Analysis of initial growth rate of COVID-19 cases in Indiana

Arny Dagsdottir, Steven Donovan, Aaron Smith, Brittany Williams, Isaac Wyatt

Department of Mathematical Sciences
Purdue University - Fort Wayne

Introduction

Towards the end of the year 2019, the world began to change massively due to the COVID-19 virus starting to break out of China (Yang, 2022). Early 2020 saw the United States begin to react by shutting down schools, enacting lockdown protocols, and classifying workers as essential or nonessential. (Decker, 2020). Once the virus started to spread, Indianapolis, one of the largest cities in the nation and located in Marion County, developed a reputation as one of the epicenters of COVID-19. King, 2021. Meanwhile, although Allen County certainly dealt with the virus, things seemed to be more or less under control in that part of the state. (ACDH, 2022). Allen County is made up of Fort Wayne and surrounding area, including a lot of rural areas and small towns, while Marion County consists mainly of Indianapolis. (Indiana.Gov, 2022). According to US Census Bureau, as of July 2021, the estimated population of Marion County was 971,102, and the estimated population of Allen County was 385,410. US Census, 2022). Due to the population difference between the two counties, we will have to consider percents in our research rather than raw numbers. We want to know if the rate of infection for Marion County was higher than the rate of infection for Allen County. We will transform data to achieve linearity and perform least squares regression to determine the infection rates for both counties.

Methods

Our team will look at rate of growth based on positive tests for COVID-19 from January 2020 through October 8, 2022, using time in daily increments as the explanatory variable and rolling 7-day infection case rate percent as the response variable. At the beginning of a pandemic, infectious disease growth is known to be exponential and so it can be modeled by $N(t) = N_0\, e^{\beta\, t}$, where $N_0 = N(0)$ is the number of cases at time $t = 0$. All the terms in this model are positive, so we can apply a log transformation and obtain

$$\log N(t) = logN_0 + \beta t.$$

If we rename the variables and assume that there is an error component to satisfy the E's hypothesis in Frees (2009), the relationship between log of the number of cases N and the time t satisfies a linear regression model:

$$y_t = \beta_0 + \beta_1 t + {}_t\,.$$

We will use least square method to obtain this equation, which is "a statistical procedure to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve" (Investopedia, 2022). This formula helps us understand and find the dependent variable and by knowledge of the population we would be able to estimate how much infection rates have improved (or gotten worse). So for our research we would be able to find out the increase in positive Covid tests by using the Least squares Method formula, and we should be able to find out how the disease grew over time and how it will increase (or hopefully decrease) over the next years.

Please see our paper for full reference citations.

Data

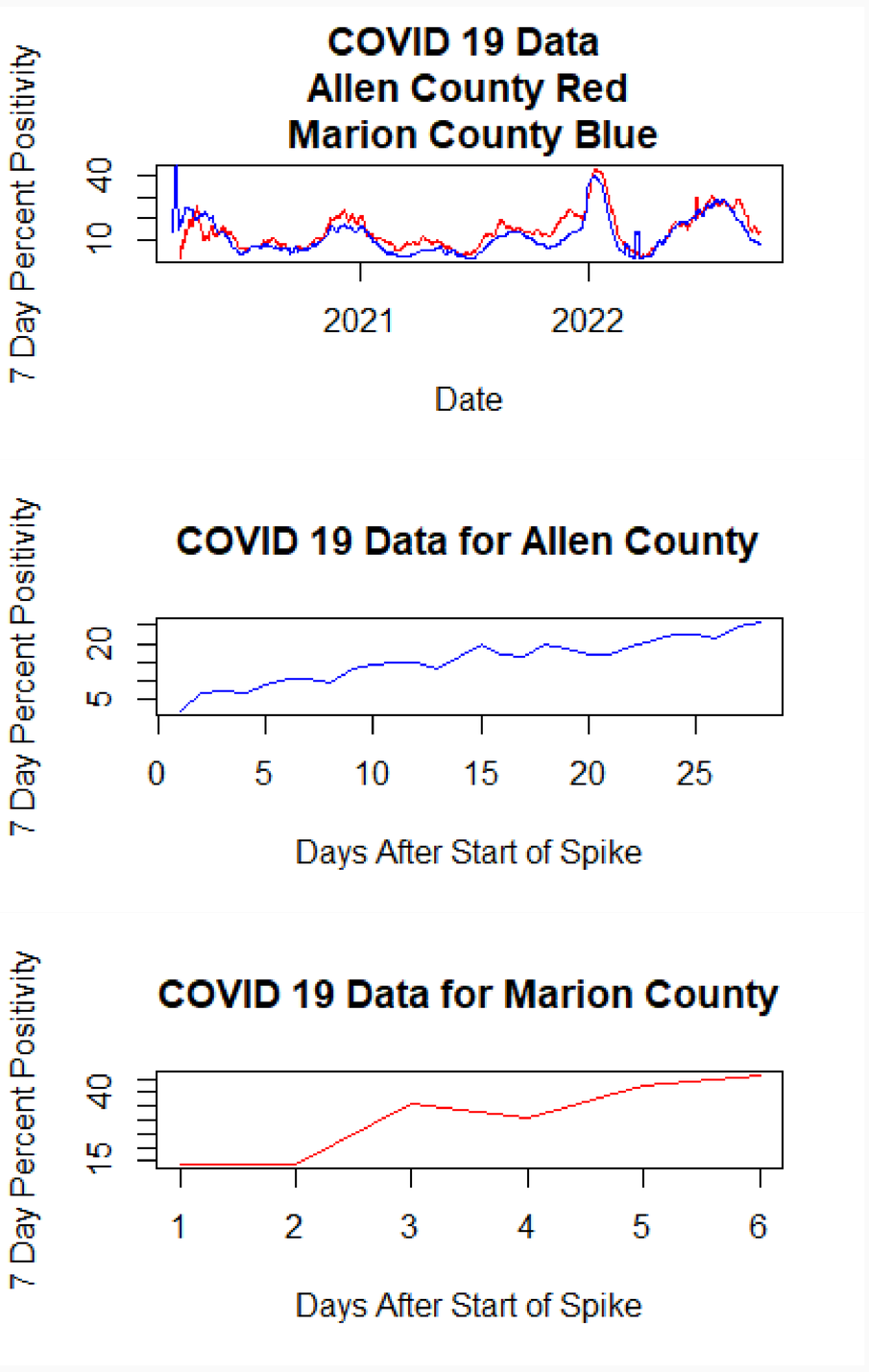
The data we will be using is publicly available, and can be found on the CDC's website (CDC, 2022). We downloaded the time series data for Marion County and Allen Count, and then modified the original spreadsheets by eliminating columns extraneous to this study for ease of management. This data will be input into RStudio, where date (month/day/year) will be defined as the explanatory variable and infection rate as the response variable, for both data sets separately. We also created a column for "days after start of the spike" for ease of computation with a least squares regression model.

The next step is to create a scatterplot for both data sets to determine if a linear model is appropriate or if the data needs transformed to achieve linearity. If needed, we will separately perform different types of transformations on the data sets to determine which type of transformation yields the best linear model for each data set. Once this method is determined, we will use RStudio to create these linear models.

A combined scatterplot will be created for both data sets, with the linear models superimposed. This final product will allow for easy visual comparison of the infection rates for Marion County and Allen County.

Assuming we are able to create models for the entire scope of the pandemic, we will then use the models to make predictions for what infection rates might look like over the next couple months and discuss both statistical and real-world dangers of this particular extrapolation in real-world context.

Plots



Analysis

We ran the data in RStudio using time as the explanatory variable and 7 Day Percent Positivity as the response variable, which produced the following time series chart. From a brief visual inspection, it is obvious that there is not a good basic function type to model the entire timespan.

When it comes to COVID-19 as a disease, an exponential function was what we expected to be used to map the infection rate for the first several weeks of the infection beginning its spread. An exponential growth function is given as $x(t) = x_0\, b^t$, but we use linear regression to find the exact growth factor of COVID-19 within specific counties. Therefore, we rewrite our formula as $log(x(t)) = log(x_0) + log(b) * t$.

An exponential model is used to explain exponential growth. Exponential growth is often a word for fast growth. In the real world, exponential growth can not carry on indefinitely but it's good in the beginning for a short period of time. When we talk about exponential growth that means that quantity increases and so do the rates at which it grows. In our case, we are talking about infectious diseases, and the cases become a sizable fraction of the total population, so the permitting population is significantly smaller and growth will be slower than exponential. That is why after the first week, we should not use the exponential model, diseases don't grow exponentially for a long time.

With the data imported into Rstudio, we can use this linear regression formula to find the exact growth factor as mentioned before, which more accurately interprets the linear model of the COVID-19 infection rate, and ultimately allows us to predict said rate. However, what we found was that the plots of the initial spikes appeared linear in nature! See Plots, above.

In response to this unexpected finding, we ran basic linear regression in RStudio and came up with the following linear models and R-squared values. For both models, y is the predicted 7 day percent positivity rate and x is the number of days after the start of COVID infections in the respective county. Note that we are comparing the initial spikes for each county, so each data set starts on a different date, and contains a different number of days, but both data sets contain the respective start dates and relative maxima for positivity rates. Marion County's spike started at 3/6/2020 and peaked at 3/11/2020, and Allen County's spike started at 3/18/2020, and peaked at 4/14/2020. basic linear regression yielded the following equations and R-Square values:

Marion: $y = 6.357 + 6.879x$, $R - Square = .8613$

Allen: $y = 5.21825 + .70505x$, $R - Square = .92$

As seen by the high R-Square values and visually linear nature of the plots, it seems reasonable to use basic linear regression with no transformations of the data in contrast to what we originally expected we would need to do.