Modelling the COVID-19 Outbreak Using a Modified Logistic Regression Model

Piotr Suder

piotr.suder@ufl.edu

Abstract

In this paper we will derive a modified logistic regression epidemic model. We will apply it then to the currently available data regarding the global COVID-19 outbreak in 2020, as of today, 4th of April 2020. We will use the model to make predictions about the number of reported cases of COVID-19 in the upcoming weeks.

The model

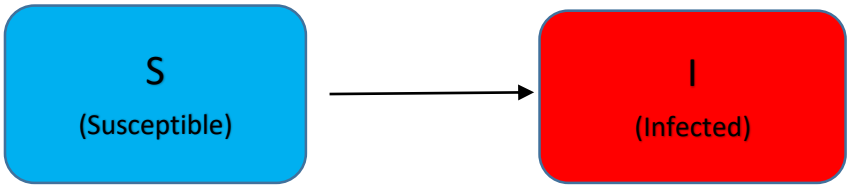
In this project I used the following Modified Logistic Regression Model:

$$I(t) = \frac{K}{1 + \exp\left(\frac{\theta}{t} + \omega t + \gamma\right)}$$

where $I(t)$ is the number of infections that were detected up to the time t, expressed in days and $K, \theta, \omega, \gamma$ are unknown constant parameters that need to be determined based on the data.

Justification of the model

We will derive the above model from a differential equation. In this model, since our goal is to predict the cumulative number of infections that will occur throughout a short period of time (a couple of weeks), we can assume that we have only two compartments: I (infected), and S (susceptible), and that the flow occurs only from S to I, as shown on the diagram below.



Notice, that in this model we are trying to predict the total number of reported infections rather than the number of people being actually infected at a given instant of time. That is why there is no outflow from the infected compartment, since once a person becomes infected this infection will be permanently recorded on the website.

The infections occur by interactions between an infected person and a susceptible person. Let $S(t)$ be the number of susceptible and $I(t)$ the number of infected people at any time t. Then the number of possible pairs consisting of a susceptible and an infected person is equal to $S(t)I(t)$. Therefore, the rate at which $I(t)$ increases at a particular instant of time is proportional to this quantity. So we can write:

$$\frac{\mathrm{d}}{\mathrm{d}t}I(t) = \beta(t)S(t)I(t)$$

where $\beta(t)$ is a factor that depends on time. $\beta(t)$ represents the fact that whenever a susceptible person meets an infected person there is some probability of an infection occurring, as well as the fact that only a portion of possible meetings between a susceptible and an infected person will take place, since it is obviously not possible for every susceptible person to meet with every infected person. We are assuming that the value of $\beta(t)$ decreases over time. The reason is that due to regulations and restrictions imposed by the government as well as increased awareness and safety measures taken by people, the proportion of the number of meetings between susceptible people and infected people that will actually take place to the number of meetings between them that could take place (given by $S(t)I(t)$) will be decreasing, as more people start practicing social distancing. Also, for the same reasons, for every meeting that does take place, the probability of an infection occurring will decrease because people for example wash hands more often.

We will give a more concrete approximation of what $\beta(t)$ might look like a bit later. For now, let us solve the differential equation without specifying it. In order to do that, we need to notice, that the susceptible and infected compartments together form the entire population, which we are assuming to be constant because of the short span of time that we are trying to capture with our model. Let therefore $N = S(t) + I(t)$ be the total population. Then $S(t) = N - I(t)$ so the equation above becomes:

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta(t)I(N - I)$$

$$\frac{1}{I(N-I)}\frac{\mathrm{d}I}{\mathrm{d}t} = \beta(t)$$

$$\int \frac{1}{I(N-I)}\mathrm{d}I = \int \beta(t)\,\mathrm{d}t$$

One can check from the integration tables that:

$$\int \frac{1}{I(N-I)}\mathrm{d}I = \frac{1}{N}\ln\left(\frac{I}{N-I}\right) + c$$

So we have

$$\frac{1}{N}\ln\left(\frac{I}{N-I}\right) = \int \beta(t)\mathrm{d}t$$

$$\ln\left(\frac{I}{N-I}\right) = N\int \beta(t)\mathrm{d}t$$

$$\frac{I}{N-I} = \exp\left(N\int \beta(t)dt\right)$$

$$I\left(1+\exp(N\int \beta(t)dt)\right) = N\exp\left(N\int \beta(t)dt\right)$$

$$I = N\frac{\exp\left(N\int \beta(t)dt\right)}{1+\exp(N\int \beta(t)dt)} = N\frac{1}{1+\exp(-N\int \beta(t)dt)}$$

Let

$$B(t) = -N\int \beta(t)dt$$

Then

$$I = \frac{N}{1+\exp(B(t))}$$

Now we need to choose a function to approximate $\beta(t)$ in order to proceed.

The assumptions we originally made about $\beta(t)$ is that it is a decreasing function. It should also be nonnegative, since we are only assuming flow from the susceptible compartment to the infected compartment, so $I(t)$ can never decrease. A function that satisfies these conditions is

$$\beta(t) = \frac{a}{t^2}$$

where a is some positive constant.

So in this case we will have

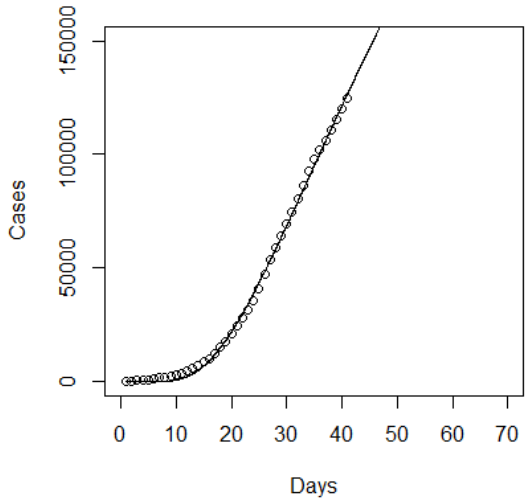
$$B(t) = -N\int \frac{a}{t^2}dt = N\left(\frac{a}{t}+c\right) = \frac{\theta}{t}+\gamma$$

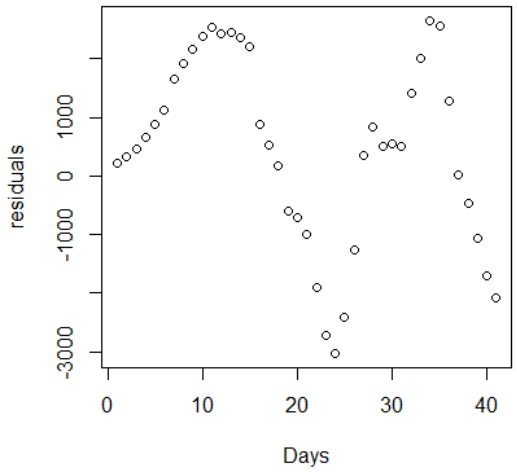
where θ and γ are some constants.

So we are getting the following model:

$$I = \frac{N}{\left(1+\exp\left(\frac{\theta}{t}+\gamma\right)\right)}$$

Unfortunately, in practice this model turns out to not fit the data well. Below is an example of fitting this model to the data for the number of infections in Italy. We can clearly see that the model does not capture the trend of the data very well and it becomes even more apparent when we plot the residuals against time. We can clearly see a pattern appearing there which suggests that the model does not provide a good fit for the data.





The problem with this model is that it does not have enough parameters that can be estimated based on the data and therefore is not well suited for capturing more complex trends in the data. So in order to improve our model we need to increase the number of unknown parameters in the model. Let us then slightly modify $\beta(t)$ simply by adding a constant to it. Let

$$\beta(t) = \frac{a}{t^2} + b$$

We can think of b as of the restriction on how much we can reduce the interactions between the susceptible and the infected due to practical reasons such as the fact that people need to do grocery and might meet an infected person in the store.

So for this modified $\beta(t)$ we have:

$$B(t) = -N\int \frac{a}{t^2} + b\, dt = N\left(\frac{a}{t} - bt + c\right) = \frac{\theta}{t} + \gamma + \omega t$$

and hence the model is:

$$I(t) = \frac{N}{1 + \exp\left(\frac{\theta}{t} + \omega t + \gamma\right)}$$

There is, however, one serious problem with this model. Since b will be positive in most cases, both in theory and in practice, ω will be negative, which means that as the value of t increases, $\left(\frac{\theta}{t} + \omega t + \gamma\right)$ will be tending to negative infinity, so $\exp\left(\frac{\theta}{t} + \omega t + \gamma\right)$ will be tending towards zero, and so in turn, $I(t)$ will be tending towards the size of the whole population N within a short period of time, which is not really supported by the data with several countries such as China, South Korea, Italy or Spain having already passed their point of maximum infections within a day and stabilizing at the number of infections being orders of magnitude smaller than the population size. This imperfection comes from the simplifications we initially made in our model, primarily the fact that we are assuming that there is no outflow from the infected compartment to either the removed compartment with people dying from the virus or recovering and gaining immunity, or back to the susceptible compartment. This simplification,

while being justified by the goal of this project and the effort to keep the model simple, turns out to be problematic, since it creates a situation when unless the value of $\beta(t)$ tends to zero, the predicted number of infections will quickly reach the size of the population. In practice this is not the case, because if an individual becomes infected he will be at risk of infecting others only for a specific period of time, before he recovers or dies, so it is enough for him to infect on average less than one susceptible person within that period and the number of infections will eventually stabilize at a lower number than the size of the population.

Therefore, in order to account for this in our model, we need to introduce one more parameter, which will rescale the predicted number of infections so that the number of infections stabilizes at a level suggested by the data rather than reach the size of the population. Thus, we will replace the size of the population N in the model by an unknown parameter K. So the model we obtain is exactly the one we introduced at the beginning:

$$I(t) = \frac{K}{1 + \exp\left(\frac{\theta}{t} + \omega t + \gamma\right)}$$

As a side note, since θ/t is not defined for $t = 0$, we will always start from $t = 1$, as for day 1.

Fitting the model to the data

Even though we are trying to predict the number of infections globally, due to recent restrictions in travel between many countries as well as cultural and economic differences between countries and the fact that different countries are at different stages of the epidemic, I have decided to try to fit the data for each country individually for as many countries as possible. Obviously, this was not possible for every single one of them, as many countries are still at early stages of the epidemic, but it turned out to be successful with around 30 countries which residents currently constitute for around 90% of cases worldwide. The benefits of such approach, in addition to the ones mentioned above, allow us to make appropriate modification to the data if necessary. For example, France recently saw a big spike in the reported cases, simply because of including the cases from nursing homes which had not been reported earlier. These modifications to the data in particular cases like this one will be discussed later.

For each of the countries we will use R tools to fit the data. In particular, we will be utilizing the nls() function, which provides least squares fit for nonlinear models such as ours. The code used for fitting and transforming the data is provided in the Appendix.

Firstly we fit the data for a big number of countries for which there are no special modifications to the data needed. Note that for every country we will only consider the data starting from the day when a given threshold number of cases was reached, due to the fact that the data from the first days after the appearance of the virus in a given country is usually rather noisy and does not provide a lot of insight. For most of the countries this threshold will be at 200 cases, but for Iran it will be placed at 20000 because there we are observing what seems to be a second wave of the outbreak which came after a period of stabilization. Thus our model would not provide a good fit for the entire data.

As mentioned before, France had its data modified to account for the spike caused by the inclusion of cases from nursing homes. The method of transforming the data used in this case is the following one. Since we know the number of cases from the nursing homes which was included on this particular day, we distribute this number proportionally over the entire period before the day of the inclusion. So, for example, let us suppose that the total number of cases which did not come from nursing homes was 80000 and 20000 cases from nursing homes was included on that day. Then we take every day before the day of the inclusion and we add to that day the number of cases from nursing home included multiplied by the ratio of the cases reported up to this particular day divided by the number of cases reported up to the day of the inclusion not coming from nursing homes. So, in our example, to a day by which we had 45000 cases, we will add $\frac{45000}{80000} \times 20000 = 11250$ cases coming from nursing homes.

We also slightly changed the model for South Korea and fitted it individually. After the rapid outbreak practically ended in this country, we are nevertheless observing a rather constant number of about 100 new cases occurring consistently every day. Therefore we used a linear function of time to model the number of cases occurring after the outbreak ended, which happened around the time that 7500 cases were reached.

The results for France and South Korea were put into the same data frame as for the other countries which were fitted individually. Here are the results:

```
> parameters_frame
        country curr. infections pred. infections       omega    gamma       theta             K
1           USA            311357     1030463.6477 -0.09183227 2.605976 31.02755673 1584108.8638
2         Spain            126168      165373.0157 -0.18561773 4.588483  8.22824307  167531.3604
3         Italy            124632      146613.9265 -0.14201664 3.931663 14.24980452  149143.8448
4          Iran             55743       80891.4419 -0.14540413 1.519411 -0.25928418   84386.9702
5       Germany             96092      139682.5428 -0.13893488 3.288858 20.90767672  145154.5198
6         China             81669       81090.6764 -0.21846914 3.949521  0.93383070   81090.6863
7            UK             41903       86366.5823 -0.19062731 5.559719  2.90173946   90082.1992
8        Turkey             23934       35042.8855 -0.29470917 4.222400  1.33292511   35153.4745
9   Switzerland             20505       25027.3211 -0.14052391 2.359180 15.58671785   25526.9730
10      Belgium             18431       25319.3084 -0.21999204 5.060192 -0.36669154   25566.8381
11  Netherlands             16627       21364.6993 -0.20035062 4.420225  0.33387370   21582.0172
12       Canada             13912       24502.6060 -0.20757424 4.317931  1.03506138   25078.3194
13       Norway              5550        7591.4901 -0.12589092 2.503342  1.01013617    7964.6033
14       Russia              4731       12682.4112 -0.24564909 4.495829 -0.41432521   13021.0129
15       Poland              3627        8267.9307 -0.16291479 3.459282  0.46309852    9023.4106
16      Ireland              4604        7257.3734 -0.17242632 2.962976  1.06467469    7502.4940
17      Czechia              4472        6126.5538 -0.18549526 2.987468  0.29403884    6232.9078
18     Malaysia              3483        4704.8394 -0.13682846 2.101748  1.05802371    4895.1988
19 Saudi Arabia              2179        3143.8246 -0.17179587 2.369883  0.53967851    3227.3911
20    Indonesia              2092        3043.4892 -0.17471922 2.934063  0.80693594    3119.9461
21      Austria             11781       12986.3135 -0.24338291 3.801242  0.43541337   13007.6662
22     Portugal             10524       14083.0157 -0.23463515 3.845222  0.63961829   14172.9108
23    Australia              5550        6078.0399 -0.26722907 3.501215  0.03369282    6084.0448
24       Israel              7851       11348.6925 -0.22289356 3.845735  0.62976432   11461.8986
25   Luxembourg              2729        3264.2548 -0.18835459 1.810852  1.33693076    3292.6842
26         Peru              1595       14281.1265 -0.13355067 6.661983 -0.10647531  149991.5926
27        Chile              4161        6555.6034 -0.20297051 3.370801  0.45297800    6685.2123
28  Philippines              3094        4793.0262 -0.26976806 4.218301 -0.98816514    4818.1321
29     Thailand              2067        2456.2673 -0.21067004 2.804808  0.91106175    2470.1371
30    Hong Kong               845        1833.4420 -0.11273302 4.108360 -1.13522809    2268.3890
31      Vietnam               240         305.6636 -0.14601300 2.217202 -0.28812015     312.6652
32    Singapore              1189        2798.4264 -0.08875503 5.459277 -1.58733847    4477.1867
33  South Korea             10156       11905.0000 -0.35500000 3.690000  0.04000000    7981.0000
34       France             10156      146286.0000 -0.15800000 4.720000  5.05000000  153000.0000
>
```

Below we also present summaries of models for a couple of different countries at different stages of the epidemic. Note that for China we utilized the same spike smoothing procedure as for France.

China

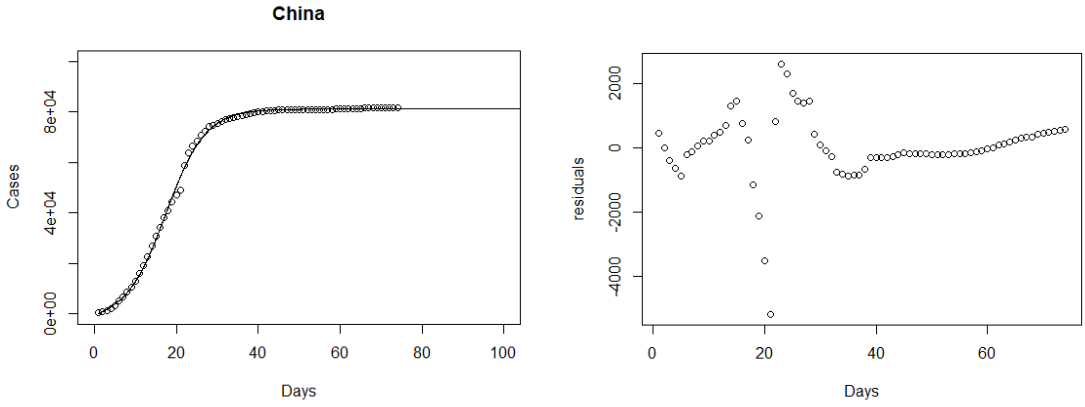
```
> summary(model)

Formula: Cases ~ K/(1 + exp(theta * (Days^-1) + (omega * Days) + gamma))

Parameters:
        Estimate Std. Error t value Pr(>|t|)
omega -2.029e-01  5.820e-03 -34.863   <2e-16 ***
gamma  3.414e+00  1.673e-01  20.405   <2e-16 ***
theta  2.925e+00  1.154e+00   2.536   0.0134 *
K      8.109e+04  1.771e+02 457.819   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1077 on 70 degrees of freedom

Number of iterations to convergence: 6
Achieved convergence tolerance: 1.235e-05
```



Italy

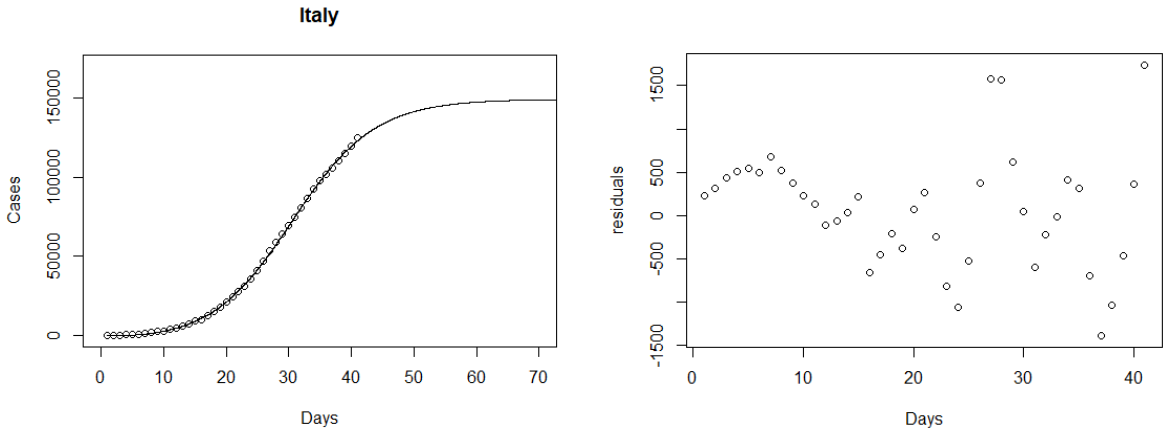
```
> summary(model)

Formula: Cases ~ K/(1 + exp(theta * (Days^-1) + (omega * Days) + gamma))

Parameters:
        Estimate Std. Error t value Pr(>|t|)
omega -1.420e-01  5.526e-03 -25.698  < 2e-16 ***
gamma  3.932e+00  2.223e-01  17.683  < 2e-16 ***
theta  1.425e+01  2.514e+00   5.668 1.77e-06 ***
K      1.491e+05  2.218e+03  67.255  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 701.8 on 37 degrees of freedom

Number of iterations to convergence: 9
Achieved convergence tolerance: 3.668e-05
```



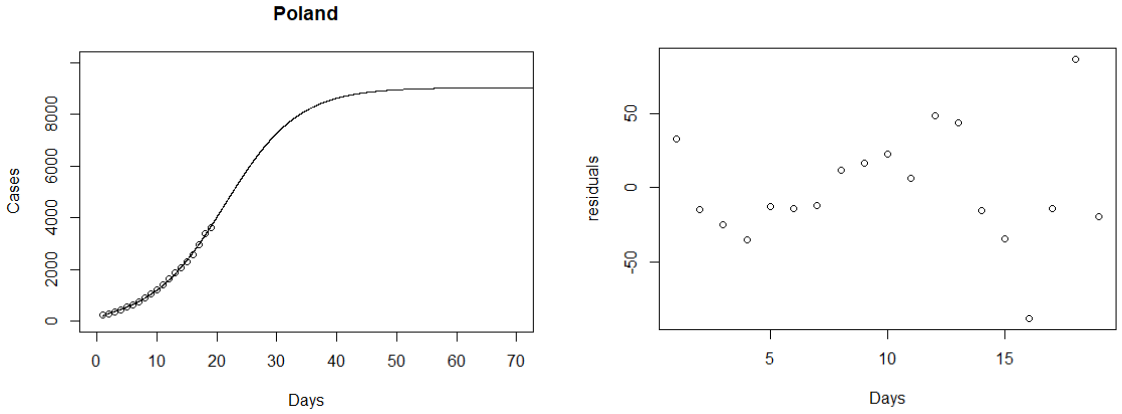
Poland

```
Formula: Cases ~ K/(1 + exp(theta * (Days^-1) + (omega * Days) + gamma))

Parameters:
        Estimate Std. Error t value Pr(>|t|)
omega -1.629e-01  9.581e-03 -17.005 3.26e-11 ***
gamma  3.459e+00  9.409e-02  36.766 4.09e-16 ***
theta  4.630e-01  2.500e-01   1.852   0.0838 .
K      9.023e+03  1.440e+03   6.267 1.51e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.89 on 15 degrees of freedom

Number of iterations to convergence: 9
Achieved convergence tolerance: 6.751e-05
```



As we can see, in most of the cases we have very low p-values for the parameters and residuals on the plots appear to not exhibit any particular patterns, at least not in the middle stage of the epidemic with has the highest number of new cases every day and thus is of the biggest interest. The only exception in this case is China, where we can see some clear patterns in the residuals, also in the middle part. This is probably at least partially caused by imperfections in the spike smoothing method used.

Rest of the world

The data was also fitted for the rest of the world, that is for the countries which turned out to be either at a too early stage of the epidemic or have too much noise in the data to successfully fit the model for them individually. Hence all of these countries were put together and treated like one country. As mentioned before, these countries in total constitute for only about 10% of cases as of today, so any inaccuracies resulting from treating them as one country should not hinder too much the accuracy of the overall prediction at least in the short term.

Here is the summary of the model:

```
> summary(model_rest)

Formula: Cases ~ K/(1 + exp(theta * (Days^-1) + (omega * Days) + gamma))

Parameters:
        Estimate Std. Error t value Pr(>|t|)
omega -1.276e-01  6.055e-03 -21.078 5.27e-12 ***
gamma  3.271e+00  1.816e-01  18.008 4.44e-11 ***
theta  1.858e-01  8.814e-02   2.108 0.053491 .
K      3.236e+05  6.929e+04   4.671 0.000361 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 662.7 on 14 degrees of freedom

Number of iterations to convergence: 9
Achieved convergence tolerance: 1.121e-05
```

and the predicted number of cases on the 21st of April.

```
> prediction_rest
[1] 248192.7
>
```

As we can see, the number of cases on the 21st of April for the rest of the world is 248192.

From the summary of the model, we can also see that almost all of the parameters have very low p-values in the t-test, which implies that for each of factors in our model them we can say with very high confidence that there is a correlation between this factor and the number of infections. The only parameter which has a relatively high p-value is θ with its p-value at around 0.053. This still, however, implies that we can say with 94% confidence level, that θ is nonzero and hence there is a correlation between $1/t$ and $I(t)$.

Conclusion

As of today, by adding the prediction for the number of cases in the countries for which we carried out individual predicitons and the prediction for the rest of the world we get that the predicted number of reported cases on the 21st of April 2020 is **2,401,724.**

```
>
> prediction <- sum(parameters_frame$`pred. infections`)
> prediction_rest <- predict(model_rest, list(Days = nrow(data) + 17), type = "response")
> total_prediction <- prediction + prediction_rest
> total_prediction
[1] 2401724
>
```