Preliminary Studies on Transmissibility of the Covid Omicron Variant

Introduction

Recently, the advent of the Omicron variant in South Africa has stirred substantial unease internationally. As a variant with numerous unexpected mutations and clinical manifestations, new discoveries are made daily and existent speculations could become obsolete. Nonetheless, to provide directions to new covid control measures, one crucial question is to estimate how transmissible the Omicron variant is compared to existent variants, which shall be the topic of this write-up. We commence with a brief recapitulation of the features of the compartmental models, and then explain how these relate to statistical methods for the estimation of Omicron's infection dynamics. Eventually, we use South Africa's covid infection data on which to apply the models, and discuss what can be said preliminarily about the Omicron variant.

Section 1. Construction of the periodic linear model

Since the advent of covid, numerous researches have used different compartmental models to study its transmission. While many researchers use 3-compartmental models such as the SIR model, attempts have made on models with 4 or more compartments in order to capture more complicated features such as reinfection and resistance. Nonetheless, as countries become more experienced and responsive to curb a potentially uncontrolled transmission, it is foreseeable that viral transmission would unlikely experience all the phases in a classical compartmental model before new policies take place. For example, upon evidence of an exponential spread for a few days or weeks, sudden and strict control measures would be instituted, upon which the exponential spread would halt in a matter of time. This means that instead of laborious attempts to model the entire process of covid dynamics, useful information could be extracted from the periods of exponential spread alone with simpler parameter estimation.

To commence, let $I(t)$ denote the number of infections at time t, modelled by the exponential spread $I(t) \cong I_0 e^{\beta t}$, where β denotes the transmission parameter, and I_0 denotes the initial number of infected individuals. The rate of detection of new infections is described by the derivative

$$I'(t) \cong \beta I_0 e^{\beta t},$$

which results in the approximation

$$R(t) \cong I(t+1) - I(t) \cong \beta I_0 e^{\beta t} \times 1 = \beta I_0 e^{\beta t},$$

where $R(t)$ denote the number of new infections in one day at time t, and is usually the most important statistics to determine the extent of viral spread. We then transform $R(t)$ to $\ln[R(t)]$, which has the form

$$\ln[R(t)] \cong \ln(\beta I_0) + \beta t.$$

Since $\ln[R(t)]$ is linearly dependent on t, it is natural to contemplate the regression model

$$\ln[R(t_i)] = \eta + \beta t_i + \varepsilon(t_i),$$

with t_i as the ith discrete time in days, η as the intercept term and $\varepsilon(t_i) \sim \mathrm{N}(0, \sigma^2)$ for all t_i, where assumptions on normality, homoscedasticity and independence need to be subsequently assessed. We note that information on I_0 is not needed, since it is fully included in the η term, whereas what we are interested in is the estimation of β.

Nonetheless, for almost all countries, a basic examination of the daily covid data would reveal that even in episodes of exponential spread, there is a clear and consistent oscillation pattern on $\ln[R(t_i)]$ with a period of one week, where the extent of this oscillation needs to be modelled as well. Such oscillation can be mechanistically explained by the fact that people's activity level varies over different weekdays, and covid detection routines may differ from weekdays to weekends. This means that the complete model should also incorporate sinusoidal functions of t with periodicity 7, which eventually has the form

$$\ln[R(t_i)] = \eta + \beta t_i + \omega_1 \sin(2\pi t_i/7) + \omega_2 \cos(2\pi t_i/7) + \varepsilon(t_i).$$

We note that in the model, we only include two sinusoidal functions, instead of addition of many sine and cosine functions of different offsets over t_i. This is because sine and cosine functions with small offsets tend to be heavily correlated, which introduces multicollinearity issues that are problematic for parameter estimation. Instead, since $\sin(2\pi t_i/7)$ and $\cos(2\pi t_i/7)$ are translations of each other by 3.5 days, different values of ω_1 and ω_2 would be adequate to account for different offsets in oscillations for different time durations. The parameter of concern, β, would naturally be estimated by the OLS method, which we denote as $\hat{\beta}_{OLS}$.

Section 2. Hypothesis tests and model diagnostics

2.1. Tests for transmission rate over reference values

Upon computation of $\hat{\beta}_{OLS}$, hypothesis tests could be performed for inference. Generally, for exponential spread, the alternative hypothesis for $\beta > 0$ is almost unequivocal, and it is more fruitful to compare $\hat{\beta}_{OLS}$ with other hypothesized values. Suppose that we want to test if β exceed some constant b, we have the hypotheses $H_0: \beta \leq b$ and $H_A: \beta > b$ with T-statistic

$$T = \frac{\hat{\beta}_{OLS} - b}{SE[\hat{\beta}_{OLS}]} \sim t_{n-4}$$

where n is the sample size. An alternative approach is to identify the maximum rate b_{max}, under which the alternative hypothesis holds at error level α, which is computed as

$$b_{max,\alpha} = \sup\left\{b \in [0, \hat{\beta}_{OLS}) : \frac{\hat{\beta}_{OLS} - b}{SE[\hat{\beta}_{OLS}]} > t_{\alpha,n-4}\right\},$$

where $t_{\alpha,n-4}$ is the upper αth quantile of the t-distribution with $n-4$ d.f.. This c_{max} then represents the maximum level which we could confidently claim β to be above. The $1-\alpha$ CI is simply constructed as

$$CI_{1-\alpha}\{\beta\} = \left(\hat{\beta}_{OLS} - t_{\alpha/2,n-4} SE[\hat{\beta}_{OLS}], \hat{\beta}_{OLS} + t_{\alpha/2,n-4} SE[\hat{\beta}_{OLS}]\right).$$

2.2. Comparative tests for transmission rates of different periods

Another important type of test concerns the comparative transmission rates of different periods. Suppose that for the same dataset, we identify two distinct periods of exponential spread, with parameter estimates $\hat{\beta}_{OLS,1}$ and $\hat{\beta}_{OLS,2}$, where we wish to assess whether one value is k times more than the other. In such cases, the hypotheses become $H_0: \hat{\beta}_{OLS,1} \leq k\hat{\beta}_{OLS,2} \Leftrightarrow \hat{\beta}_{OLS,1} - k\hat{\beta}_{OLS,2} \leq 0$ and $H_A: \hat{\beta}_{OLS,1} > k\hat{\beta}_{OLS,2} \Leftrightarrow \hat{\beta}_{OLS,1} - k\hat{\beta}_{OLS,2} > 0$. Since time durations for two different exponential spreads never overlap, the errors that arise from parameter estimates can be viewed as independent, where we have the test statistic

$$z = \frac{\hat{\beta}_{OLS,1} - k\hat{\beta}_{OLS,2}}{\sqrt{\widehat{\mathbb{V}}[\hat{\beta}_{OLS,1}] + k^2\widehat{\mathbb{V}}[\hat{\beta}_{OLS,1}]}} \sim \mathrm{N}(0,1).$$

Similarly, we can compute a maximum multiplicative factor k_{max} under which the alternative hypothesis holds, where we have

$$k_{max,\alpha} = \sup\left\{k \in \left[1, \frac{\hat{\beta}_{OLS,1}}{\hat{\beta}_{OLS,2}}\right) : \frac{\hat{\beta}_{OLS,1} - k\hat{\beta}_{OLS,2}}{\sqrt{\widehat{\mathbb{V}}[\hat{\beta}_{OLS,1}] + k^2\widehat{\mathbb{V}}[\hat{\beta}_{OLS,1}]}} > z_\alpha\right\},$$

where z_α is the upper αth quantile of the standard normal distribution. Clearly, the test on whether one value exceeds another is simply a special case of the above-mentioned test with $k = 1$.

2.3. Model diagnostics

As is applicable to all regression models, the periodic linear model should be tested for the assumptions on normality, homoscedasticity and independence of error terms. In the current case, we shall use the Shapiro-Wilk's test to assess the normality of residuals. Assessments on homoscedasticity is done via routine inspection of the residual plot, as is the case for independence of error terms, where special focus should be made to assess whether inclusion of sine and cosine functions in the model effectively removes periodicities in the data.

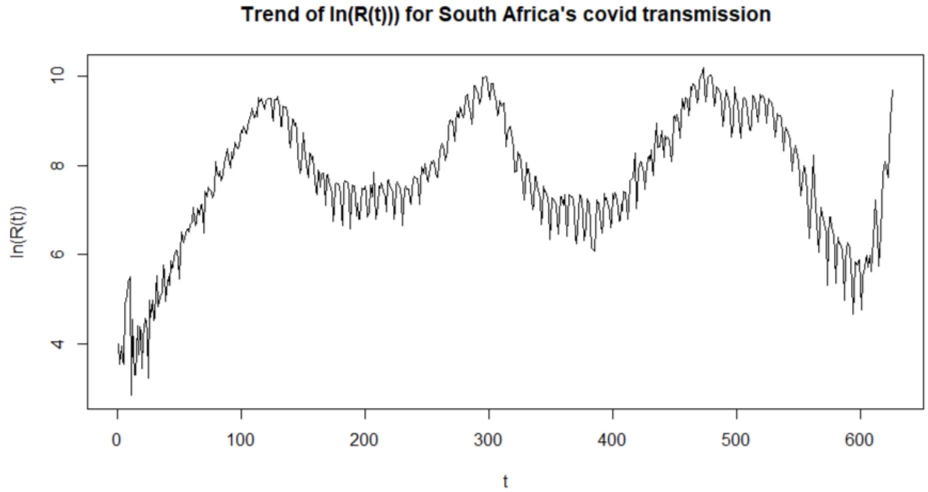
Section 3. Case study on South Africa's covid infection data

This section includes the application of the previous discussions in the evaluation of South Africa's covid infection data. We shall commence with the construction and evaluation of periodic linear models for previous episodes of exponential spread and for the most recent Omicron variant. Subsequently, we shall compare the characteristics of transmission of the Omicron variant with previous covid spreads, and interpret what we can say about the transmissibility of this new variant.

3.1. Description of transmission data

The comprehensive dataset for daily covid updates of all countries is downloaded from Our World in Data website, with source included in the references. For South Africa, the dataset includes covid updates from 2020-02-07 to 2021-12-04, of which we extract the historical records for newly detected covid cases for analysis. Unfortunately, the 1st 40 elements are either empty or have many zeros that are not amenable for analysis, and the last element is zero as well, so that the actual date available for analysis commences from 2020-03-18 to 2021-12-03, which amounts to a duration of 626 days. Subsequently, there are 6 isolated zero entries for $R(t_i)$ with problematic $\ln[R(t_i)]$, such that we shall interpolate their values with data from the days immediately before and after the zero record. The preprocessed historical trend for newly detected covid cases is displayed in Plot 3.1.1.

Plot 3.1.1: Historical covid infection trend in South Africa from 2020-03-18 to 2021-12-03



Visually, we could approximately locate four episodes of exponential spread, where $\ln[R(t_i)]$ increases linearly with t. Three episodes correspond to the traditional variant and the Delta variant, whereas the most recent exponential spread corresponds to the new Omicron variant. A weekly oscillation pattern is consistent over the entire duration of covid infection, which indicates that the periodic linear model with sinusoidal functions is appropriate. Table 3.1.1 provides a detailed breakdown of the four episodes of interest.

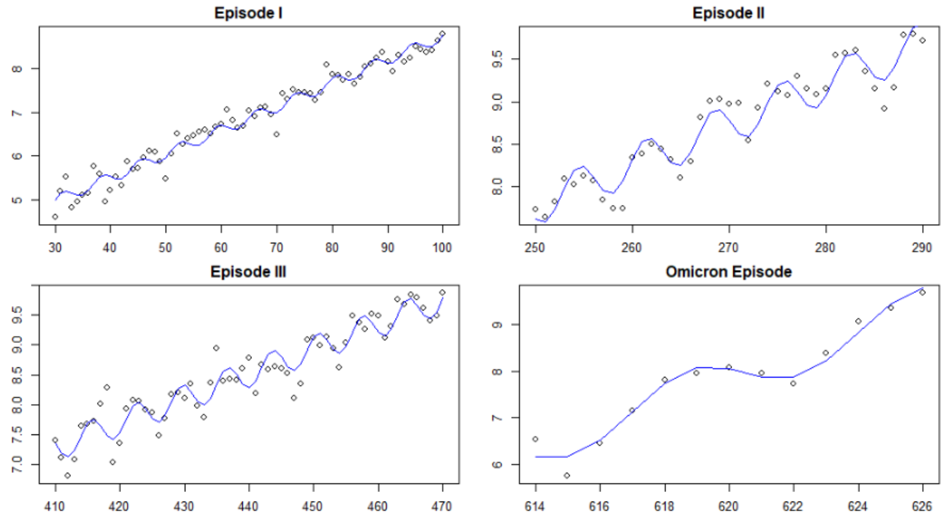
Table 3.1.1: Approximate time windows of exponential spreads

Denomination	Start Date	End Date	Duration
Episode I	2020-04-16	2020-06-25	71
Episode II	2020-11-22	2021-01-01	51
Episode III	2021-05-01	2021-06-30	31
Omicron Episode	2021-11-21	2021-12-03*	13
*: This is only the end of data, but not the end of actual spread for the Omicron variant.			

3.2. Estimation of transmission parameters

The periodic linear model is then computed for the four episodes of interest, where actual data and model predictions are displayed in Plot 3.2.1.

Plot 3.2.1: Actual transmission data and model predictions over time for different episodes



Parameter estimates and their statistical characteristics are displayed in Table 3.2.1.

Table 3.2.1: Estimated transmission parameters and statistical characteristics

Episode	$\hat{\beta}_{OLS}$	$SE[\hat{\beta}_{OLS}]$	$b_{max,0.05}$	$CI_{0.95}\{\beta\}$
Episode I	0.0541	0.00122	0.0520	(0.0516, 0.0565)
Episode II	0.0479	0.00223	0.0440	(0.0433, 0.0524)
Episode III	0.0415	0.00174	0.0386	(0.0381, 0.0450)
Omicron Episode	0.2445	0.01818	0.2112	(0.2034, 0.2857)

Generally, for Episodes I, II and III, estimates for $\hat{\beta}_{OLS}$ have small standard errors, whereas the estimate for the recent Omicron variant has relatively more standard error, which is natural since we do not have many data points for estimation. Obviously, the estimated transmission rate for Omicron is several times that of all previous episodes, which raises the question on the extent that Omicron could be more transmissible than the other variants. To proceed, we conduct the hypothesis tests for $H_0: \hat{\beta}_{OLS,1} \leq k\hat{\beta}_{OLS,2}$ and $H_A: \hat{\beta}_{OLS,1} > k\hat{\beta}_{OLS,2}$, and compute $k_{max,\alpha}$, which results in

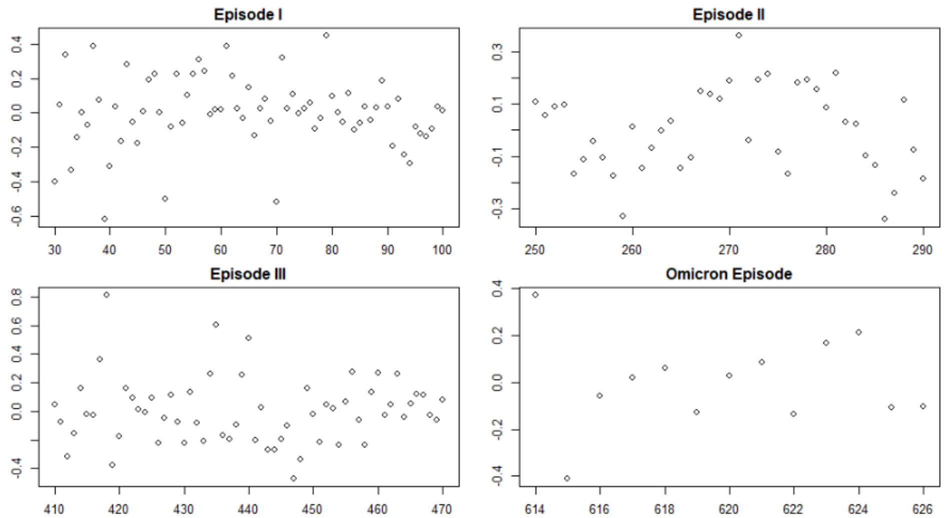
$$k_{max,0.05} = \sup\left\{k \in \left[1, \frac{0.2445}{0.0415}\right) : \frac{0.2445 - 0.0415k}{\sqrt{0.1818^2 + 0.0174^2 k^2}} > z_\alpha\right\} = 5.08,$$

which means that the recent transmission of the Omicron variant is probably 5 times faster than the previous episode.

3.3. Assessment of model assumptions

Subsequently, model assumptions are evaluated with visual examination of the residual plots, as well as the Shapiro-Wilk's test as the formal test for normality. The residual plots are displayed in Plot 3.3.1.

Plot 3.3.1: Residual plots over time for different episodes



For independence of errors, Episode I, Episode III and the Omicron Episode have residuals that are randomly distributed about zero without any temporal patterns, which indicates that sine and cosine functions in the model have completely removed all periodic patterns in errors. For Episode II, the residuals do not show any weekly oscillations, but do demonstrate ambivalent evidence of some temporal relationship.

For error variance, all plots do not show any systematic increase or decrease of variance over time, which indicates that the homoscedasticity assumption approximately holds. However, Episode I and Episode III each has three potential outliers at the lower and upper side of the residual plot respectively. Since sample sizes are small, we are not able to conclusively state whether these are true outliers or contributors to potential heteroscedasticity.

Finally, Shapiro-Wilk's tests conducted on the four episodes produce p-values of 0.0267, 0.5990, 0.0167 and 0.8749 respectively, which indicates that error terms for Episodes I and II may have deviations from normality. Such deviations may have been caused by the presence of the above-mentioned outliers, which contributes to overdispersion and skewness. While normality appears to hold for Omicron Episode for now, more data is needed to verify this conclusion, because Shapiro-Wilk's test tends to be lenient on small sample sizes.

3.4. Interpretation and discussion

Based on the results obtained, we can see that the periodic linear model is a useful approach to model the transmission of covid infection in the exponential phases, with consideration on weekly oscillations in transmission. Nonetheless, residual plots and the Shapiro-Wilk's tests show that model assumptions on normality, homoscedasticity and independence of errors do not hold universally for all episodes, such that each exponential episode should be interpreted on its own. Despite the possibility to apply more complicated transformations or covariates, we hold that the current model is still useful for analysis, because the linear dependence of $\ln[R(t_i)]$ on t, as well as the presence of weekly oscillations, have clear mechanistic explanations in the real world, whereas excessive sophistication may undermine model interpretability. In South Africa's case, current evidence indicates that the most recent episode on the Omicron variant has a 5 times faster transmission rate compared to last exponential spread at 5% error level.

Nonetheless, it is important to realize that an estimated $\hat{\beta}_{OLS}$ is not useful on its own, but must always be viewed comparatively with other episodes. This is because $\hat{\beta}_{OLS}$ does not measure the intrinsic transmission rate of covid variants, but rather captures the compounded result from the interplay of intrinsic viral characteristics, local population characteristics, as well as the intensities of covid control measures issued. Instead, comparison of $\hat{\beta}_{OLS}$ estimates from different covid episodes in the same country provide a clue on the relative rate of covid transmission controlled for the relative intensity of covid control measures, which is of value for policy decisions. For example, the fact that the most recent Omicron case is 5 times transmissible implies that covid control measures must be stricter to reduce individuals' covid exposure by 5 times as well, in order to suppress the transmission rate to historical levels. Alternatively, other countries that pursue the same policy trend as South Africa in the recent months should also revise their current covid measures to implement stricter controls, or else these countries would experience a 5 times transmission rate in the foreseeable future and suffer substantial socioeconomic and medical disruptions.

Conclusion

This small write-up has provided the basic tools for a periodic linear model to describe the exponential phase of covid transmission, and applied them to the analysis of the recent Omicron variant that surfaced in South Africa. Our preliminary results indicate a 5 times transmission rate for the Omicron variant, which is of much concern. While the estimations themselves are not direct descriptors for Omicron's intrinsic transmission rate, they still offer useful clues on the prospects of covid infection relative to the intensities of covid control measures. Since presence of the Omicron variant has already been detected in many countries, the current evidence indicates an immanent need to implement much stricter covid control measures, so that the same exponential spread in South Africa would not be repeated.

References

1. South Africa: Coronavirus Pandemic Country Profile. Data accessible and downloadable from: https://ourworldindata.org/coronavirus/country/south-africa?country=~ZAF.