MATH35600 Assessed Practical 2020

Danielle Bueno (db17072), Yuan Chen (yc17451), Ren Cheong (rc17719)

March 28, 2020

1 Summary

The UK is currently experiencing the coronavirus (COVID-19) pandemic. Our team was assigned to analyse data on UK COVID-19 cases from 31/01/2020 to 11/3/2020 to see whether the rate of the COVID-19 spread is slowing or not. This knowledge could be useful in terms of learning the danger of this pandemic through how much it spreads. As a result, it allows the government to make decisions on the preventative next steps to tackle the pandemic to be better informed.

From the data, we constructed a model to simulate the number of daily new cases. The model which best fits the data suggests that there is some evidence that the spread would gradually slow down over time.

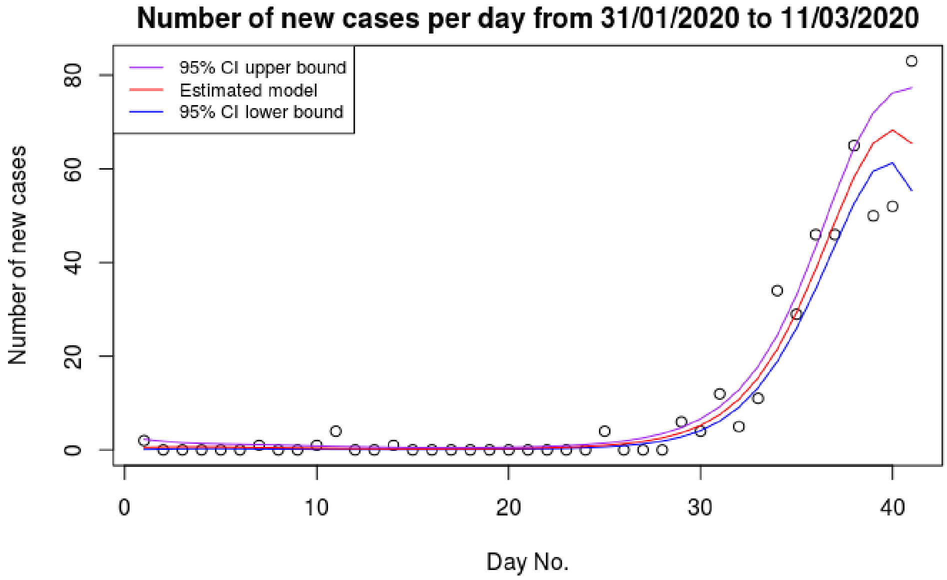

Figure 1: Graph showing model of new daily COVID-19 cases with 95% confidence intervals

However, the evidence may not be reliable mainly due to the number of data-points gathered which only spans 41 days. As the confidence interval generated by the model relies on large amounts of data to achieve accurate bounds. Therefore, any conclusions made from it cannot be sufficiently backed up without more data.

Furthermore, due to the nature of reporting, the number of new cases each day actually accounts for the number of new cases on the previous day. Looking at data that is updated daily from https://www.gov.uk/ [1], there is a notable decrease in the number of new cases reported on Mondays, which suggests that there is a decrease in the number of new cases on Sundays. However, on Tuesdays the reported number of new cases resumes the existing trend giving an overall picture of an increasing number of new cases. A possible reason for this could be that there are fewer tests being conducted on weekends as fewer health staff are working during weekends compared to that of the weekdays. This also affects our model prediction as it does not take into account the decrease in cases during the weekends.

2 Report

2.1 Model

The data for the analysis is the UK COVID-19 cases up as of 11/03/2020 published at https://www.gov.uk/ [1]. Since the number of daily new cases is countable, it can naturally be thought to follow a Poisson distribution. Moreover, we have that the expectation of a Poisson random variable is restricted to be positive, so a reasonable model to use for the observed data at time t is $y_t \sim Poisson(N(t))$ where N(t) is the following:

$$N(t) = e^{n_0 + \sum_{i=1}^{K} \beta_i t^i} \tag{1}$$

where t represents the t-th day since first occurrence of COVID-19 in the UK on 31/1/2020, $n_0 = \log N(0)$ and β_i are parameters to be estimated. In order to find the parameters which best fits the observations, we are going to choose parameter values that maximizes the likelihood of the observations.

To achieve this, we will perform numerical optimisation: Quasi-Newton Method to the negative log likelihood of our model to find a good estimate for the parameters. Note that minimising negative log likelihoods is equivalent to maximising log likelihoods. While it is difficult to find the actual Maximum Likelihood Estimate (MLE), we can get a sufficiently good estimate for most purposes that require it using optimisation. In the optimisation step, we note that since linear rescalings of time are equivalent, we would use a time vector in equal steps between $[-1, 1]$ to represent t (vs using the actual number of days) in the model to assist the performance of the optimisation function.

Here an unbounded exponential growth would then be a model with $K = 1$ (hence 2 parameters). We shall fit models with varying $K \in \{1, 2, 3, 4, 5\}$ and test which model is the best according to some criteria. By controlling K, we could change the dimension of the parameter vector, resulting in $K + 1$ parameters for the model, including n_0.

2.2 Model Selection

One approach to selecting the best model would be to choose the model with the smallest Residual Sum of Squares (RSS). Minimising the RSS is the same as minimising the differences between the observed data and the fitted values. However, RSS always decreases with more parameters, no matter how useful that variable is for prediction which can result in over-fitting.

Therefore, we have decided to select a K value by minimising the Akaike information criterion (AIC). The aim of AIC is to find a model that is the closest to the true model in the probabilistic average sense. This also allows us to test all the models simultaneously and selecting the best one.

The AIC is defined as:

$$AIC = -2l(\hat{\theta}) + 2p \tag{2}$$

where p is the number of parameters in the model, and $l(\hat{\theta})$ is the value of the log-likelihood at MLE of θ.

Compared to RSS, AIC imposes a penalty ($2p$) on the number of parameters used in the model and thus ensures that we only add useful parameters to the model.

After computing the AIC value for each $K \in \{1, 2, 3, 4, 5\}$, we see that the model with $K = 4$ (i.e. 5 parameters) has the lowest AIC of 170.38. This implies the model with $K = 4$ is the best at predicting the mean $\mathbb{E}(y_t) = N(t)$ of the model.

Model with different K values	AIC (2d.p)
K = 1	218.23
K = 2	215.05
K = 3	179.75
K = 4	170.38
K = 5	171.36

Figure 2: AIC values for models with different K values from 1 to 5

2.3 Model Checking

We now want to check that our selected model with $K = 4$ cannot be simplified further. This could be done by conducting a hypothesis test with

$$H_0 : The\ model\ with\ K = 3\ (Model\ A)\ is\ a\ good\ fit\ to\ the\ data$$

$$v.s$$

$$H_1 : The\ model\ with\ K = 4\ (Model\ B)\ is\ a\ good\ fit\ to\ the\ data$$

.

To assess the goodness of fit of two models, where Model A is nested within Model B, we could use the Generalised Likelihood Ratio Test (GLRT) with test statistic $T = 2\{l(\hat{\theta_1}) - l(\hat{\theta_0})\}$ where l is the log-likelihood function and $\hat{\theta_0}$ and $\hat{\theta_1}$ is the MLE under H_0 and H_1 respectively. We have T has an asymptotic ${\chi_r}^2$ distribution under the Null Hypothesis with $r = 1$. This is because Model A imposes 1 restriction on the parameter vector.

By calculating $\mathbb{P}(\chi_1^2 \geq T)$, we get a p-value of 0.00075 which is very small. This suggests the observed values are very unlikely to occur under the Null Hypothesis, hence we reject it in this case. Therefore, we could conclude that Model B is the best model that describes the observations which is conclusive with AIC.

2.4 Confidence Intervals

In addition to the point estimates for the model, we would also like to generate interval estimates for our selected model with $K = 4$. Since the standard deviation of $log(N(t))$ is unknown, we could provide an approximate $(1 - 2\alpha)$ confidence interval for $log(N(t))$ at each time step t using the interval:

$$\hat{y_t} \pm t_{n-p}(\alpha)\hat{\sigma}_{\hat{y_t}} \tag{3}$$

where $\hat{y_t}$ is the fitted value of $log(N(t))$ from the model, $\hat{\sigma}_{\hat{y_t}}$ is the estimate standard deviation of $\hat{y_t}$ and $t_{n-p}(\alpha)$ denote the point above which a t_{n-p} random variable lies with probability α with $n = 41$ and $p = 5$ in this case.

We first notice that $log(N(t)) = n_0 + \sum_{i=1}^{K} \beta_i t^i$ is a linear function of the powers of t. Hence taking $A = \begin{bmatrix} 0 & t & t^2 & t^3 & t^4 \end{bmatrix}$ and $\theta = \begin{bmatrix} n_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{bmatrix}^T$, we have $log(N(t)) = A\theta$.

Then, an unbiased estimator of $var(\hat{y_t})$ could be obtained as:

$$v\hat{a}r(\hat{y_t}) = v\hat{a}r(A\hat{\theta}) = Ac\hat{o}v(\hat{\theta})A^T \quad by\ linear\ transformation\ of\ covariance\ matrices \tag{4}$$

where $\hat{\theta}$ is the MLE of θ and $c\hat{o}v(\hat{\theta}) = M^{-1}$ where M is the Hessian matrix of the log likelihood function of the model evaluated at $\hat{\theta}$.

Therefore, an unbiased estimator of $\sigma_{\hat{y_t}}$ is

$$\hat{\sigma}_{\hat{y_t}} = \sqrt{v\hat{a}r(\hat{y_t})} \tag{5}$$

Using the fitted parameters from the model, we can thus calculate the approximate 95% confidence interval of $log(N(t))$.

Since we are interested in whether there is any indication that the rate of transmission is slowing over time, we would like to calculate the 95% confidence interval of $N(t)$. This gives us a range of values such that we are 95% certain it contains the true mean $N(t)$ for each time step t.

Let's consider the function $g : \mathbb{R}_{>0} \longrightarrow \mathbb{R}_{>0}$ defined as $g(\alpha) = e^{\alpha}$. Since g is increasing, bijective and continuously differentiable, a direct transformation of the confidence interval $[L(log(N(t))), U((log(N(t)))]$ is

$$[g(L(log(N(t)))), g(U(log(N(t)))] \tag{6}$$

where $L(log(N(t)))$ and $U(log(N(t)))$ is the lower bound and upper bound of $log(N(t))$ respectively.

As a result, we obtained the confidence interval of $N(t)$ seen in Figure [1].

2.5 Results

By making statistical inference of our chosen model:

$$N(t) = e^{n_0 + \sum_{i=1}^{4} \beta_i t^i} \tag{7}$$

we conclude that there is a departure from unbounded exponential growth. From the 'Estimated model' curve in Figure 1, we see that it is curving down which suggests that the number of daily new cases is decreasing and thus the rate of transmission of COVID-19 is slowing down over time. Similarly, the lower bound for 95% confidence interval suggests the rate of transmission is reduced. To contrast, we observe that the upper bound for 95% confidence interval is still increasing, but at a smaller rate. The graph is starting to change from an exponential curve to look more like a logistic curve, which suggests that the rate of transmission is starting to reach saturation. This also back up our conclusion that the rate of transmission is slowing down.

However, we note that since there are few data-points (41 days), the asymptotic properties do not necessarily hold via the Central Limit Theorem. This is problematic because we have used the Central Limit Theorem in our calculations of confidence intervals which assumes the MLEs having a normal distribution centred on the true parameter value, in the large sample limit. Therefore, we are less confident in our observed 95% confidence intervals.

Moreover, there are significant outliers in the number of cases particularly on weekends which the model does not account for. Thus, the model would be influenced by a significant proportion by these outliers which could reduce the accuracy of the model in predicting the true trend.

References

[1] Public Health England. Covid-19: track coronavirus cases. https://www.gov.uk/government/publications/covid-19-track-coronavirus-cases.