STATISTICS IN TRANSITION new series, June 2022
Vol. 23 No. 2, pp. 197–208, DOI 10.2478/stattrans-2022-0024
Received – 20.06.2020; accepted – 28.07.2021

On the quick estimation of probability of recovery from COVID-19 during first wave of epidemic in India: a logistic regression approach

Hemlata Joshi[1], **S. Azarudheen**[2], **M. S. Nagaraja**[3], **Singh Chandraketu**[4]

ABSTRACT

The COVID-19 pandemic has recently become a threat all across the globe with the rising cases every day and many countries experiencing its outbreak. According to the WHO, the virus is capable of spreading at an exponential rate across countries, and India is now one of the worst-affected country in the world. Researchers all around the world are racing to come up with a cure or treatment for COVID-19, and this is creating extreme pressure on the policy makers and epidemiologists. However, in India the recovery rate has been far better than in other countries, and is steadily improving. Still in such a difficult situation with no effective medicine, it is essential to know if a patient with the COVID-19 is going to recover or die. To meet this end, a model has been developed in this article to estimate the probability of a recovery of a patient based on the demographic characteristics. The study used data published by the Ministry of Health and Family Welfare of India for the empirical analysis.

Key words: COVID-19, epidemic, coronavirus disease, recovery estimation, logistic regression, logit analysis.

1. Introduction

Coronaviruses are the group of related RNA viruses which has ribonucleic acid as its genetic material. These viruses cause diseases in humans, other mammals and birds and sickness may range from common cold to severe respiratory diseases. COVID-19

[1] CHRIST (Deemed to be University), Bangalore, India. E-mail: hemlata.joshi28@gmail.com. ORCID: https://orcid.org/0000-0002-4051-6330

[2] CHRIST (Deemed to be University), Bangalore, India. E-mail: azarudheen.s@christuniversity.in. ORCID: https://orcid.org/0000-0001-7568-4273.

[3] CHRIST (Deemed to be University), Bangalore, India. E-mail: nagaraja.ms@christuniversity.in. ORCID: https://orcid.org/0000-0002-6900-8436.

[4] CHRIST (Deemed to be University), Bangalore, India. E-mail: chandraketu.lko@gmail.com. ORCID: https://orcid.org/0000-0003-2367-5396.

is the most recent disease that has jumped off to humans. Initially the eruption of the novel coronavirus was documented in China's Wuhan at the beginning of December 2019 and then circularized all across the world. Often during coughing or sneezing, the infection of coronavirus disease disseminates from one human to others via droplets raised from the respiratory system of the infected humans (WHO, 2020). The COVID-19 symptoms generally include fever, dry cough, tiredness, and in severe cases, infection can lead pneumonia, shortness of breath, chest pain, loss of speech or movement, kidney failure, and even death (WHO, 2020), but approximately 20 percent of the cases have been deemed to be severe (Singh et al., 2020). The World Health Organization (WHO) announced this COVID-19 a pandemic on 11 March 2020 and ingeminated the call for countries to take quick actions and scale up response to treat, detect and reduce transmission to save people's lives. The developed countries such as the United States of America, Italy, Spain, France, UK, etc. are struggling to overcome the disease spreading by novel coronavirus. According to WHO, by the end of May 2020 it has spread in around 188 countries, the total number of cases have exceeded 6 million and approximately 3.7 lakh deaths worldwide. In India, the first case of coronavirus infection was observed in Kerela on 30 January 2020 and for the two months, the spread of the coronavirus disease was extremely slow may be due to the strict nationwide lockdown. After that, the Government of India gave the conditional relaxation in the nationwide lockdown and during this period of lockdown, the coronavirus cases started increasing with the exponential rate. Although the incubation period for the coronavirus disease has not been confirmed yet, from the pooled analysis it is seen that the symptoms may appear in 2 days to 14 days (Singhal, 2020) and the Government of India has declared minimum 14 days quarantine period for the suspected cases. In the absence of any efficacious medicine or vaccination, the social distancing has been consented as a most efficient scheme for cutting the severity of this coronavirus disease all across the globe (Ferguson et al., 2020; Singh et al., 2020).

As India is the second largest most populated country and majority of the population live under the inadequate hygiene and with insufficient medical facilities such as lack of testing kits, labs and health personnel, etc., and with the relaxation in lockdown, the coronavirus disease may start spreading at community level. In the middle of June, the total confirmed COVID-19 cases crossed 3.43 lakh with an increase of more than ten thousand cases in a single day and the new cases was rising at the record pace while the deaths have come up to 9900 with 380 fatalities. If the same rate continues, India will reach the sixth position in the most affected countries by COVID-19, and presently India is the 7th worst affected country after the USA, Brazil, Russia, UK, Spain and Italy (WHO), and in terms of the fatality rate, India is at the twelfth position while it is ranked 8th in terms of recovery rate from coronavirus disease currently.

The Prime Minister of India Mr. Narendra Modi stated that currently India is being listed amongst the countries with the least number of deaths due to coronavirus and also said that the death rate can still be reduced if we all follow all the guidelines suggested by WHO. PM Modi also said that the decision of nationwide lockdown on time served better in controlling the speed of spreading of coronavirus disease in India. According to the ICMR's serological survey, about 0.73% of the population was exposed to the virus by the mid-June and India could have 200 million COVID infected people by September. The Indian Council of Medical Research (ICMR) said that India was not in community transmission yet but a large chunk of the population is at risk and physical distancing and other similar measures need to continue. The return of millions of migrants to villages in Bengal, UP, Bihar, Orissa, Chhattisgarh, Jharkhand, etc. will lead to a surge of infections in these rural hinterlands.

As COVID-19 is a new pandemic, it has become a challenging task in front of the scientists and researchers to fight with this coronavirus disease in the absence of vaccine. Thus, to know its behaviour and nature a lot of research is being done all across the globe, so that it could help the scientists or epidemiologists to possibly cure humans from its infections. The published data on COVID-19 pandemic are analysed by many researchers by using various mathematical modelling approaches (Rao et al., 2020; Chen et al., 2020). Huang et al. (2020) worked on the clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Modelling and forecasting of the COVID-19 pandemic is done by Anastassopoulou et al. (2020), Corman et al. (2020), Rothe et al. (2020) and Gamero, J. et al. (2020) and many interesting results have been obtained using the principles of mathematical modelling. Nikolay et al. (2020) used the coronavirus data and compared the Verhult model with the half-logistic curve of growth with polynomail variable transfer model. Further, they have compared the Verhulst growth model with Verhulst curve of growth with polynomail variable transfer model on the Covid-19 data and also have studied the intrinsic properties of some models of growth with polynomial variable transfer that give a very good approximation of the specific data on the pandemics in Cuba. Zaliskyi et al. (2020) built a mathematical model for COVID-19 data of European countries. In this article, an effort has been made to estimate the probability of recovery from the coronavirus disease using the indirect method of estimation. For this a logistic regression techniques has been used and for the empirical analysis, the available information about the demographic variables such as age and gender of the patients, which was published by the Ministry of Health and Family Welfare, Government of India, is utilized.

2. The Model and Methodology

Here, the variable of the interest is the status of the patient whether the patient recovered or deceased after the infection of COVID-19. The status of the patient can take only two values – either 0 if the patient deceased due to COVID-19 or 1 if recovered, and we want to estimate the probability of dying or survived after getting the infection of COVID-19 as a function of the indicator variables such as gender (male or female) and various age groups (0–20, 21–40, 41–60 and 60 and over). Since the response variable is of a dichotomous type, the logistic regression modelling technique is applied for the estimation of the probability whether the patient will die or recover by using various age groups and gender of the patients.

Let π denote the probability of recovery from the corona disease of a patient for the given values of p predictor variables and the relationship between the probability π and p predictors can be represented by the logistic model (see Chatterjee, S. and Hadi, Ali S. (2006)), i.e.

$$\begin{aligned}\pi &= Pr(Y = 1 \mid X_1 = x_1,..., X_p = x_p) \\ &= \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2 ....., \beta_pX_p+\varepsilon}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2 ..., \beta_pX_p+\varepsilon}}\end{aligned} \tag{1}$$

The function given in equation (1) is the logistic regression function. It is non-linear in the regression coefficients $\beta_0, \beta_1 ... \beta_p$ and it is linearised by the logit transformation, i.e. if the probability of an event that the patient recovers from the corona disease is π then the ratio $\frac{\pi}{1-\pi}$ obtained is the odds for the recovery from the coronavirus disease.

Since

$$\begin{aligned}1-\pi &= Pr(Y=0 \mid X_1 = x_1,...,X_p = x_p) \\ &= \frac{1}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+.....+\beta_px_p+\varepsilon}}\end{aligned} \tag{2}$$

Then,

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1x_1+\beta_2x_{2+}+.....+\beta_px_p+\varepsilon} \tag{3}$$

Taking natural log both sides, we get

$$\begin{aligned}logit(\pi) &= log(\frac{\pi}{1-\pi}) \\ &= \beta_0+\beta_1x_1+\beta_2x_2+.....+\beta_px_p+\varepsilon\end{aligned} \tag{4}$$

Here, the function $logit\ (\pi)$ in equation (4) is a linear function of explanatory variables x_i (i=1,…,p) in terms and it is called the logit function and the range of π in equation (1) is between 0 and 1 while the range of the values of $log(\frac{\pi}{1-\pi})$ is between $-\infty$ and ∞, which makes the logits more appropriate linear regression fitting, and the disturbance term ε satisfies all the basic assumptions of ordinary least squares.

Now, our predictor variables are categorical type so the dummy variables are created for each of the categorical predictors. If the regression model contains an intercept term, the number of dummies defined should always be one less than the number of categories of that variable. Let G be the dummy variable for the gender of the patient which have only two categories (male and female), i.e. $G=1$ if the patient is male, 0 otherwise. Similarly, the dummy variables for the age having four age groups is $A_t; t=1,2,3$ and it can be defined as

$$\begin{aligned} A_1 &= 1; \text{ If the patient lies in the age group } 21-40 \\ &= 0; \text{ Otherwise} \\ A_2 &= 1; \text{ If the patient lies in the age group } 41-60 \\ &= 0; \text{ Otherwise and} \\ A_3 &= 1; \text{ If the patient lies in the age group 60 and above} \\ &= 0; \text{ Otherwise} \end{aligned} \tag{5}$$

Here, the female category in the dummy variable G and the age group 0–20 in the A_t dummy variable are taken in the reference category and the logit model can be written as

$$\begin{aligned} logit(\pi) &= log(\frac{\pi}{1-\pi}) \\ &= \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \beta_3 A_3 + \beta_4 G + \varepsilon \end{aligned} \tag{6}$$

3. Empirical Analysis

For the estimation of the probability of recovery of a patient infected by coronavirus disease in India, the data issued by the Ministry of Health and Family Welfare (MoHFW, India) are utilized. In the analysis, 427 patients have been included due to the lack of availability of data on all the patients and the data on the patients' status from all over India are taken from between 30 January 2020 to 30 May 2020, which is shown in Table 1. From the available data, an effort has been made to estimate the probability of recovery from coronavirus disease in India. For this, the

logistic regression technique is used and the developed model is shown in equation (6), where age group and gender of the patients are the indicator variables and π is the probability of recovery of a patient from coronavirus disease. The analysis is done using *RStudio* (R Core Team (2020)) and the results obtained are shown in Table 2. The estimated model is given as

$$\begin{aligned} logit\ (\pi) &= log\ (\frac{\pi}{1-\pi}) \\ &= 0.0401 + 0.9346\ A_1 - 1.5913\ A_2 - 2.0101\ A_3 - 0.1071\ G \end{aligned} \tag{7}$$

Now, from Table 2, it can be seen that the age groups 41–60 and 60 and over are significant at 0.05 level of significance as their p-values are smaller than the 0.05and the log odds of recovery from the corona disease are −1.5913 and −2.0101 for the age group 41–60 and 60 and over respectively. For a better understanding of the results, the exponentiated terms of the regression coefficients has also been computed, which is shown in Table 3. If we look at the exponentiated terms of these log odds of significant variables, i.e. $exp(-1.5913) = 0.20365$ and $exp(-2.0101) = 0.13397$, these exponentiated terms show the odds of recovery from the coronavirus disease means that recovery odds for the patients in the age group 41–60 years is equal to 0.2036 times the recovery odds for the patients in the age group 0–20 years. Similarly, the patients aged 60 and over have 0.13397 times the odds of being recovered from Covid-19 disease compared to the patients in the age group 0–20 years on average, holding all else constant. From these two odds ratios, it can also be discovered that the odds of recovery from the corona disease is higher in the patients aged between 41–60 than the patients whose age is 60 and over. From Table 2, it can be assured that for the patients in the age group 0–20 and who are male, the probability of recovery from coronavirus disease is 0.6597 and the probability of recovery for the male patients aged between 41–60 is 0.6818. Also, the predicted recovery probability from coronavirus disease of patients aged 60 and over is 0.6746, which is slightly lower than the patients aged between 41–60 and higher than the patients of aged between 0–20. But on average, it can also be seen that the probability of recovery from coronavirus disease during the first wave of pandemic is almost same in all the patients and lies between the probability 0.6597–0.6818. If we look at the coefficient of gender (male) in Table 2, which is also statistically insignificant, it means there is no strong evidence for a gender difference in risk of dying due to coronavirus disease. This implies that the probability of recovery from coronavirus disease is same in males and females, keeping all else constant.

To test the goodness of fit of the model to the data, the log likelihood ratio R^2, sometimes called McFadden R-squared, the C-Statistic (Concordance Statistic)

and Chi-Square goodness of fit test, has been used. The McFadden R-square is defined as:

$$R^2_{MF} = 1 - \frac{LL_{full}}{LL_0} \tag{8}$$

where LL_{full} is the full log likelihood model and LL_0 is the log likelihood function of the model with the intercept only. Backhaus et al. (2000) suggested that a McFadden R^2 value is in the range 0.2–0.4 indicates a good fit of the model and the obtained value of the R^2_{MF} is 1-384.12/482.96= 0.20465463 and shows the model is sufficiently well fitted to the data and the C-statistics can be computed by considering all possible pairs consisting of one patients who recovered from the coronavirus disease and one patients who deceased. The obtained C-statistics is the proportion of such pairs in which the patients who experienced a recovery from coronavirus disease had a higher estimated probability of experiencing the recovery than the patients who did not experience the recovery from the coronavirus disease. The value of C-statistics can lie between 0.50 to 1.00 The closer the C-statistic is to 1, the better a model is able to classify outcomes correctly. The value of C-statistics between 0.70 and 0.80 signals the model is good fitted to the data and the value between 0.50 to 0.70 indicates poor models (Hosmer & Lemeshow, 2000). Here, the obtained C-statistic is 0.7599994, which also indicates that the model is good enough and is able to classify outcomes correctly.

The Chi-square goodness of fit test is also used to test the goodness of fit of the model. For this, the standardized residuals are calculated as

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i(1-\hat{y}_i)}}$$

And then the Chi-squared statistics is obtained as

$$\chi^2 = \sum_{i=1}^{n} r_i^2$$

The χ^2 statistics follows a χ^2 distribution with n-(p+1) degree of freedom, where p are the number of covariates. The obtained χ^2 value is 427.228 with 422 degree of freedom and the corresponding p-values is 0.4199021. This indicates that we cannot reject the null hypothesis that the model is exactly correct and it shows that the model fits the data well. From Figures (1 and 2), it can also be seen that the observed and expected number of cases of recovered and deceased is almost same, which also indicates that the model fits the data well.

4. Conclusion

The coronavirus has wreaked havoc all across the world with the rising cases of COVID-19 every day and with the absence of any effective treatment. In these gravedigger circumstances, the Government of India adopted many preventive steps such as lockdown, social distancing and urging people to live with extra cleanliness and India benefited somewhat from the strict lockdown but this nationwide lockdown cannot be continued for so long as it is not the solution for this pandemic, and it also not good for the country's economy. Hence, it is necessary to estimate the probability of recovery from the coronavirus disease as most of the Indian population is living in poor hygienic conditions. In this article, a probability model is developed using the indirect method of estimation based on some demographic factors, and it is found that the probability of recovery from coronavirus disease is statistically same in both males and females. Also, the coronavirus patients in the age group 0–40 years have almost equal probability of being recovered from this disease. In the patients aged between 41–60, the odds of recovery from the coronavirus disease is equal to 0.2036 times the recovery odds of the patients of the age group 0–20 years, while the patients aged 60 and over have 0.13397 times the odds of recovery from coronavirus compared to the patients of the age group 0-20 years on average. Also, the odds of recovery from coronavirus is higher in the patients of the age group 41–60 years than in the patients aged 60 and over.

References

Anastassopoulou, C., Russo, L., Tsakris, A. and Siettos, C., (2020). Data-based analysis, modelling and forecasting of the COVID-19 outbreak, *PLOS ONE*, 15(3), e0230405. https://doi.org/10.1371/journal.pone.0230405.

Backhaus, K., Erichson, B., Plinke, W. and Weiber, R., (2000). *Multivariate analysenmethoden*, Berlin: Springer.

Chen, Yi C., Lu, P. E., Chang, C. S. and Liu, T. H., (2020). *A Time-dependent SIR model for COVID-19 with Undetectable Infected Persons*, http://gibbs1.ee.nthu.edu.tw/A Time Dependent SIR Model For Covid 19.pdf.

Chatterjee, S. and Hadi, Ali S., (2006) Regression analysis by example. *John Wiley* & *Sons*, Inc., Hoboken, New Jersey.

Corman, V. M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D. K., Bleicker, T., Brunink, S., Schneider, J. and Schmidt, M. L., (2020). *Detection of 2019 novel coronavirus (2019-ncov) by realtime rt-pcr, Euro surveillance*, 25(3), 2000045.

Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubã, Z., Cuomo-Dannenburg, G., Dighe, A. Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L. C., Elsland, S. V., Thompson, H., Verity, R., Volz, E., Wang H., Wang, Y., Walker, P. Gt., Walters, C., Winskill, P., Whittaker, C., Donnelly, C. A., Riley, S. and Ghani, A. C., (2020). *Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand*, Imperial College COVID-19 Response Team.

Gamero, J., Tamayo, J. A. and Martinez-Roman J. A., (2020). *Forecast of the evolution of the contagious disease caused by novel corona virus (2019-ncov) in China*, arXiv preprint ar Xiv: 2002, 04739.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J. and Gu, X., (2020). *Clinical features of patients infected with 2019 novel coronavirus in Wuhan*, China, The Lancet, 395(10223), pp. 497–506.

Hosmer Dw, Lemeshow S., (2000). *Applied Logistic Regression (2nd Edition)*, New York, NY: John Wiley & Sons;.

Nikolay K., Anton I. and Asen R., (2020). On the half–logistic model with ”polynomial variable transfer”. Application to approximate the specific ”data corona virus”. *International Journal of Differential Equations and Applications*, 19(1), pp. 45–61.

Nikolay K., Anton I. and Asen R. (2020). On the Verhulst growth model with polynomial variable transfer. Some applications. *International Journal of Differential Equations and Applications*, 19(1), pp. 15-32.

Maksym Z., Roman O. B., Yuliia P., Maksim I. and Irakli P., (2020). Mathematical model building for COVID-19 diseases data in European Countries. *IDDM'2020: 3rd International Conference on Informatics & Data-Driven Medicine*, November 19–21, 2020, Växjö, Sweden, Session 1: Artificial intelligence, CEUR Workshop Proceedings.

Rao Srinivasa A., S. R., Krantz S., G., Kurien T. and Bhat R., (2020). Model based retrospective estimates for COVID-19 or coronavirus in India: continued efforts required to contain the virus spread. *Current Science*, 118(7), pp. 1023-1025.

R Core Team, (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria, URL https://www.R-project.org/.

Rothe, C., Schunk, M., Sothmann, P., Bretzel, G., Froeschl, G., Wallrauch, C., Zimmer, T., Thiel, V., Janke C. and Guggemos, W., (2020). Transmission of 2019-ncov infection from an asymptomatic contact in Germany. *New England Journal of Medicine*, 382(10), pp. 970-971.

Singh, B. P., Singh, G., (2020). *Modeling Tempo of COVID-19 Pandemic in India and Significance of Lockdown*, https://doi.org/10.1002/pa.2257.

Singh, B. P., (2020). *Forecasting Novel Corona Positive Cases in India using Truncated Information: A Mathematical Approach*, medRxiv preprint, doi: https://doi.org/10.1101/2020.04.29.20085175.

Singh, R., Adhikari, R., (2020). *Age-structured impact of social distancing on the COVID-19 epidemic in India*, arXiv: 2003, 12055.

Singhal, T., (2020). A review of coronavirus disease-2019 (COVID-19). *The Indian Journal of Pediatrics*, pp. 1-6.

World Health Organization, (2020). *Updated WHO advice for international traffic in relation to the outbreak of the novel coronavirus 2019-nCoV*, Available at: https://www.who.int/ith/2020-24-01-outbreak-of-Pneumonia-caused-by-new-coronavirus/en/ (accessed January 2020).

World Health Organization, (2020). *Coronavirus disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update*, Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (accessed March 2020).

Appendix

Table 1. Number of patients deceased or recovered from the corona disease in India during 30 January 2020 to 30 May 2020

Age Group	Patient Status				Total
	Deceased		Recovered		
	Female	Male	Female	Male	
0-20	6	2	4	4	16
21-40	6	15	21	31	73
41-60	48	104	11	19	182
60 and over	51	87	6	12	156
Total	111	208	42	66	427

Table 2. β Coefficients showing the log odds ratios of recovery from the coronavirus disease

Deviance Residuals:				
Min	1Q	Median	3Q	Max
-1.61	-0.59	-0.51	0.8	2.1
Coefficients:				
Group	Estimate	Standard Error	z value	Pr(>|z|)
Intercept	0.0401	0.5103	0.0790	0.9373
21-40	0.9346	0.5676	1.6470	0.0996
41-60	-1.5913	0.5441	-2.9250	0.0034*
60 and over	-2.0101	0.5632	-3.5690	0.0003*
Gender(Male)	-0.1071	0.2695	-0.3970	0.6913
Null Deviance:	482.96 on 426 degree of freedom			
Residual Deviance:	384.14 on 422 degree of freedom			
AIC:	394.14			
Number of Fisher scoring iterations: 4				

The p-values denoted by * are significant at 0.05 level of significance

Table 3. Exponentiated estimated coefficients showing the odds ratios and their respective confidence intervals

Group	Estimates	95% Confidence Interval	
		Lower limit	Upper limit
Intercept	1.04	0.38	2.89
21-40	2.55	0.83	7.89
41-60	0.20	0.07	0.60
60 and over	0.13	0.04	0.41
Gender (Male)	0.90	0.53	1.53

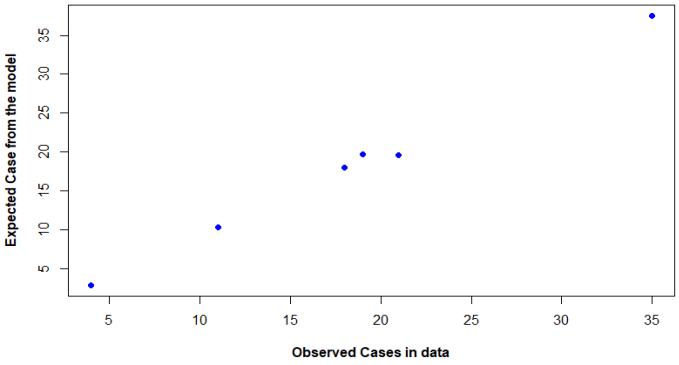

Figure 1. Observed and expected number of cases recovered from the corona disease in groups

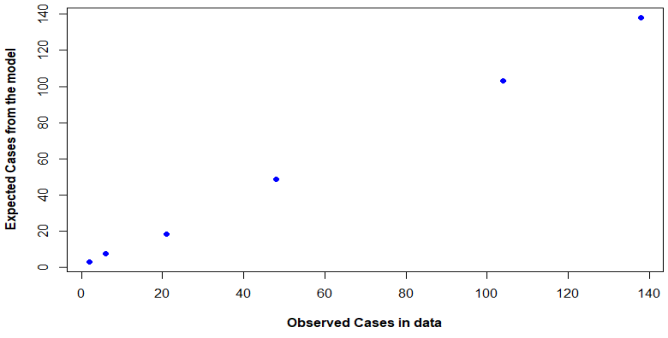

Figure 2. Observed and expected number of cases deceased from the corona disease in groups