The Journal of Applied Science **Vol. 20 No. 2: 23-33 [2021]**
วารสารวิทยาศาสตร์ประยุกต์ doi: 10.14416/j.appsci.2021.02.003
ISSN 1513-7805 Printed in Thailand ISSN 2586-9663 (Online)

Research Article

Cubic B-spline and Generalised Linear Models for COVID-19 Patients in Thailand

Orathai Polsen[1], Pianpool Kamoljitprapa[1*]

[1] Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
***E-mail**: pianpool.k@sci.kmutnb.ac.th
Received: 18/03/2021; Revised: 29/05/2021; Accepted: 03/06/2021

Abstract

Thailand is one of the countries which has been affected by the coronavirus disease (COVID-19) pandemic as it is known a severe acute respiratory disease. COVID-19 was first emerged in Thailand in January, 2020 and the number of infected people on a daily basis has increased significantly over a year. A valid data set of COVID-19 cases in Thailand was collected by the Department of Disease Control, Ministry of Public Health. This paper, therefore, was subjected to the report of cases to the estimation of statistical models using count autoregressive regression based on Poisson distribution and cubic B-spline regression analysis to quantify patterns in the incidence for different groups of COVID-19 patients in Thailand. The findings in Phase I (12th January, 2020 – 19th December, 2020) show that male patients with specific age groups influence the spread of infected disease. In addition, the cubic B-spline with optimal tuning equal to 8 is the best-fitting model in Phase II (20th December, 2020 – 15th February, 2021). The models may be used to predict numbers of COVID patients overall, as well as in key subgroups. In addition, the findings will better inform the planning of the national strategies in preventing and controlling the pandemic.

Keywords: Autoregressive Model, Coronavirus, Cubic B-spline, Generalised Linear Model, Poisson Regression

Introduction

The coronavirus disease (COVID-19) has currently spread across the world and has become a worldwide pandemic which totaled a number of confirmed cases to exceed hundred thousand and is still significantly increasing daily. According to the World Health Organization (WHO) report of 15th March, 2020, the five countries with the highest cumulative number of confirmed cases of COVID-19 infections are China with 81,048, Italy with 21,157, Iran with 12,729, Republic of Korea with 8,162 and Spain with 5,753 ("Coronavirus disease 2019 (COVID-19) Situation Report – 55", 2020). The virus spread among the people through breathing, sneezing, coughing and talking in close contact. The most common symptoms of COVID-19 are high fevers, coughs, body aches and shortness of breath. Other symptoms include the loss of the sense of smell, breathing difficulties and chest pain, however, in some cases, they may display no symptoms at all. The severe of COVID-19 is lung affected and multi-organ failure. In order to prevent the spreading of the virus, we need to increase our personal hygiene: washing hands with soap frequently, having a face mask on and also social distancing (Grant et al., 2020; Singhai, 2020).

The first case of COVID-19 pandemic in Thailand was reported on 12th January, 2020, when a Wuhan resident who travelled to Bangkok tested positive for the virus and within the first month, the confirmed cases were 19. Eventhough the initial number was not as high as in China or other countries, the confirmed cases increased dramatically to 1,609 at the end of March. The government announced the locking down of the country and controlled people who wanted to commute from hometown to other provinces. In March and April, the government allowed Thais in abroad countries to come back to Thailand and which showed an increasing number of new cases who had been infected from outside countries. To control the spreading of the disease, they had to state quarantine for 14 days before going back to their hometown and that has contributed to a reduction in the number of new cases infected in Thailand in May and later. The majority number of cases who were infected in Thailand, exclude Thais, is Burma who most working in Bangkok, Nonthaburi, Samut Sakhon and Songkhla. Whereas the source of infected cases from foreign countries into Thailand, the top five are India, France, US, UK and China.

According to the reports from Department of Disease Control, Ministry of Public Health, the outbreak of pandemic had two phases, the first one is covering the period from the announced date on 12th January, 2020 to 19th December, 2020 while the second period covers from the announced date on the 20th of December to the present day. The daily confirmed cases used in this paper were collected from the first case until 15th February, 2021, which totals 24,714 in all ("Open government data of Thailand", 2020).

The number of daily confirmed cases is count-valued time series. The methodologies based on generalised linear model were introduced for modelling on count time series. Zeger (1988) proposed the method for time series of counts and explored the relationship between a response variable with a within-group correlation structure and a set of independent variables. The model was used for investigating trends in US polio incidence. Fokianos & Fried (2012) considered an autoregressive Poisson model for count time series data. Tsay (1984) considered a linear regression model for the autocorrelation function of observed series using the method of least squares. The models due to Kirdwichai (2017) are based on Poisson and negative binomial distributions that dealing with a count data in incidence of AIDS in Bangkok.

Indeed, many ongoing researches have made use of splines to model the continuing time periods within non-linear time series models. Huang & Shen (2004) explored a smoothing method using polynomial spline to fit regression models for non-linear time series. Amorim et. al. (2008) proposed a method for estimating time-varying coefficients in the rates model using regression B-spline. We also considered spline regression to investigate the time-varying autoregressive model for counts to study the spread of COVID-19 in Thailand.

This paper attempts to model the confirmed cases for two phases which the analysis used R software (R Core Team, 2016). For the primary model, we wish to identify the group of gender and age that affect COVID-19 incidences using generalised linear model (GLM). In the second phase, the gender and age are not fully reported. As a result, the GLM cannot be implemented to estimate incidences for different subgroups. The cubic B-spline regression is therefore considered as a statistical method to fit the model of cumulative number of cases instead.

Materials and Methods

One of the most statistical techniques for investigating and modelling the relationship between response and explanatory variables is regression analysis which can be linear and non-linear. A regression is called linear when it is linear in parameters. A regression model for response $\boldsymbol{y}_i$ measured at n_i time points for the group is

$$\boldsymbol{y}_i = \boldsymbol{f}(\boldsymbol{x}_i;\boldsymbol{\beta}) + \boldsymbol{\varepsilon}_i,\ i = 1,2...,r \tag{1}$$

where $\boldsymbol{f}(\boldsymbol{x}_i;\boldsymbol{\beta}) = [f(x_{i,1},\boldsymbol{\beta}),..., f(x_{i,n_i},\boldsymbol{\beta})]'$ is the vector for function f of the explanatory variables $\boldsymbol{x}$ and unknown parameter vector $\boldsymbol{\beta} = (\beta_0\ \beta_1 \cdots \beta_k)'$, and $\boldsymbol{\varepsilon}_i$ is a vector of errors with covariance matrix $\sigma^2\boldsymbol{V}$ (Kirdwichai, 2017). The general form of the model using matrix notation is

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of responses, X is an $n \times (k+1)$ design matrix, $\boldsymbol{\beta}$ is a $(k+1) \times 1$ vector of unknown parameters and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors. The estimation of parameters can be accomplished by a method of ordinary least squares (OLS). The least-squares function is expressed as

$$S(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta})$$

where X is a design matrix. To find the vector of least-squares estimators, $\hat{\boldsymbol{\beta}}$, the function S is minimised with respect to unknown parameters and equate the derivatives to zero. The least-squares estimators must satisfy

$$\left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = -2X'\boldsymbol{y} + 2X'X\hat{\boldsymbol{\beta}} = \boldsymbol{0}$$

(Montgomery et al., 2012).

The hypothesis testing for the individual parameter, β_j, will be tested under the null hypothesis

$$H_0 : \beta_j = 0.$$

Generalised Linear Model

The number of COVID 19 confirmed cases is longitudinal count data which can be classed in Poisson distribution. The normality assumption for the model in Equation (1) is therefore invalid. The generalised linear models (GLMs) allow response variable from different distributions (McCullagh & Nelder, 1989). The model can be expressed in general form as shown in Equation (2). The parameters can be estimated by a method of maximum likelihood. The likelihood function based on Poisson $(\exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}))$ distribution (Cameron & Trivedi, 2013) is

 doi: 10.14416/j.appsci.2021.02.003

$$L(\beta_0,\beta_1,...,\beta_k)=\prod_{i=1}^{n}\exp\left[-\exp\left(\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}\right)\right]\times\frac{1}{y_i!}\times \left[\exp\left(\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}\right)\right]^{y_i}.$$

In this paper the number of the patients over time was investigated. This study examined the model of COVID-19 patients where monthly count of cases is response variable and month and gender are independent variables.

Cubic B-spline Model

The polynomial regression model is one of the importance class where the response is curvilinear. Durbin (1960) researched on regression model with an autoregressive model of order 1 (AR (1)). The model shown in Equation (3) is the polynomial regression of degree k when the response at a time T depends on previous time $T-1$:

$$y=\sum_{j=0}^{k}\beta_j x^j+\varepsilon\ , \tag{3}$$

where y is a response, β_j, $j=0,1,...,k$ are unknown parameters and ε is an error which has normal distributed with mean 0 and variance σ^2 (Montgomery et al., 2012).

The polynomial regression can sometimes be poor to fit to the data when the function behaves differently in different parts of the range of x. Piecewise polynomial spline is one of the techniques that has been able to deal with this issue. Splines (Rodríguez, 2001; Fan & Gijbels, 1996) are piecewise polynomial of order k. The cubic spline (k = 3) is usually adequate for most practical problems. The basic cubic spline model with h knots,

$$t_1<t_2<\cdots<t_h,$$

with continuos first and second derivatives is

$$E(y)=S(x)=\sum_{j=0}^{3}\beta_{0j}x^j+\sum_{i=1}^{h}\beta_i(x-t_i)_+^3$$

where

$$(x-t_i)_+=\begin{cases}(x-t_i) & \text{if } x-t_i>0\\ 0 & \text{if } x-t_i\le 0.\end{cases}$$

However, the number of knots is still left behind to be an issue and if the number of knots is large, the matrix $X'X$ becomes ill-conditioned. The method called the cubic B-spline can overcome this obstacle (Montgomery et al., 2012; Eubank, 1999). The cubic B-spline model with sequence h knots is shown in Equation (4).

$$E(y) = S(x) = \sum_{i=1}^{h+4} \gamma_i B_i(x), \tag{4}$$

$$B_i(x) = \sum_{j=i-4}^{i} \left[\frac{(x - t_j)_+^3}{\prod_{\substack{m=i-4 \\ m \neq j}}^{i} (t_j - t_m)} \right], \quad i = 1, 2, ..., h+4$$

where γ_i, $i = 1, 2, ..., h+4$, are parameters to be estimated.

Since a spline is a piecewise polynomial and its shape depends on the degree of the spline function and the number of internal breakpoints. Although, the cubic spline (degree = 3) is practical use, the tuning parameters which are degree of freedom and number of knots to obtain the best has to be investigated. Sookkhee et al. (2021) evaluated the optimal parameters in the spline regrssion by tuning the degree of freedom in genome-wide study. The finding of optimal parameter using bs() function in R language (Perperoglou et al., 2019) depending on tuning parameters which are degree of freedom (df), knots and degree. The coefficients were computed for regression quantile B-spline with fixed knots and varying the different df (Molinari et al., 2004; Sookkhee et al., 2021; Sookkhee et al., 2021).

In this paper, a cubic B-spline was selected to fit the cumulative number of confirmed cases in Phase II of the disease spreading due to the announced information is incomplete in gender and age. The optimal parameters are an extension of the study to obtain the best fit to the data.

Results and Discussion

The COVID-19 population of 24,714 studied here have been divided into two phases, the early phase of the outbreak, Phase I, has 4,331 (17.52%) cases while the new emerging, Phase II, has 20,383 (82.48%) cases. All 4,331 confirmed cases in Phase I consisted of 2,356 (54.40%) males and 1,975 (45.60%) females, while gender class and age group in Phase II, are not available online. The patient statistics in Phase I, the numbers and the percentages of confirmed cases, are shown in Table 1. In this research the average age of patients is 38 years and the most confirmed age group is between 25-29 years as can be seen in Figure 1.

Confirmed Cases in Phase I

The cumulative numbers of monthly confirmed cases for both males and females in each age group are shown in Figure 2 and Figure 3, respectively.

We applied GLM based on Possion distribution with a log link function to fit the models for monthly cumulative number of cases by gender. The estimates, standard errors and p-values are shown in Table 2. The analysis results indicate that the fitted model is appropriate to the data. A scatter plot of the monthly cumulative number for two gender groups, male and female, and fitted models are shown in Figure 4.

In order to investigate the effect of age subgroups in the groups of gender, the Poisson regression is also applied. The p-values in Table 3 reveal the only specific age subgroups that are significant at the level 0.05.

Vol. 20 No. 2: 23-33 [2021]
doi: 10.14416/j.appsci.2021.02.003

Table 1 The number of confirmed cases (percentage) in Phase I for each age groups.

Age group (years)	***Sex***		***Total***
	Male	***Female***	
≤ 4	34 (53.97%)	29 (46.03%)	63 (1.45%)
5 - 9	25 (55.56%)	20 (44.44%)	45 (1.04%)
10 - 14	22 (41.51%)	31 (58.49%)	53 (1.22%)
15 - 19	58 (48.33%)	62 (51.67%)	120 (2.77%)
20 - 24	223 (42.88%)	297 (57.12%)	520 (12.01%)
25 - 29	305 (46.42%)	352 (53.58%)	657 (15.17%)
30 - 34	290 (51.06%)	278 (48.94%)	568 (13.11%)
35 - 39	271 (53.88%)	232 (46.12%)	503 (11.61%)
40 - 44	249 (57.11%)	187 (42.89%)	436 (10.07%)
45 - 49	232 (61.38%)	146 (38.62%)	378 (8.73%)
50 - 54	185 (60.86%)	119 (39.14%)	304 (7.02%)
55 - 59	170 (65.38%)	90 (34.62%)	260 (6.00%)
≥ 60	292 (68.87%)	132 (31.13%)	424 (9.79%)
	2,356 (54.40%)	**1,975 (45.60%)**	**4,331**

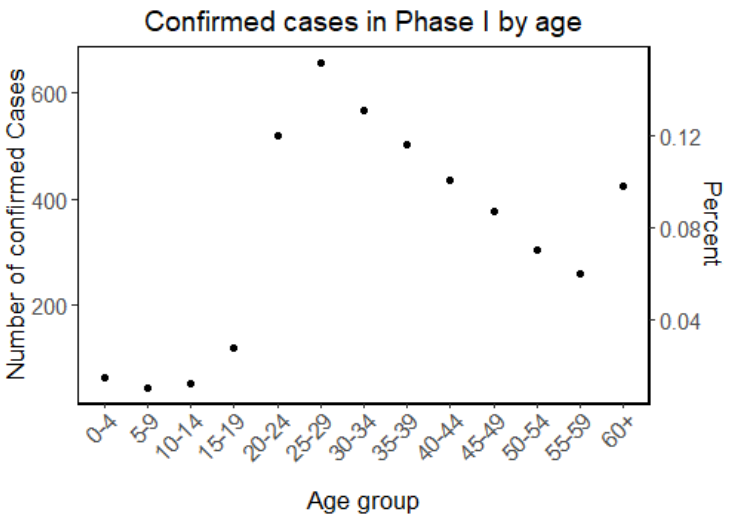

Figure 1 Scatter plot of COVID-19 confirmed cases in Phase I by age.

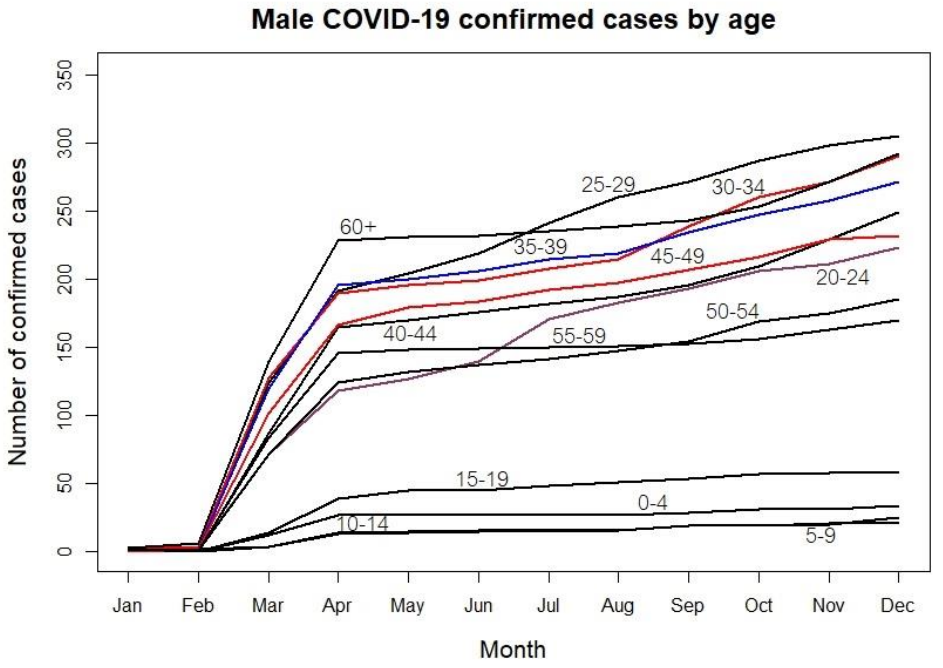

Figure 2 Profile plot of cumulative number of male COVID-19 patients in Phase I by age.

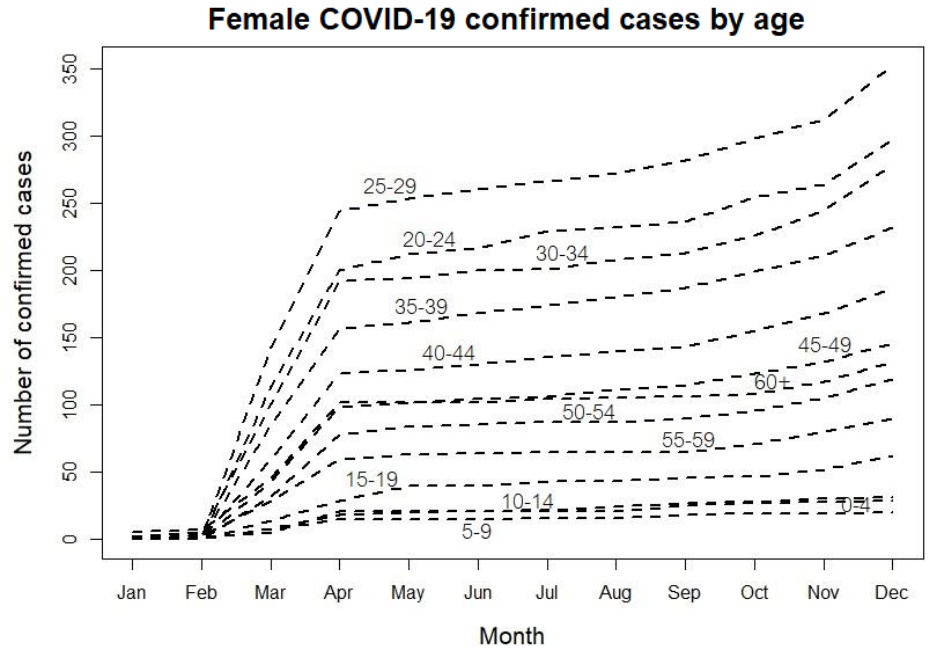

Figure 3 Profile plot of cumulative number of female COVID-19 patients in Phase I by age.

Table 2 Summary statistics for the Poisson regression model.

Parameter	***Estimate***	***Standard error***	***p-value***
β_0	6.2740	0.0145	0.0000
β_1	0.2254	0.0110	0.0000
β_2	0.1341	0.0017	0.0000

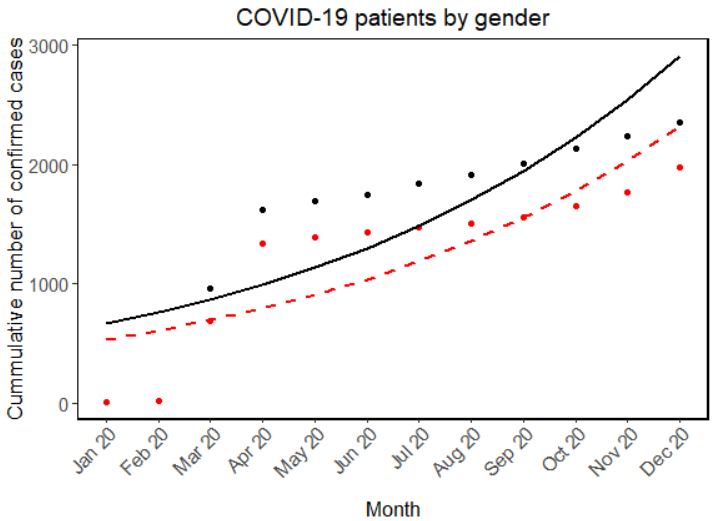

Figure 4 Scatter plot of the monthly cumulative number of COVID-19 patients in Phase I. The lines are fitted models for two gender groups, male (solid line) and female (dashed line).

Table 3 The p-values for age groups that are significant.

Age group	***p-value***
15-19	0.0113
30-34	0.0443
35-39	0.0478
40-44	0.0000
50-54	0.0013
55-59	0.0000

A possible explanation why gender is significant for some specific age groups is due to the fact that it does not take into account many other factors that influence the spread of infected disease, for example, the behaviour and life style of individuals (e.g. travel, social, careers and government policies, etc.)

Confirmed Cases in Phase II

In order to find a good fit model for the cumulative number of daily confirmed cases, we follow the model, given in previous section by cubic B-splines. To obtain the optimal parameters of the spline, the tuning df in the range of 3 - 10 was done using bs() function in R software. The Akaike's information criteria (AIC) was chosen to be a criterion to obtain the optimal tuning. In this study, the optimal df = 8 makes the best fit to the data and gives AIC = 780.3175. The parameter estimates, standard errors and p-values in cubic B-spline model are shown in Table 4.

It can be seen from the results (Table 4), the individual p-values of the coefficients are significant at the level of 0.05 and these estimates given the model has R^2 = 0.9998 and adjusted-R^2 = 0.9997. A scatter plot of daily cumulative number of COVID-19 patients in Phase II and the fitted line is shown in Figure 5.

Table 4 Summary statistics for the cubic B-spline model.

Parameter	***Estimate***	***Standard error***	***p-value***
γ_1	2171.5	171.7	0.0000
γ_2	1099.4	182.8	0.0000
γ_3	6046.5	149.6	0.0000
γ_4	7675.4	143.1	0.0000
γ_5	9438.5	151.0	0.0000
γ_6	20831.3	189.0	0.0000
γ_7	19311.8	193.3	0.0000
γ_8	20505.7	152.4	0.0000

Conclusion

This paper examined the good fitting models and analysed incidents of COVID-19 in Thailand for early and emerging pandemic phases. The GLM based on Poisson distribution with log link function is chosen to fit the monthly cumulative number of cases for gender group in the first phase to quantify trends in incidence. In addition, the incidence in each age subgroup were explored. In the second pandemic phase, the cubic B-spline regression was used to find the best model that fit the daily cumulative number of confirmed cases. Extending the evaluation, the issue of the optimal parameter for the spline was also explored. In this paper, we use R programming language to perform statistical analysis to obtain the model that fit the data.

In the first phase, the analysis of GLM based on Poisson distribution shows that male patients with subgroups of age 15-19, 30-44 and 50-59 influence the spread of infected disease. Unfortunately, the complete daily information of patients, gender and age has not been reported in the second phase, therefore, the cubic B-spline regression model was used to fit the model instead. Investigation on the optimal tuning of spline function found that df = 8 gives the best-fitting model as it has the lowest AIC and high R^2.

The models are beneficial for predicting the numbers of COVID patients and identifying the key subgroups that are affected by the disease. Moreover, the findings will alarm the government, setting the policies in order to control the pandemic. For further research, it could be interesting to apply the approaches on other count data which has subgroups. The investigation of other techniques for fitting a model could be also a possible topic of research.

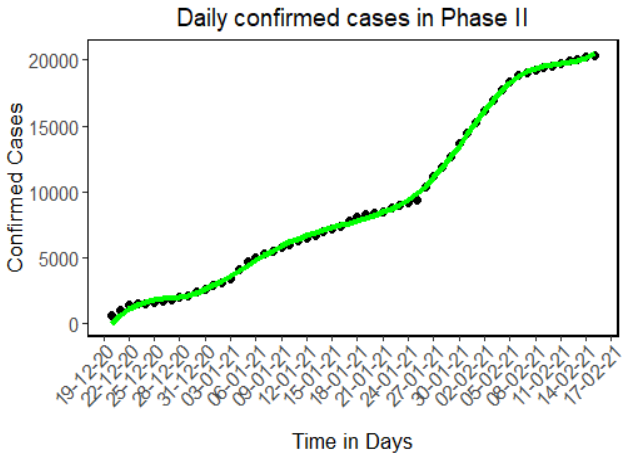

Figure 5 Scatter plot of daily cumulative number of COVID-19 patients in phase II. The solid line is fitted model using cubic B-spline.

Acknowledgements

We extend our sincere thanks to Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, who funded and contributed resources to this study (Fiscal year 2021, contract no. 642060). We also thanks to Mr. Natapat Kirdwichai for proof reading the article.

References

Amorlim, L. D., Cai, J., Zeng, D., & Barreto, M. L. (2008). Regression splines in the time-dependent coefficient rates model for recurrent event data. *Statistical in Medicine*, *27* (28), 5890–5906.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). New York: Cambridge University Press.

Coronavirus disease 2019 (COVID-19) Situation Report – 55. (2020). https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200315-sitrep-55-covid-19.pdf (Accessed: 18 March 2021).

Durbin, J. (1960). Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Society: Series B*, *22*, 139–153.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing* (2nd ed.). New York: Marcel Dekker, Inc.

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications* (1st ed.). London: Chapman and Hall.

Fokianos, K., & Fried, R. (2012). Interventions in log-linear Poisson autoregression. *Statistical Modelling*, *12* (4), 299–322.

Grant, M. C., Geoghegan, L., Arbyn, M., Mohammed, Z., McGuinness, L., Clarke, E. L., & Wade, R. G. (2020). The prevalence of symptoms in 24,410 adults infected by the novel coronavirus (SARS-CoV-2; COVID-19): A systematic review and meta-analysis of 148 studies from 9 countries. *PLOS ONE*, *15* (6). https://doi.org/10.1371/journal.pone.0234765

Huang, J. Z., & Shen, H. (2004). Functional coefficient regression models for non-linear time series: a polynomial spines approach. *Scandinavian Journal of Statistics*, *31* (4), 515–534.

Kirdwichai, P. (2017). Identifying trends and patterns in incidence of AIDS in Bangkok using generalised linear mixed models. In *proceedings of the World Congress on Engineering 2017 Vol II, 5-7 July 2017*, (pp. 574-578). London. http://www.iaeng.org/publication/WCE2017/WCE2017_ pp574-578.pdf

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear model* (2nd ed.). New York: Chapman and Hall.

Molinari, N., Durand, J., & Sabatier, R. (2004). Bounded optimal knots for regression splines. *Computer Statistics and data analysis*, *45*, 159–178.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). New Jersey: Wiley.

Open government data of Thailand. (2020). https://data.go.th/dataset

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A review of Spline function procedures in R. *BMC Medical Research Methodology*, *19*, 1–16.

R Core Team. (2016). R: *A Language and Envirnonment for Statistical Computing, R Foundation for Statistical Computing*. https://www.R-project.org/

Rodríguez, G. (2001). *Smoothing and Non-Parametric Regression*. http://data.princeton.edu/eco572/ smoothing.pdf

Singhai, T. (2020). A review of Coronavirus disease-2019 (COVID-19). *Indian J Pediatr*, *87* (4), 281-286. https://doi.org/10.1007/s12098-020-03263-6

Sookkhee, S., Kirdwichai P., & Baksh, M. F. (2021). The efficiency of single SNP and SNP-set analysis in genome-wide association studies. *Songklanakarin J. Sci. Technol*, *43* (1), 243-251.

Sookkhee, S., Kirdwichai P., & Baksh, F. (2021). The optimal parameters of Spline regression for SNP-set analysis in genome-wide association study. *Science & Technology Asia*, *26* (1), 39–52.

Tsay, R. (1984). Regression models with time series errors. *Journal of the American Statistical Association, 79* (385), 118-124.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika*, *75* (4), 921–929.