What Affect the Death of COVID-19 in the begining of worldside outbreak

Ningjing Gao

University of oklahoma

May 4, 2020

Abstract

This article will focus on analyzing the results of COVID-19 deaths to summarize which kind of people are more likely to recover. Also, this article will review several factors that lead to the death of coronavirus patients, and how different characteristics of these patients affect the number of deaths. The dataset are from kaggle which including the age, gender, whether been to or from Wuhan, and whether been to the hospital as the independent variable. Make linear regression of the logarithm of the deaths of coronavirus confirmed cases in January and February as the dependent variable.

1 Introduction

In December of 2019 that time close to the Chinese New Year, a new coronavirus was found from the seafood wholesale market in Wuhan. The World Health Organization named it COVID-19. From the first confirmed case was found to the end of February, confirmed cases and death cases in Wuhan accounted for 61 percentage and 76.5 percentage of the total cases in China. [5]

The whole world is under the threat of COVID-19 until today. The epidemic caused by the new coronavirus has spread to more than 100 countries around the world. Also, over 2,628,527 people have been confirmed in worldwide. The United States has over 100,000 people confirmed. The number of deaths due to COVID-19 has exceeded 200,000 in worldwide. It is meaningful to observe the factors that affect death.

This time, the virus spreads from person to person, causing the number of patients to grow exponentially. [4] Patients need to be quarantined because of the virus is extremely infectious, which makes the demand for isolation beds in the hospital very huge. Even the government begin suggesting patients who do not have heavy symptoms could stay at home and recover by themselves to relieve medical pressure.

Earlier studies found that the majority of coronavirus patients are elderly. But in the later period, because of the highly infectious of COVID-19, the rate of young people getting confirmed positive also increased. We started to wonder what is the connection between death and age. [2] When faced with the threat of coronavirus, people talk more about death rate and recovery rates. But people do n't spend time paying attention to which kind of people are more likely to die. Death may be determined by certain characteristics of the patient. As the epidemic continues to ferment globally, the factors that led to the death of coronavirus patients from previous patient data may help reduce the number of deaths. [1]

2 Literature Review

The coronavirus occurred in Wuhan, China at the end of 2019. The World Health Organization proposes that the spread of coronavirus through humans causes the disease to spread more widely. [10] The upsurge of the Spring Festival travel has led thousands of people to come and from Wuhan to other cities and countries. [3]

COVID-19 experienced three outbreaks in Wuhan:

1. In the early days, most cases were found to be in contact with the South China Seafood Market in Wuhan. A local outbreak has occurred in people who have been in contact with this place.

2. Infectors who were related to the seafood market in the early days do community transmission without knowing it. Spreading infections have occurred in many communities and families in Wuhan.

3. Coronavirus was discovered before the Chinese New Year. Through the large-scale national population movement, the epidemic spread to areas outside Wuhan sharply. Even it has moved overseas including Thailand, South Korea, Japan, Singapore, and other countries have successively reported the number of confirmed COVID-19. Early cases in these Asian countries were found to have at least a little bit link to the flow of Wuhan population [7]

The coronavirus spreads from person to person, and the infected person may have symptoms such as fever, cough, fatigue, and so on. Reports indicate that older men who suffer from other diseases are more likely to be infected. [8] Most elderly patients in ICU are suffering from other diseases, this situation cause severe symptoms are more likely to occur after infection, and even death. [3]

In China, more than 1 billion people are restricted to stay at their residences for more than two months. According to statistics, China's epidemic situation continues to grow from January 10 to 22. However, by tracking cases and isolating suspected contacts in a timely manner, the number of newly confirmed persons gradually decreased after peaking on the 23rd. From the early days, 2,000 new cases were confirmed daily, then reduced to 400 new cases per day. The real decline of this data is the result of effectively curbing the spread of the coronavirus. The situation in China becomes flat. [2]

Internet technology has been used to detect infectious diseases since about 20 years ago. In early research, most of the data came from mainstream media and authoritative organizations. But now, the data source has been extended to major social media. People can do keyword searches and use online maps to track frequency and transmission routes. A comparison of influenza data in the United States with Google 's estimates, the result shows the prediction of the internet technology is close to real data. And the use of statistical Twitter data makes predictions more rapid and real-time. Internet technology provides important information for infectious disease management and decision-making. Like this time, AI technology has played a great role in the prediction and management of this epidemic. But at the same time, this prediction model that relies on the Internet also has certain drawbacks. Its sensitivity and accuracy need to be improved. Therefore, it cannot directly replace the traditional prediction model, and can only be used as an auxiliary expansion tool. [9]

3 Data

The data set in this article is mainly from kaggle which includes global patient data for January and February. We used information about 1085 new patients with COVID-19 including their age, gender, whether they have been to the hospital, whether they were from Wuhan or have been to Wuhan, and whether they died. But the data set is incomplete at the beginning, especially the variable age, and multiple fields are not available and are marked as empty. The missing data reduces the accuracy of the prediction. We fill with the median of age to fill in 242 missings. And data preprocessing can convert the original data into a format that can be easily understood. The entire data set is processed into a unified format. The dummy variables 0, 1 are referenced here. Dummy variables have been introduced in the data of whether you have been to Wuhan and whether you have been to a doctor because they contain categorical data. Obviously, gender includes men and women. Age is years, death tells the confirmed patient whether to die. 0 means alive, 1 means not alive. hospited means whether they went to the hospital,1means went, 0 means never gone.Visiting.Wuhan: 1means visited, 0means never visited. From.Wuhan: 1means yes, 0means no.

there are two reasons why use variable connected with Wuhan. Figure 1 shows the number of confirmed diagnoses in each country before March. It can be seen from the visualization in graph 1 that China had the highest confirmed number before March. It is easy to see, China, Japan, and South Korea in Asia are all countries with a large number of confirmations. Everyone knows that the first confirmed case is from Wuhan. And In the early days, coronavirus cases were linked to Wuhan or China. so it seems very important whether we have been to Wuhan or from Wuhan in our data set. We can see from Figure 2 that the death rate is highest among people around the age of 50. The number of deaths in the elderly is much higher than in young children. And we also tested that all other variables and death are Highly Predictive.

4 Empirical Method

The major method used in this article is linear regression. Logistic regression is a predictive method for regression analysis of dependent variables. The logistic regression equation is used to explain the relationship between the dependent variable and multiple independent variables. This article analyzes the relationship between age, gender, whether you have been to or from Wuhan, and whether you have been to the hospital and the death of coronavirus patients. In this case, the reliability of the model is high in the early stage, but not in the middle and late stages. The reasons are as follows:

1. There is basically no control of the disease in the early stage of transmission, but the community and medical units have strictly controlled the disease in the middle and late stages, and the transmission may be reduced.

2. The infection base is large, and some cases die or heal, reducing the number of diagnoses. The logistic model does not take into account.

3 The detection period of the kit and the amount of the kit used for detection are changing. [6]

Therefore, the logistic growth model only predicts the disease and cannot accurately judge it, and not the best model for CORVID-19. My linear regression model is as follows:

$$Log(death) = \beta_0 + \beta_1 age + \beta_2 gender + \beta_3 hospitalized + \beta_4 from.Wuhan + \beta_5 visting.Wuhan, \quad (1)$$

The subject of this study is how the characteristics of patients with early coronavirus death affect. Obviously, Patient's physical fitness is the most important factor for recover, but this cannot be measured. But the difference of physical fitness between men and women, person at different ages is true. This model excludes some variables that cannot be measured, such as the date of confirmation and the date of visiting.Wuhan. These variables are converted into dummy variables in a unified manner. Usually, people with infectious diseases can get good treatment when they go to the hospital. But this time, it is a regional or even national outbreak. As a result, the needs of medical facilities and medical personnel have exceeded the scope of social commitment. Wuhan is also a special case. The medical power from whole of China is concentrated in Wuhan. And Chinese people actively cooperate to keep social distance. Therefore, these predictions cannot be directly used in data of other countries.

In addition to linear regression, we also used a random forest algorithm to detect the accuracy of the model.

5 Research Findings

We used the lm () function when fitting a linear regression model to the data. The results of the linear regression model are logical. People from Wuhan account for an absolute proportion of COVID-19 death cases. while older age, men, and been to hospital may lead to an increased chance of death. Cases that have been to the hospital may had severe typical symptoms or become infected in the hospital. But what is surprising is that visiting Wuhan does not lead to an increased chance of death. Other external factors are also negatively related to the death of COVID-19.

$$\begin{aligned} Log(death) = & -0.168 + 0.003age + 0.032gender + 0.021hospitalized \\ & + 0.168from.Wuhan + 0.017visting.Wuhan, \end{aligned} \tag{2}$$

We can see deviance residuals from Table 1, which is a measure of model fitting. The first quartile (1Q) and third quantile (Q3) of the residuals are quite different, which means that the bell-shaped distribution is asymmetric. The age, from. Wuhan prediction ability in the model is very strong; The R value of 0.1474 is actually quite good. Considering the lack of detailed information about patients, the size of some of the errors needs attention, but it is not surprising.

when we do linear regression of independent variable and the dependent variable, this relationship is assumed to be linear. This is not correctly every time. For example, for all age values, the effect of age on death may not be constant. So far, we have only considered the individual effects of each feature on the results. What if certain characteristics have a comprehensive effect on the dependent variable? For example, age and from Wuhan both increase the impact of death, but it is reasonable to assume that their combined effects may be worse than the sum of their individual effects. In order to reflect the interaction between age and Wuhan, a formula can be written in this form:

$$age * from.Wuhan, \tag{3}$$

The linear regression equation after our improvement is:

$$\begin{aligned} Log(death) =& \beta_0 + \beta_1 age + \beta_2 gender + \beta_3 hospitalized + \beta_4 from.Wuhan \\ & + \beta_5 visting.Wuhan + \beta 6 Hospied^2 + \beta 7 age * from.Wuhan, \end{aligned} \tag{4}$$

Relative to our first model, the R-squared value increased from 0.1417 to about 0.2334. In addition, the interaction between age and Wuhan also has a certain significant effect. The elderly in Wuhan will increase the risk of death by an additional 1 percentage. Interestingly, the square term hospitalized2 is still insignificant. This may indicate that the medical burden of the hospital is heavy. Patients may not be able to get good treatment in the hospital. Or for COVID-19, the medical level is not capable of confrontation. Therefore, the death of the diagnosed patient is not affected by whether been to see a doctor.

We changed 75 percentage of the data set into the training set and 25 percentage into the test set. Use the data of whether the COVID-19 patient died or not to apply to the random forest algorithm to see how accurate the model is. We see in Figure 3 that the gender of the patient in the model is the most important, followed by from.Wuhan, gender, Hospied and visiting.Wuhan. From the results of the random forest algorithm, the prediction accuracy of the model is above 95 percentage.

Some unfriendly people appeared in the news and do racial discrimination because of the COVID-19 outbreak. In order to understand people's attitude towards COVID-19, I use twitter API to get the words related to coronavirus that people use. And I show these related information with figure 3. Surprisingly, people's positive attitude towards COVID-19 far exceeds the negative attitude. Although there are also a large number of people who still have negative or skeptical attitudes towards the epidemic. Perhaps people with a good transfer of the Chinese epidemic have an optimistic attitude.

6 Conclusion

From the above research, I found that Wuhan was experiencing a very serious epidemic. If the patient is from Wuhan, the probability of death will increase by 40 percentage. Older men with other diseases need pay more attention. In the face of unknown viruses, the hospital's treatment is not so helpful. Therefore, it is still necessary to focus on prevention. Avoiding go to crowded places, washing hands frequently, and maintaining a safe social distance. Using the twitter API, I found that most people have positive emotions about COVID-19.Then in front of the virus, people will actively cooperate to defeat COVID-19.

The coronavirus spread model is very complicated, and there are still many problems that have not been solved. It is not simply a multiple regression equation that can be understood clearly. More papers have focused on fitting various new models to predict new cases. This will make more sense.

References

[1] Milan Batista. *Estimation of the final size of the coronavirus epidemic by the logistic model*. 2020.

[2] David Baud et al. “Real estimates of mortality following COVID-19 infection”. In: *The Lancet infectious diseases* (2020).

[3] Muhammad Fawad et al. “An Updated Trend Analysis Representing the Outbreak of Novel Coronavirus (2019-nCoV) in 16 Cities of Hubei Province, China Using Logistic S-Curve Model”. In: *China Using Logistic S-Curve Model (3/6/2020)* (2020).

[4] Dmitry Ivanov. “Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case”. In: *Transportation Research Part E: Logistics and Transportation Review* 136 (2020), p. 101922.

[5] Lin Jia et al. “Prediction and analysis of Coronavirus Disease 2019”. In: *arXiv preprint arXiv:2003.05447* (2020).

[6] Wei Liu et al. “Analysis of factors associated with disease outcomes in hospitalized patients with 2019 novel coronavirus disease”. In: *Chinese medical journal* (2020).

[7] World Health Organization et al. “Coronavirus disease 2019 (COVID-19): situation report, 72”. In: (2020).

[8] Tanu Singhal. “A review of coronavirus disease-2019 (COVID-19)”. In: *The Indian Journal of Pediatrics* (2020), pp. 1–6.

[9] Tianzhi Wu et al. “Open-source analytics tools for studying the COVID-19 coronavirus outbreak”. In: *medRxiv* (2020).

[10] Hengbo Zhu, Li Wei, and Ping Niu. “The novel coronavirus outbreak in Wuhan, China”. In: *Global health research and policy* 5.1 (2020), pp. 1–3.

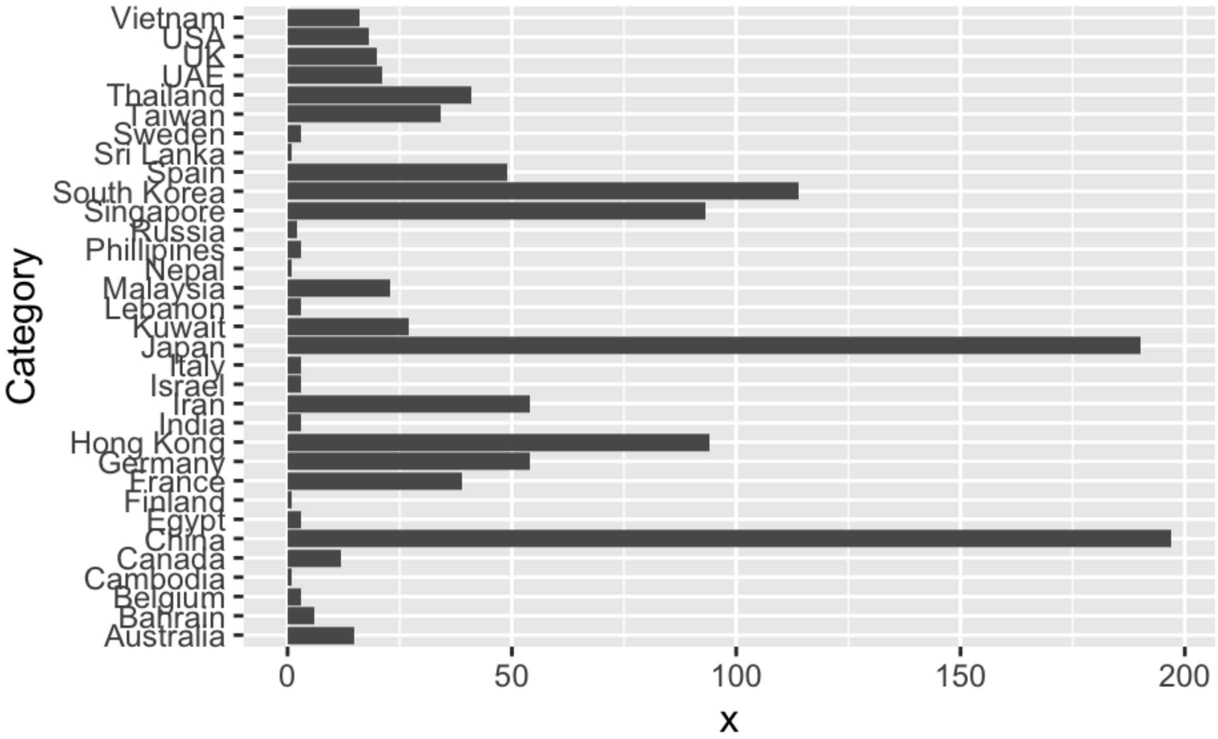

Figure 1: The Number of Diagnoses in Various Countries

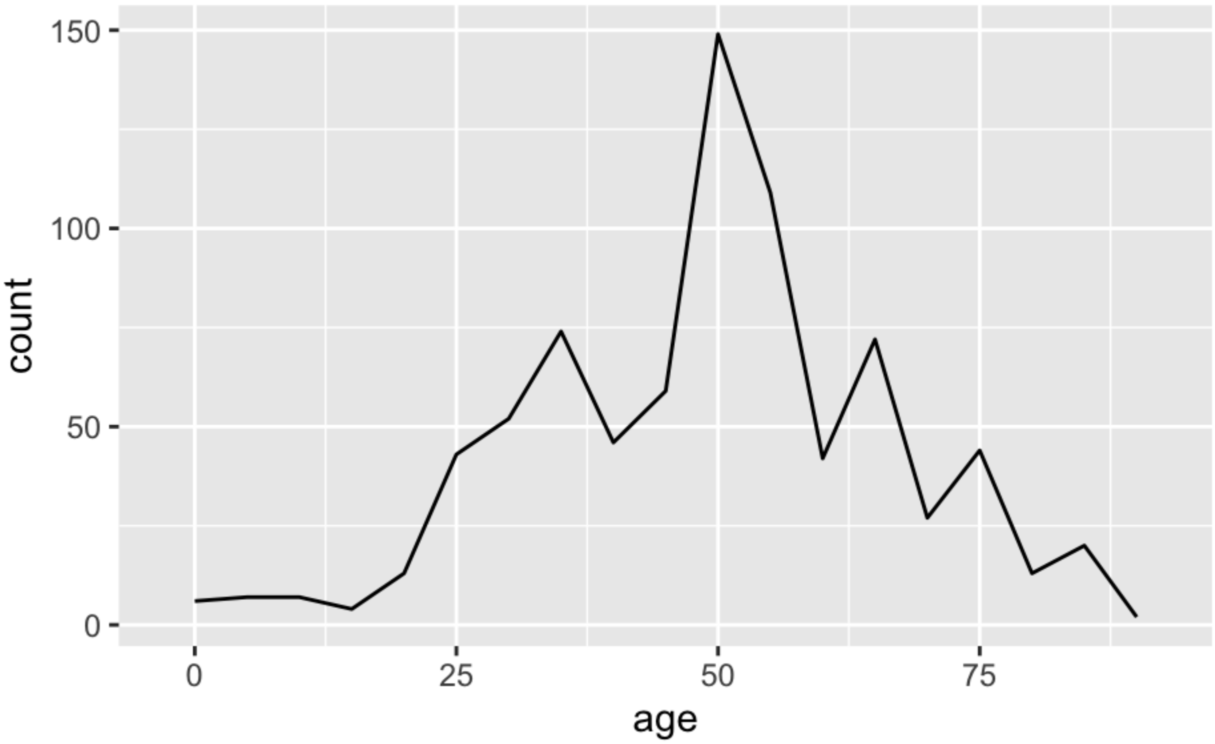

Figure 2: Age distribution of dead patients

Figures and Tables

Table 1

	Dependent variable:
	death
age	0.004***
	(0.0004)
gender1	0.032**
	(0.013)
visiting.Wuhan	−0.017
	(0.018)
from.Wuhan	0.168***
	(0.019)
Hospied	0.021
	(0.014)
Constant	−0.168***
	(0.024)
Observations	1,085
R^2	0.147
Adjusted R^2	0.143
Residual Std. Error	0.217 (df = 1079)
F Statistic	37.316*** (df = 5; 1079)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2

	Dependent variable:
	death
age	0.002^{***}
	(0.0004)
gender1	0.026^{**}
	(0.013)
visiting.Wuhan	-0.029^{*}
	(0.017)
from.Wuhan	-0.412^{***}
	(0.056)
Hospied	0.018
	(0.013)
age:from.Wuhan	0.011^{***}
	(0.001)
Constant	-0.060^{**}
	(0.025)
Observations	1,085
R^2	0.233
Adjusted R^2	0.229
Residual Std. Error	0.205 (df = 1078)
F Statistic	54.707^{***} (df = 6; 1078)
Note:	*p<0.1; **p<0.05; ***p<0.01

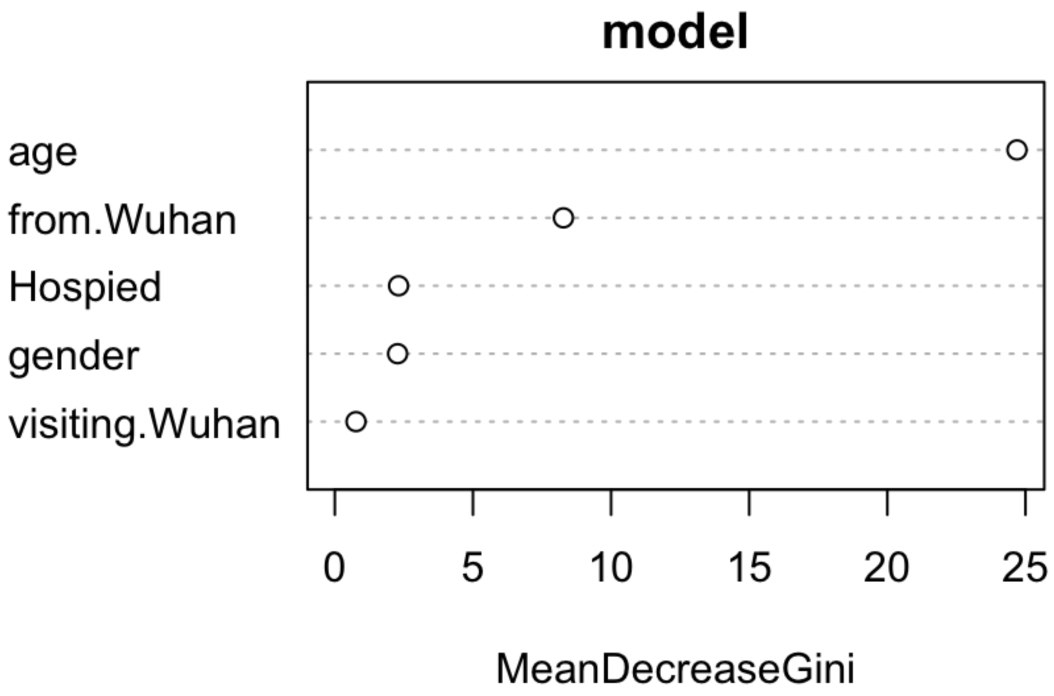

Figure 3: Variable importance map

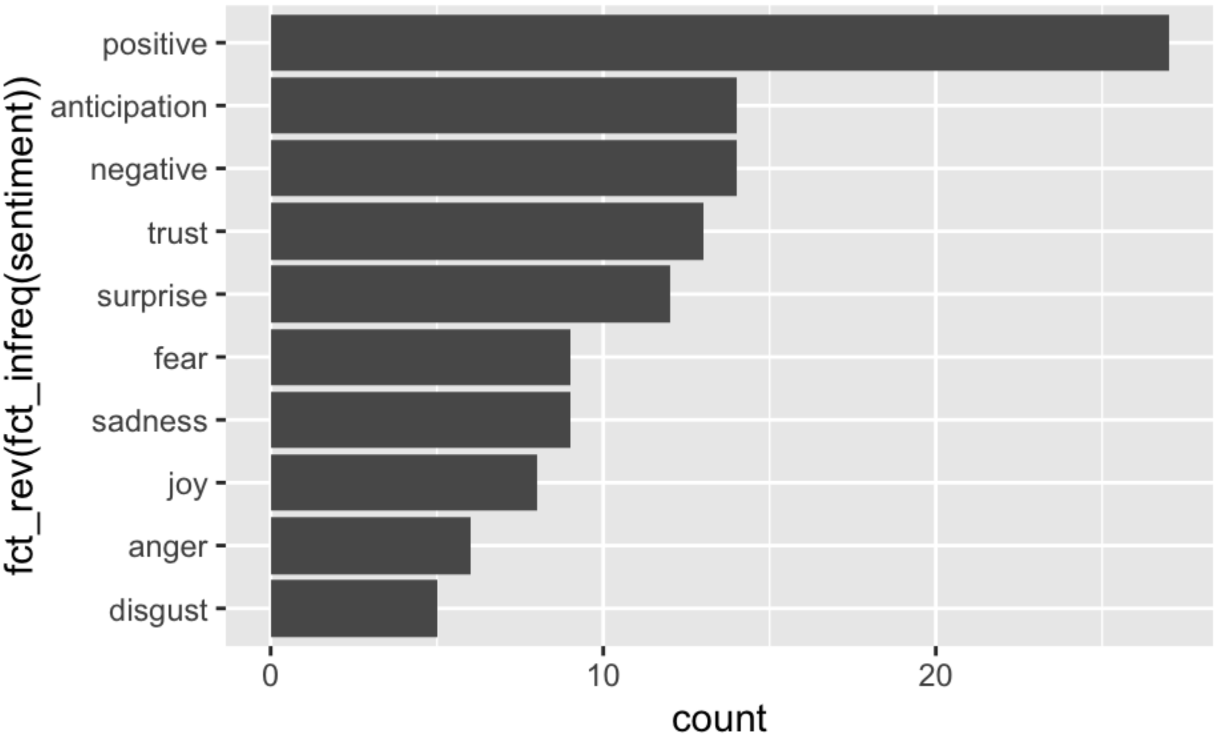

Figure 4: People’s feelings for COVID-19 on Twitter