the relevant unknown parameters. The simulation results with $n = 1500$, $B = 200$ are summarized in Table 3.

From Table 3, for the case of fixed p, it can be seen that a departure from the true value p_0 of p will bring some bias. It should be noted that the greater the departure, the larger will be the bias. This is also reflected in Figure 1. For the case of a fixed range, when this range contains p_0, the MLE produces unbiased estimates. When the range does not contain p_0, the estimates will have some bias. Furthermore, the greater the departure of the range from p_0 is, the larger will be the bias.

For the case with covariates, we still take Example 1 with the 10 random initial value setting: $\lambda_I = U(0,1)$, $\alpha_I = U(1,10)$, $\beta_{Ii} = U(-5,5)$ $(i = 1,2)$, $\theta_{Ik} = U(a,b)$ $(k = 0,1,2)$. The maximization step is also constrained in $\{\theta_i \in [a,b], i = 1,2,3\}$. The results are summarized in Table 4. From this table, it can be seen that compared with the result with no constraint, no matter how large the fixed range of $\boldsymbol{\theta}$ is, when it contains $\boldsymbol{\theta}_0$, the estimates of $\lambda, \alpha, \boldsymbol{\beta}$ are very

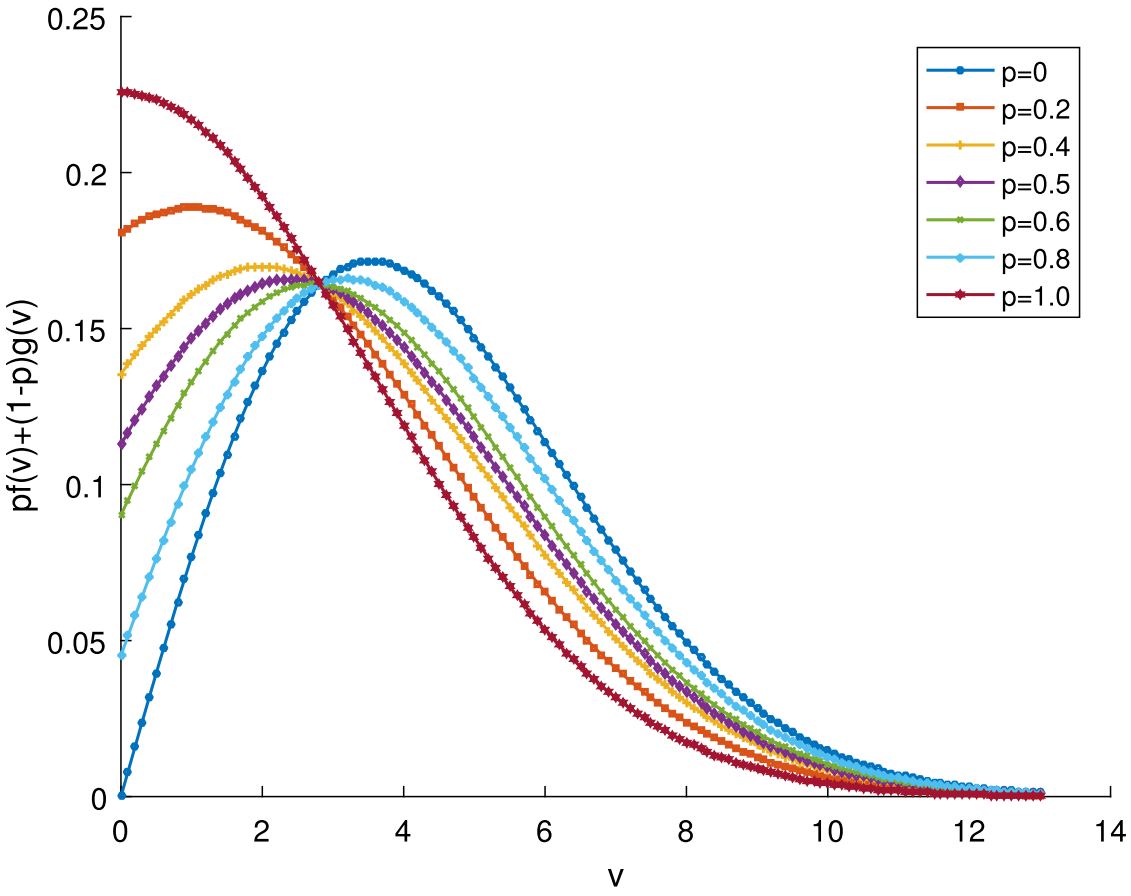

Figure 1: Density function $m(v; \lambda, \alpha, p) = pf_{\lambda,\alpha}(v) + (1-p)g_{\lambda,\alpha}(v)$ for $\lambda = 0.2, \alpha = 2$.

stable and are still unbiased. These results indicate the greater reliability of the proposed method. However, when the true value of θ_i is not contained in the fixed range, such as $(-2, -0.5)$, $(-1, 0)$, the greater the departure of the range from $\boldsymbol{\theta}_0$, the larger will be the bias of the MLE of $\boldsymbol{\beta}$. For example, the departure of $(-2, -0.5)$ from $\boldsymbol{\theta}_0$ is greater than that of $(-1, 0.5)$, and the results in the $(-2, -0.5)$ case have larger bias than those in the $(-1, 0.5)$ case. The reason behind this is that $\boldsymbol{\beta}$ is actually combined with the scale parameter λ in the mixture model. Thus, the sensitivity of the estimate of $\boldsymbol{\beta}$ in the mixture regression model is just a reflection of the sensitivity of the estimate of the scale parameter in the no-covariate case. Therefore, in practice, a larger fixed range is preferred when permitted. This is also what we have done in the real data analysis.

3.3 Hypothesis Testing

In Section 4, l.r.t.'s will be used to make inferences for the parameters of interest. However, in Wolfe (1971) and the book by Everitt and Hand (1981), it is noted that standard l.r.t.'s may fail for the mixture model since the asymptotic distribution is no longer a χ^2 distribution. Aitkin and Rubin (1985) also pointed out this problem and placed a prior distribution on the mix proportion to make the asymptotic distribution of l.r.t.'s approximate the classical standard

Table 3: Sensitivity analysis for $m(v; \lambda, \alpha, p)$ with $\lambda = 0.2, \alpha = 2, p = 0.5, n = 1500, B = 200$.

Fixed value of p		0.1		0.2		0.4		0.5	
Quantity	True value	MLE	SE	MLE	SE	MLE	SE	MLE	SE
λ	0.2	0.1553	0.0032	0.1675	0.0033	0.1897	0.0037	0.2001	0.0038
α	2.0	2.6868	0.1535	2.4619	0.1236	2.1330	0.0869	2.0055	0.0752
Fixed value of p		0.6		0.8		0.9		1.0	
Quantity	True value	MLE	SE	MLE	SE	MLE	SE	MLE	SE
λ	0.2	0.2102	0.0040	0.2303	0.0044	0.2404	0.0045	0.2507	0.0046
α	2.0	1.8916	0.0656	1.6864	0.0499	1.5894	0.0426	1.4901	0.0351
Fixed range of p		(0.1,0.4)		(0.3,0.6)		(0.7,0.9)		(0,1)	
Quantity	True value	MLE	SE	MLE	SE	MLE	SE	MLE	SE
λ	0.2	0.1890	0.0043	0.1988	0.0090	0.2203	0.0042	0.1993	0.0098
α	2.0	2.1436	0.0950	2.0287	0.1319	1.7856	0.0573	2.0240	0.1392
p	0.5	0.3934	0.0218	0.4901	0.0802	0.7008	0.0073	0.4954	0.0912

Table 4: Results of sensitivity analysis for Example 1

10 random initial points setting							
Fixed range (a, b) of $\boldsymbol{\theta}$		(−3,3)		(−5,5)		(0.5,2)	
Parameter	True value	Est	SE	Est	SE	Est	SE
λ	0.2	0.1981	0.0082	0.1983	0.0084	0.2046	0.0063
α	2.0	2.0354	0.0741	2.0325	0.0758	2.0051	0.0636
β_1	0.8	0.8043	0.0250	0.8043	0.0252	0.7934	0.0214
β_2	0.6	0.5959	0.0175	0.5959	0.0174	0.5868	0.0152
θ_0	0	−0.0561	0.4086	−0.0592	0.4247	0.5146	0.0576
θ_1	1.0	1.0811	0.4233	1.0904	0.4158	0.8534	0.3073
θ_2	1.0	0.9194	0.3419	0.9356	0.3715	0.7126	0.2447
Fixed range (a, b) of $\boldsymbol{\theta}$		(-1,0.5)		(-2,-0.5)		No constraint	
Parameter	True value	Est	SE	Est	SE	Est	SE
λ	0.2	0.1991	0.0062	0.1638	0.0078	0.1978	0.0079
α	2.0	2.0787	0.0585	3.0110	0.2014	2.0372	0.0716
β_1	0.8	0.7925	0.0194	0.7427	0.0241	0.8045	0.0249
β_2	0.6	0.5887	0.0153	0.5357	0.0181	0.5962	0.0173
θ_0	0	0.4327	0.1333	−0.5672	0.3104	−0.0686	0.3975
θ_1	1.0	0.4935	0.0340	−0.5000	0.0000	1.0731	0.4203
θ_2	1.0	0.4943	0.0277	−0.5040	0.0463	0.9086	0.3312

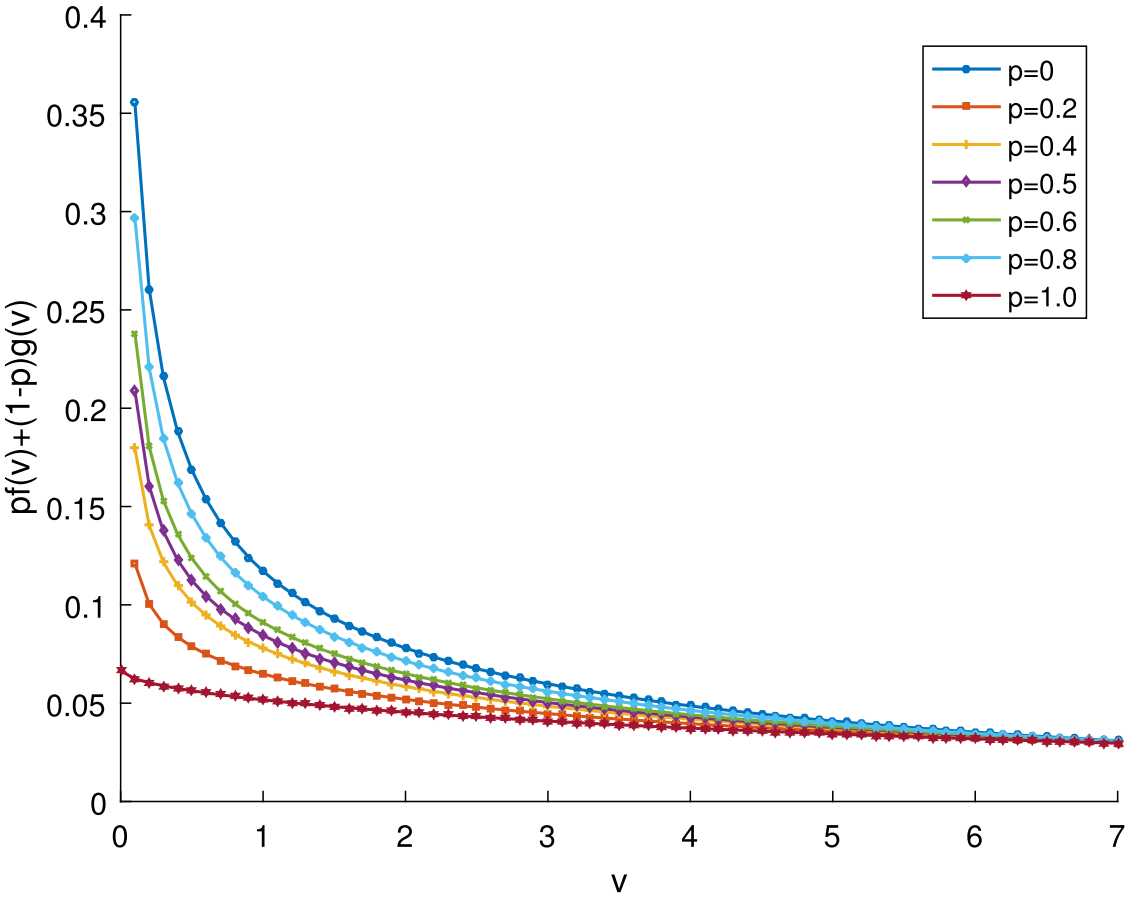

Figure 2: Density function $m(v;\lambda,\alpha,p) = pf_{\lambda,\alpha}(v) + (1-p)g_{\lambda,\alpha}(v)$ for $\lambda = 0.1, \alpha = 0.6$.

asymptotic distribution. Thus, it is necessary to test whether the l.r.t.'s and the corresponding classical standard asymptotic distribution work for our model. In this subsection, we adopt the model from Example 1 with different $\boldsymbol{\beta}$ and test three null hypotheses:

$$H_{10} : \boldsymbol{\beta} \equiv \mathbf{0}, \qquad H_{20} : \beta_1 = 0, \qquad H_{30} : \beta_2 = 0.$$

We still use the EM algorithm to obtain the estimates $\{\hat{\lambda}_{i0}, \hat{\alpha}_{i0}, \hat{\boldsymbol{\beta}}_{i0}, \hat{\boldsymbol{\theta}}_{i0}\}$ of all parameters under the constraint H_{0i} $(i = 1, 2, 3)$ and the estimates $\{\hat{\lambda}_0, \hat{\alpha}_0, \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\theta}}_0\}$ of all parameters without constraints. The same random initialization setting as that in Example 1 is adopted, but we use 20 random initial values here. The maximization step also is constrained in $\{\theta_i \in [-5, 5], i = 1, 2, 3\}$. For $H_{i0}(i = 1, 2, 3)$, the l.r.t. statistic $l_{H_{i0}}$ is

$$l_{H_{i0}} = -2\log\left[\frac{\mathcal{L}(\hat{\lambda}_{i0}, \hat{\alpha}_{i0}, \hat{\boldsymbol{\beta}}_{i0}, \hat{\boldsymbol{\theta}}_{i0})}{\mathcal{L}(\hat{\lambda}_0, \hat{\alpha}_0, \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\theta}}_0)}\right].$$

Here, $\mathcal{L}(\lambda, \alpha, \boldsymbol{\beta}, \boldsymbol{\theta})$ is given by (2), and the p-value is calculated as $p_{H_{i0}} = P\{l_{H_{i0}} \leqslant \chi^2_{dim_i}\}$, where $\chi^2_{dim_i}$ is the chi-square random variable with degree dim_i equal to the number of constrained parameters in H_{i0}. In our simulations, as illustrations, we reject the null hypothesis H_{i0} if

Table 5: Results of hypothesis testing simulation with different $\boldsymbol{\beta}$.

$\boldsymbol{\beta}$	NH		
	H_{10}	H_{20}	H_{30}
$[0.8, 0.6]'$	100%	99.5%	100%
$[0.8, 0]'$	100%	100%	2.5%
$[0, 0]'$	3.0%	4.0%	3.0%

Table 6: Results of hypothesis testing simulation with $\boldsymbol{\beta} = [0.1, 0]'$ and different n.

n	NH		
	H_{10}	H_{20}	H_{30}
500	21.5%	28.0%	5.0%
1000	37.0%	52.5%	5.0%
1500	62.0%	72.5%	6.5%
2000	77.5%	86.5%	3.5%
2500	86.0%	94.5%	3.0%
3000	90.5%	96.0%	3.5%

$p_{H_{i0}} \leqslant 0.05$. We calculate and present the percentage RP of rejected null hypotheses among B replications in the relevant tables, i.e.,

$$RP = \frac{\text{number of rejected null hypotheses among } B \text{ replications}}{B}.$$

First, we set $n = 500, B = 200$ and simulate with different $\boldsymbol{\beta} = [0.8, 0.6]', [0.8, 0]', [0, 0]'$. The results are summarized in Table 5, where NH denotes the null hypothesis. From this table, we can see that the l.r.t.'s control the type I error well, since when the null hypothesis is true, the false rejection rate is less than 5%. Even better, in such a case, the type II error is also very small. The reason for this is that the regression coefficients β_1, β_2 are relatively large and thus the signal is strong.

Consequently, as an another test, we set $B = 200, \boldsymbol{\beta} = [0.1, 0]'$, which means that X_2 has no effect and X_1 has a very weak effect. Since generally test power depends on sample size, we simulate with various n to show the effect of sample size. The simulation results are summarized in Table 6, where NH again denotes the null hypothesis. From this table, it can be noted that the type-I error is still controlled well and is reasonable for the classical standard l.r.t.'s. Compared with the results for $\boldsymbol{\beta} = [0.8, 0]'$, we see that small but nonzero regression coefficients lead to a sharp increase in the type II error, but increasing the sample size can increase the power of the proposed test and reduce the type II error. All of these results show that the l.r.t.'s and the corresponding classical standard asymptotic distributions work well for our model. It is therefore appropriate to adopt these l.r.t.'s and asymptotic distributions.

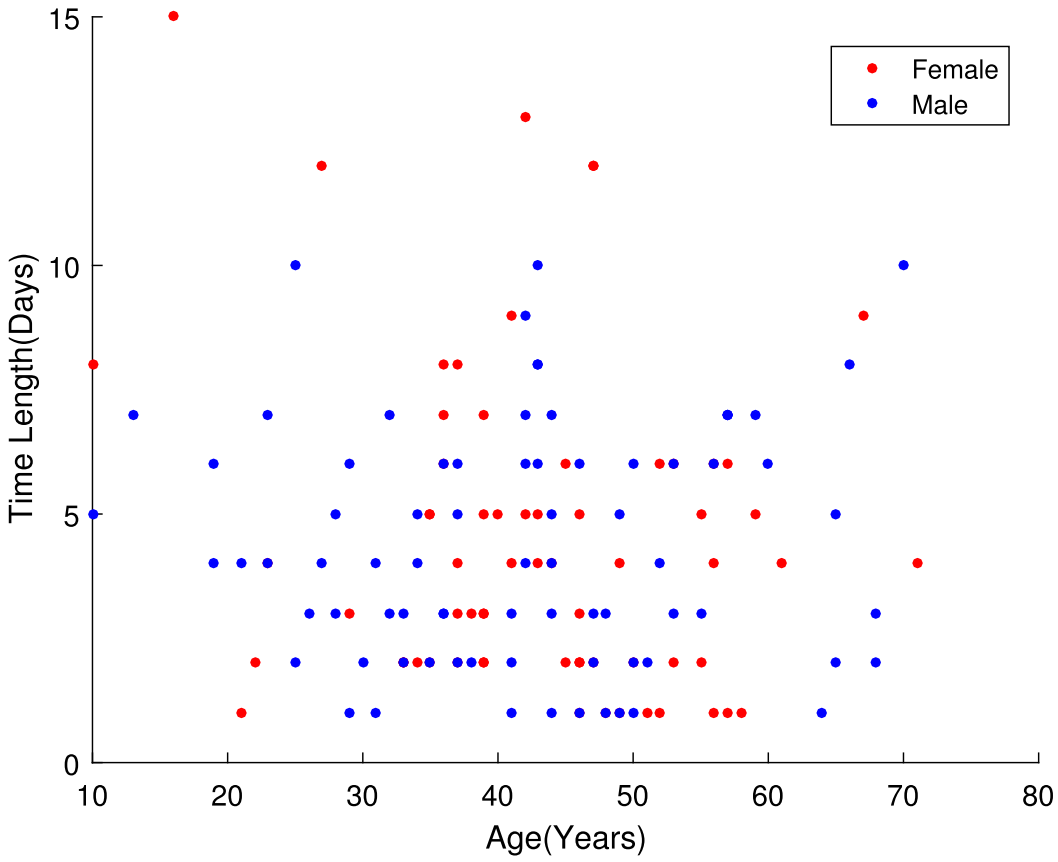

Figure 3: Data description.

4 Real Data Analysis

In this section, the proposed model and method are applied to a real dataset. Our real data analysis is based on 143 observations of time lag between departure from Wuhan and onset of symptoms, which is of count type but is treated as continuous for our method. These 143 cases left Wuhan between January 19 and January 23, 2020 for Zhejiang. The observed risk factors are *Age* and *Gender* (Female is set as 1). There are 66 female cases and 77 male cases. Figure 3 presents a description of our data, where Time Length is the time lag between the departure of a case from Wuhan and their onset of symptoms. Some summary statistics for the real data are also presented in Table 7.

In the estimation of $\lambda, \alpha, \boldsymbol{\beta}$ in the model (1) for our real data, through an EM computational procedure starting with $\lambda_I = 0.1, \alpha_I = 2.0, \boldsymbol{\beta}_I = [0; 0], \boldsymbol{\theta}_I = [0; 0; 0]$ without other constraints, it is found that the estimate of $(\lambda, \alpha, \beta_{gender}, \beta_{age})$ is $(0.1849, 1.6713, -0.0673, 0.0029)$.

4.1 Sensitivity Analysis

Actually, without any other constraints, the estimate of $\boldsymbol{\theta}$ is very unstable. However, $\boldsymbol{\theta}$ is not a parameter in which we are interested. For sensitivity analysis, we take a 20 random initial values computational procedure and choose the final estimate as we have done in the simulations.

Table 7: Data summary statistics.

Variable	Gender	*Mean*	*Min*	*Max*	25%	*Median*	75%	90%	*Std*
Age	Female	43.08	10.00	71.00	37.00	43.00	51.00	56.90	11.45
	Male	41.77	10.00	70.00	32.00	42.00	50.00	59.80	13.47
Time Length	Female	4.45	1.00	15.00	2.00	4.00	6.00	8.90	3.23
	Male	4.35	1.00	10.00	2.00	4.00	6.00	7.00	2.42

Table 8: Results of sensitivity test for a fixed range of θ.

Value of l	ML	λ	α	β_{gender}	β_{age}	θ_1	θ_2	θ_3
3	−329.9835	0.1849	1.6713	−0.0673	0.0029	1.5369	1.5986	1.0841
5	−329.9835	0.1849	1.6713	−0.0673	0.0029	2.4741	1.3965	1.0296
10	−329.9942	0.1861	1.6826	−0.0785	0.0027	9.9893	−4.6786	−0.0475

ML is the corresponding maximum likelihood.

Specifically, the random initialization setting is $\lambda_I = U(0,1), \alpha_I = U(1,10), \beta_{Ii} = U(-1,1), i = 1,2, \theta_{Ik} = U(-l,l), k =, 0, 1, 2$, where l is a pre-assumed positive constant. The maximization step is also constrained in $\{\theta_i \in [-l,l], i = 0,1,2\}$. The results of the sensitivity test are summarized in Table 8. From this table, we can see that with different constrained ranges of $\boldsymbol{\theta}$ when randomly initializing and implementing the maximization step, the estimate of $\boldsymbol{\theta}$ is unstable, but $\lambda, \alpha, \boldsymbol{\beta}$ and the corresponding maximum likelihood are robust to these settings. This makes the estimates of interest $\lambda, \alpha, \boldsymbol{\beta}$ more reliable.

4.2 Hypothesis Test for the Effect of Age and Gender

Finally, as the simulation results in Section 3.3 have shown, l.r.t.'s are allowed for β_{gender} and β_{age}. We gives the corresponding test results under three null hypotheses: $H_1 : \beta_{gender} = \beta_{age} = 0$, $H_2 : \beta_{gender} = 0$, and $H_3 : \beta_{age} = 0$. The computation procedure is implemented with the setting $l = 5$ in Section 4.1. The results are summarized in Table 9. All these results show that it is statistically not rejected that age and gender have no effect on the incubation period. Based on this on-hand evidence, a common quarantine period for the whole population is reasonable and there is no need to specify different quarantine measures for different groups.

As an illustration of this result, we can estimate the incubation periods for groups with various ages and genders. Suppose that, given $X = \boldsymbol{x}$, the conditional density of V is

$$\pi(\boldsymbol{x},\boldsymbol{\theta}) f_{\lambda,\alpha}(v \exp(\boldsymbol{x}'_{new}\boldsymbol{\beta}) \exp(\boldsymbol{x}'_{new}\boldsymbol{\beta}) + (1 - \pi(\boldsymbol{x},\boldsymbol{\theta})) g_{\lambda,\alpha}(v \exp(\boldsymbol{x}'_{new}\boldsymbol{\beta})) \exp(\boldsymbol{x}'_{new}\boldsymbol{\beta}), \tag{3}$$

where $\boldsymbol{x} = (x_{Gender}, x_{Age})'$, $\boldsymbol{x}_{new} = (x_{Gender}, I\{x_{Age} > 45\})'$, and $x_{Gender} = 1$ when the case is female. The other notation is as before. We still adopt the 20 random initial values setting with $l = 10$. The results are summarized in Table 10. From these, we can see that the estimates of the Weibull distribution parameters and the mean and quantiles of the incubation periods

Table 9: Results of hypothesis test.

Hypothesis	H_1	H_2	H_3
CML	−330.5694	−330.2039	−330.3030
LLRS	1.1719	0.4408	0.6390
p-value	0.5566	0.5068	0.4241

CML is the corresponding constrained maximum likelihood.
LLRS is the log-likelihood ratio statistic.

Table 10: Results for the new model (3).

		Female				Male		
Age	> 45		$\leqslant 45$		> 45		$\leqslant 45$	
Quantity	Est	CI	Est	CI	Est	CI	Est	CI
λ	0.210	(0.161,0.266)	0.184	(0.148,0.228)	0.227	(0.178,0.273)	0.199	(0.168,0.228)
α	1.673	(1.538,1.993)	1.673	(1.538,1.993)	1.673	(1.538,1.993)	1.673	(1.538,1.993)
$Mean$	4.254	(3.362,5.539)	4.857	(3.908,6.036)	3.935	(3.276,4.987)	4.493	(3.930,5.307)
$Q_{0.05}$	0.807	(0.615,1.249)	0.922	(0.720,1.361)	0.747	(0.572,1.166)	0.853	(0.676,1.209)
$Q_{0.25}$	2.262	(1.803,3.055)	2.583	(2.112,3.368)	2.092	(1.693,2.897)	2.389	(2.032,3.013)
$Q_{0.50}$	3.826	(3.033,4.997)	4.368	(3.565,5.479)	3.539	(2.930,4.618)	4.041	(3.506,4.893)
$Q_{0.75}$	5.789	(4.563,7.460)	6.610	(5.277,8.140)	5.355	(4.469,6.599)	6.115	(5.352,7.125)
$Q_{0.90}$	7.840	(6.141,9.802)	8.951	(7.126,10.78)	7.252	(6.046,8.711)	8.280	(7.196,9.472)
$Q_{0.95}$	9.175	(7.166,11.33)	10.48	(8.258,12.69)	8.487	(7.048,10.11)	9.691	(8.369,11.06)
$Q_{0.99}$	11.86	(9.175,14.41)	13.55	(10.50,16.27)	10.97	(8.991,13.01)	12.53	(10.60,14.33)

CI is the 95% confidence interval obtained from 1000 times bootstraps.
Q_p is the quantile corresponding to probability p of Weibull distribution with the estimated parameter.

of different groups are similar. There are some differences in the quantile estimates, but these are due to the limited sample size. The confidence intervals of all the quantiles contain the corresponding quantile estimates of the different groups. Thus, these estimates should make only a little difference, and this implicitly confirms the hypothesis test result.

5 Concluding Remarks

In this paper, we have proposed a novel mixture regression model to analyze the effects of age and gender on the incubation period of COVID-19. An EM method is used to obtain estimates of the parameters of interest, and the simulation results show that the proposed method outperforms the simple regression method and has robustness. The hypothesis test simulations also show that the application of l.r.t.'s and the corresponding classical standard asymptotic distribution is a

reasonable approach. It should be noted, however, that we use a simple random initialization method for the EM algorithm. To the best of our knowledge, the way in which a good starting value should be chosen for the EM algorithm remains a difficult and unsolved problem deserving of further work in the future.

By applying the proposed method to a Zhejiang COVID-19 dataset, it has been found that age and gender statistically have no effect on the incubation period of COVID-19. Thus, the quarantine period currently in operation, which is the same for everybody, is reasonable. This result has direct significance for COVID-19 prevention work and future economic recovery. However, our sample size still seems relatively limited, and so in the future it will be worthwhile collecting more data to further confirm our results.

Supplementary Material

The Supplementary Material including the detailed proofs of Theorems 1 and 2, can be found on the *Journal of Data Science* website. The data/code used in the analyses can be found at https://github.com/SimonsZheng/Assessment-of-Effects-of-Age-and-Gender-on-COVID-19.

Acknowledgments

The authors would like to thank the editor and two anonymous referees for their insightful comments. Zhou's work was supported by the State Key Program of the National Natural Science Foundation of China (71931004).

References

Aitkin M, Rubin D (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society, Series B*, 47(1): 67–75.

Backer JA, Klinkenberg D, Wallinga J (2020). Incubation period of 2019 Novel Coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Euro Surveill*, 25(5): 2000062.

Boldea O, Magnus J (2009). Maximum likelihood estimation of the multivariate normal mixture model. *Journal of the American Statistical Association*, 104(488): 1539–1549.

Chen JH (2017). Consistency of the MLE under mixture models. *Statistical Science*, 32(1): 47–63.

Dimitris K, Evdokia X (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3–4): 577–590.

Everitt BS, Hand DJ (1981). *Finite Mixture Distributions*. Chapman and Hall, London.

Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. (2020). Clinical characteristics of Coronavirus Disease 2019 in China. *The New England Journal of Medicine*, 382: 1708–1720.

Jiang WX, Tanner AM (1999). Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. *Annals of Statistics*, 27(3): 987–1011.

Khalili A, Chen JH (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102(479): 1025–1038.

Lauer S, Grantz K, Bi QF, Jones F, Zheng QL, Meredith H, et al. (2020). The incubation period of Coronavirus Disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172(9): 577–582.

Li Q, Guan XH, Wu P, Wang XY, Zhou L, Tong YQ, et al. (2020). Early transmission dynamics in Wuhan, China, of Novel Coronavirus infected pneumonia. *The New England Journal of Medicine*, 382(13): 1199–1207.

Liang FM, Liu CH, Carroll RJ (2010). *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*. Wiley, New York.

Linton NM, Kobayashi T, Yang YC, Hayashi K, Akhmetzhanov A, Jung SM, et al. (2020). 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.

McLachlan GJ, Peel D (2000). *Finite Mixture Model*. Wiley, New York.

Qin J, You C, Lin QS, Hu TJ, Yu SC, Zhou XH (2020). Estimation of incubation period distribution of COVID-19 using disease onset forward time: A novel cross-sectional and forward follow-up study. *Science Advances*, 6(33): eabc1202.

Qin YC, Priebe CE (2013). Maximum L_q-likelihood estimation via the expectation-maximization algorithm: A robust estimation of mixture models. *Journal of the American Statistical Association*, 108(503): 914–928.

Wolfe J (1971). A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of normal distributions. Naval Personnel and Training Research Laboratory San Diego, Technical Bulletin STB 72-2.