Disease Mapping & Prediction (DRAFT– 50% complete)

Sara Pessognelli

November 16, 2023

Abstract

In this project we will explore modelling, regression, and Bayesian Inference techniques that can be used to not only map and describe the spread of disease throughout a population, but also determine the susceptibility of individuals to certain diseases.

Secondarily, we will explore how disease spreads using branching processes, specifically the Galton-Watson process. As case studies, we will use well known, recent epidemics and pandemics– like Ebola, COVID-19, and swine flu (H1N1)– that have much of the necessary modelling information and data sets available.

Secondarily, we will use data sets that link certain genetic and health markers (things like familial histories, heart health indicators, vaccination status, etc.). We want to use Bayesian inference and regression models (like naive Bayes classifiers and Bayesian neural networks) to make accurate predictions about the outcomes of disease.

Work to be added:

- R_0 estimation for H1N1 data
- Python models of the Galton-Watson process, compared to the SIR model
- Bayseian AIML model to determine lethality of diseases
- Determine if R_0 really changed after lockdown, using Bayesian inference method from [1]
- (Maybe) discuss η, the probability of extinction.
- (Maybe) Reorganize to sort by disease, rather than topic.

Contents

1 Modelling the Spread of Disease.

1.1 Early Stages: The Galton-Watson Process.

The spread of disease, especially the early stages of epidemics, is easily modeled using branching processes—namely, the Galton-Watson process. A branching process is a stochastic process, and the Galton-Watson has two distinct features: a) all individuals give birth according to the same probability law independently of each other, and b) the number of offspring produced by an individual is independent of the number of individuals in that generation [4].

With respect to epidemiology, the number of individuals in generation n, X_n, is equivalent to the number of **new cases** during a given infectious period, and offspring refers to the number of new people, $Y^(j)_k$ each individual in a generation infects, where j refers to the generation and k to the individual of that generation. So, clearly

$$X_{n+1} := Y_1^{(n)} + Y_2^{(n)} + ... + Y_{X_n}^{(n)}$$

[32].

It is not unreasonable to model the number of offspring using a Poisson distribution [31]:

$$Y_K^{(j)} \in Po(\lambda)$$

And in fact, this simplifies some of the math. When discussing the spread of disease, one of the relevant variables of interest is called R_0. This value, R_0, is the **basic reproductive number**, and is a metric used to describe the transmissibility of a disease [11]. Referring back to Ahlberg's paper

$$R_0 = EY$$

which, in the case of the Poisson distribution, is simply the Poisson parameter, λ. In other words,

$$R_0 = EY = \lambda$$

Intuitively, this makes sense. The offspring, $Y_k^{(j)}$, really just represents the number of infectious contacts of k^{th} individual in generation j. So, if we think about the number of infectious contacts as being Poisson distributed, then we already know that λ is the mean number of infectious contacts, by definition of the Poisson parameter.

1.2 Late Stages: The SIR Model.

Interestingly, it is well documented that probabilistic models, like the Galton-Watson branching process, are really only necessary in the **beginning** stages of an outbreak. As time progresses, the simpler SIR model–which is a deterministic ODE model[32][29]. In the SIR model, "S" stands for **susceptible**, "I" for **infected**, and "R" for **recovered**. From Smith and Moore of the Mathematical Association of America, we therefore know we can model an outbreak of disease over time as the following:

$$s(t) = \text{proportion of population that is susceptible at time t.}$$

$$i(t) = \text{proportion of population that is infected at time t.}$$

$$r(t) = \text{proportion of population that is recovered at time t.}$$

And the differential equations related to the SIR model are as follows:

$$\frac{ds}{dt} = -b \cdot s(t)i(t)$$

$$\frac{dr}{dt} = k \cdot i(t)$$

$$\frac{di}{dt} = b \cdot s(t)i(t) - k \cdot i(t)$$

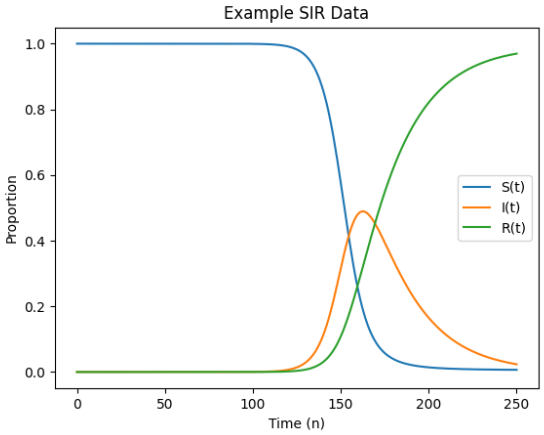
Where b is the proportion of infected-susceptible interactions per day, that result in infection, that are sufficient to spread the disease, and k is the fraction of the infected group that will recover during a given generation. For our purposes, k will equal one. Again, intuitively, these equations make sense. Of course the number of individuals leaving susceptible status is proportional to -b, the number of susceptible people and the number of infected people, at time t. And then, of course, the number of infected increases by the same, and decreases by the number of infected at time t, when $k = 1$. And of course the number of recovered is just the number of people leaving infected status at time t.

A natural question is how the parameters b and k relate to R_e or R_0. We can define R_e as follows

$$R_e = \frac{b \cdot s(t)}{k}$$

So, R_e is **decreasing** overtime, and when once $R_e < 1$, the infected population will decrease to 0[33]. Something interesting to note about this model is that frequently the parameters b, k, and $i(0)$ are typically determined by a trial and error method [33]. Looking ahead to the section where we work out R_e for various

Figure 1



counties in the US, we can see the orange line on the chart above, $I(t)$, shares a similar shape to what we see in the "New Cases by day" charts, as we would expect.

2 Parameter Estimation Using Baye's Theorem.

2.1 What is R_0?

If you are at all familiar with epidemiology or the study of disease in general, then you've probably heard of the value R_0. Simply stated R_0 is the average number of people an infected person will infect during their infectious period. The value varies by disease, and it depends on many factors, but in general the higher the R_0 value the more likely the disease is to spread through a population.

In the context of this paper, we can think of R_0 as the expected number of children in the Galton-Watson process we will use to describe the spread of an epidemic. It is widely accepted that the number of infectious interactions a carrier will have is Poisson distributed. So, if we let $Y_k^{(j)}$ be the infected cases ("offspring") caused by individual k in generation j, then

$$Y_k^{(j)} \in Po(\lambda)$$

We'll talk more about the Galton-Watson Process later. For now, we will see if we can estimate $R_0 = \lambda$. Below are currently (as of November 2023), recent studies have found the accepted R_0 values for various strains of SARS-CoV-2 [10].

COVID-19 Strain	R_0
Alpha	1.22
Beta	1.19
Gamma	1.21
Delta	1.38
Omnicron	1.90

But, at the beginning of the pandemic, the literature has initial calculations for R_0 much higher and were much more varied. One Chinese study[8] compared 12 studies published in the first few months of 2020 and found R_0's ranging from 1.5 to 6.68. Especially in the early stages, under reporting due to lack of awareness could account for artificially low R_0s. The great variation in early reports can be explained somewhat by the behaviors and mitigation method certain populations take. R_0 is the average number of infectious **interactions**, so societies that reduce their interactions via lockdown or other means will skew their local R_0 while societies that make no changes will see high R_0s. For example, an Italian model estimated R_0 from 2.76 to 3.25 initially but after mitigation measures saw a decrease in R_0 [11].

2.1.1 R_0 vs. R_e.

R_0 is the basic reproductive number and is the reproductive ratio that most people are familiar with, but there is also the effective reproductive number, R_e, that changes as the immunity of the population changes. Where you might consider R_0 to be the "true" reproductive ratio, R_e can be thought of as describing how the disease acts in reality. For example, R_0 is affected purely by the infectiousness of the organism and the rate of recovery and death during an outbreak. While, R_e is affected by herd immunity, either through natural immunity of significant proportions of the population contracting the disease or through immunization efforts. Obviously, such efforts will affect the rate at which the disease spreads, but we don't describe this using R_0, we described this using R_e [11].

Within the confines of this paper, we will be attempting to estimate R_0 values of diseases in various counties and locations. We will also be seeing how certain efforts, namely lockdown efforts, affect the spread of disease. In these cases, we will be estimating R_e and the effect that mitigation efforts have on R_e. Interestingly, this distinction also comes into play with our discussion of Ebola and how that disease was able to ravage the western coast of Africa from 2014 to 2016.

2.2 λ as a Random Parameter

Can I condition on the assumption that $\lambda \in Po(i_{\text{infectious}})$ where $i_{\text{infectious}}$ is a random parameter representing the number of infectious interactions an infected person has? Is this too complicated?

2.3 Estimating R_0 using Baye's Theorem.

Bearing all this in mind, we will try to estimate R_0 using Baye's Theorem during the early stages of the pandemic. Using CDC data which was collected daily from January 22, 2020 to August 7, 2020 that contains the number of confirmed cases and confirmed deaths by county in the US. First, lets' talk about estimating Poisson Parameters. Let n be the generation. We can think of each generation as being the **infectious period** of each infected person. From the most recent guidance form the CDC, we know that each generation, n, is 5 days [9][13].

Let $X(n)$ be the number of newly infected people in a generation. Then

$$X(n) = Y_1^{n-1} + Y_2^{n-1} + ... + Y_{X(n-1)}^{n-1}$$

where $Y_k^{(j)}$ is infectees ("offspring") of the k^{th} individual from the $n-1$ generation, and $X(n-1)$ is the total number of infected people in generation $n-1$. It is very typical to estimate the number of offspring as i.i.d Poisson random variables:

$$Y_k^{(j)} \in Po(\lambda)$$

Therefore, R_0 is the expected value of $Y_k^{(j)}$, which from Gut's textbook we know to be

$$R_0 = E[Y_k^{(j)}] = \lambda$$

Dr. Jarad Niemi of Iowa State University has many lectures on this topic. Below, is a synthesized version of his lectures if estimating Poisson parameters, applied specifically to the case of evaluating R_0 [7].

Let y be the observed confirmed cases of SARS-CoV-2 in the CDC dataset and recell $R-) = \lambda$. Baye's Theorem tells us

$$P(\lambda|y) = \frac{P(y|\lambda) \cdot P(\lambda)}{P(y)}$$

or even more simply

$$P(\lambda|y) = C \cdot P(y|\lambda) \cdot P(\lambda)$$

where $C = \frac{1}{P(y)}$ is a proportionality constant. This leaves us with:

$$P(\lambda|y) \propto P(y|\lambda) \cdot P(\lambda)$$

Because we have assumed all $Y_k^{(j)}$ to be independent, we can represent $P(y|\lambda)$ as the **likelihood**, which is product of the individual observations:

$$P(y|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} \cdot e^{-\lambda}}{y_i!}$$

Where y_i represents the observed value infectees for a given observed infected individual. Again using proportionality,

$$P(y|\lambda) \propto \lambda^{\sum^{n} y_i} \cdot e^{-n\lambda}$$

So, we can see that the likelihood, $P(y|\lambda)$ takes the form corresponding to $y|\lambda \in \Gamma(\sum^{n} y_i, \frac{1}{n})$. So, we can therefore choose $\lambda \in \Gamma(\alpha, \frac{1}{\beta})$ as our **conjugate** prior, where a conjugate prior is simply one that has the same type of distribution as the posterior. In fact, we can find the form of the posterior pretty simply like so

$$P(\lambda|y) = C' P(y|\lambda) \cdot P(\lambda)$$

$$P(\lambda|y) \propto P(y|\lambda) \cdot P(\lambda) = \lambda^{\alpha+\sum^{n} y_i - 1} \cdot e^{-(\beta+n)\lambda}$$

So, clearly our posterior is given by

$$\lambda|y \in \Gamma(p = \alpha + \sum^{n} y_i, a = \frac{1}{\beta + n})$$

even cooler, from Gut's textbook, because $\hat{R_0}$ is Gamma-distributed, we know

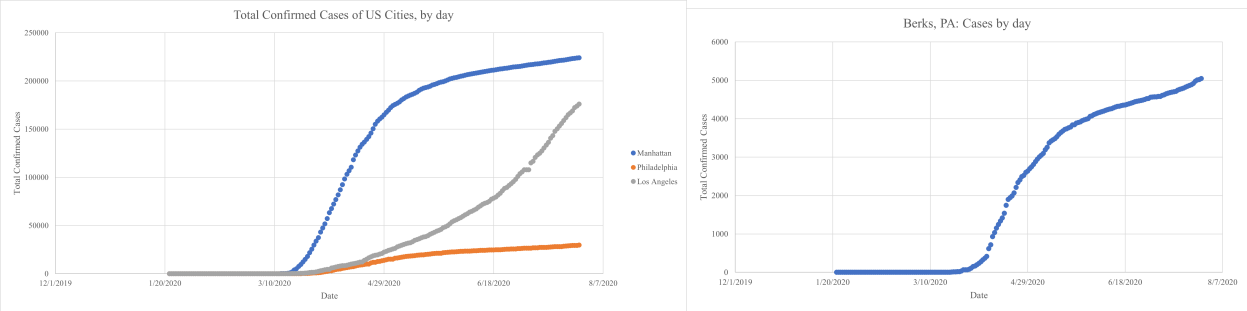
$$\hat{R_0} = E\lambda|y = p \cdot a = \frac{\alpha + \sum^{n} y_i}{\beta + n}$$

Now, let's apply all this to some data!

2.3.1 COVID-19 Data By County.

The data from the CDC, WHO, and CSSE at Johns Hopkins gives us the total number of confirmed cases of COVID-19 and deaths by COVID-19 as a function of date [21]. The data starts on January 22, 2020 and ends on July 27, 2020. During this timespan, the CDC tracked these numbers by county for every county in the United States.

For the purposes of this analysis we'll choose several individual counties from throughout the country to work with individually. Those counties will be Berks County, Pennsylvania, Manhattan county New York, Los Angeles County California, and Philadelphia county, Pennsylvania. The charts below show the cumulative cases and deaths for each of these specified counties:



(a) Total Confirmed Cases Major US Cities, by day

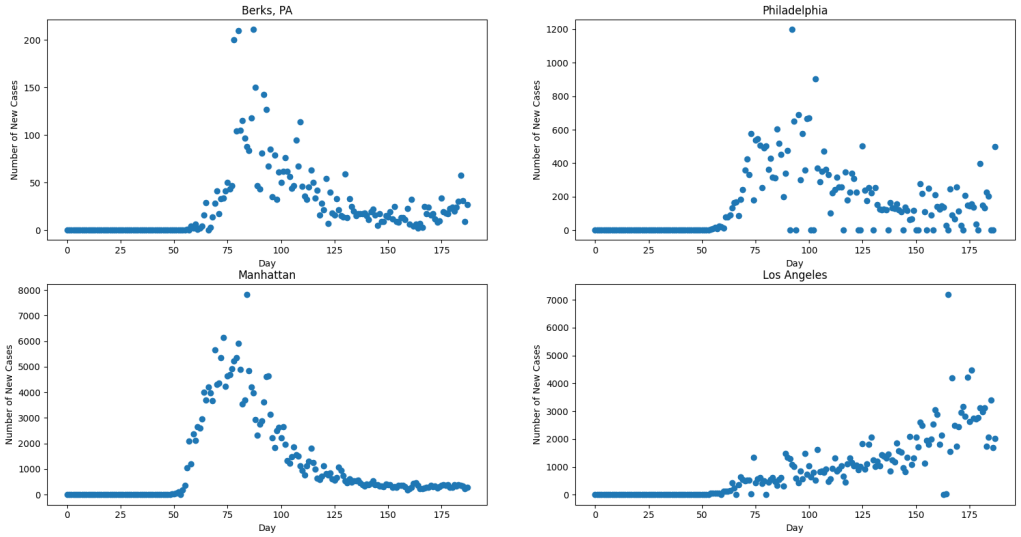
(b) Total Confirmed Cases Berks, PA, by day

Really, what the above graphs are showing is T_i– that is, the total number of cases each day, cumulatively. What we are really interested in is X_i– or the number of new cases each generation. Because the data is taken each day, we simply find X_i to be

$$X_i = T_i - T_{i-1}$$

The CDC data was transformed this way and X_i for the various US counties of interest are plotted below: What's interesting about each graph, is that even though three of them have very similar shapes the scales

Figure 2: New Cases Major US Counties, by day



are drastically different. Not to mention one county, that being Los Angeles, actually has a very different

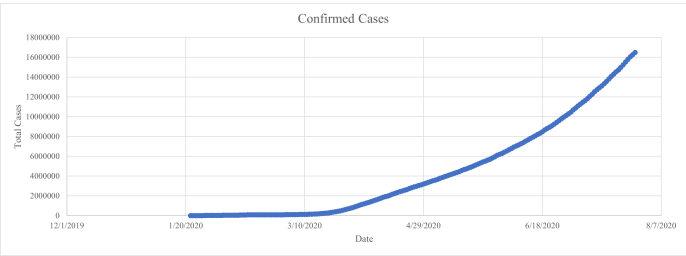
shape than the other three counties in terms of X_n versus N. This was a little surprising to see, since we would have expected similar R0 values for each county considering we're dealing with the same disease in each. But there are some rational reasons why this could have occurred. For one, access to testing resources could have artificially deflated the number of detected cases in certain areas. Of the four counties that we've chosen, Berks County is by far the smallest and least densely populated. R_0 Can be thought of as a random variable in its own right. Really, it represents the number of interactions that people have each day, which itself can be modeled probabilistically, multiplied by the probability of infection of each interaction.

$$\lambda = \text{interactions} * p_{\text{infection}}$$

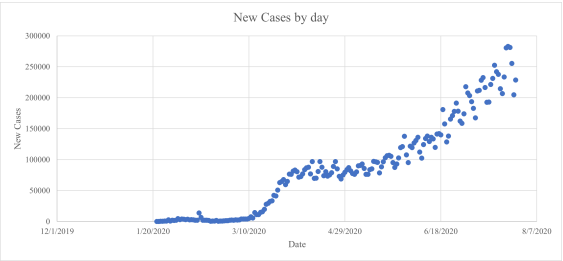
So in densely populated counties like Los Angeles or Manhattan, you would expect to see more interactions per day therefore driving up the effective R_0, or R_e. This could explain why those two counties have many more cases in pure numbers than places like Berks county.

2.3.2 COVID-19 Data Worldwide.

The CDC & JHU also tracked the number of cumulative cases and cumulative deaths by day similarly to how the CDC did for each county in the United States [21]. From this data, using a similar method as before, we will again determine R_0– but this time, for the entire world. Again, the data is visualized below:



(a) Total Confirmed Cases Worldwide, by day



(b) New Confirmed Cases Worldwide, by day

2.4 Ebola Data Set.

Similar to our previous work examining the COVID-19 pandemic, we have our T_i data (i.e., the total rolling reported number of cases) for the 2014-16 West Africa Ebola Outbreak [25]. Looking at the below graphs, we can see a few peculiarities:

- The data comes from the CDC and unlike COVID-19 in the US, the number of newly infected Ebola patients was not recorded daily, so there are some several day gaps in the data.
- At times, the total case count decreases by a few patients day-to-day. It is unclear how this is possible, and is not discussed on the CDC's site. Most likely, some previous suspected cases are ruled out as being Ebola.
- Given the date and location of this outbreak, if anything the number of reported cases is well under the true number of cases.
- Unlike the COVID-19 data that was granular to the scale of individual counties, the Ebola data is totalled across all three countries: Guinea, Sierra Leone, and Liberia.

Figure 3: Total Confirmed Ebola Cases in West Africa (2014-16), using the raw CDC data set.

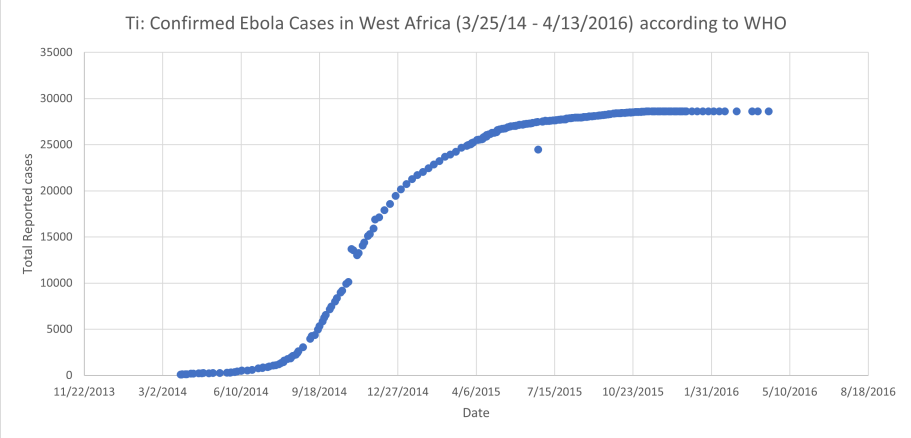
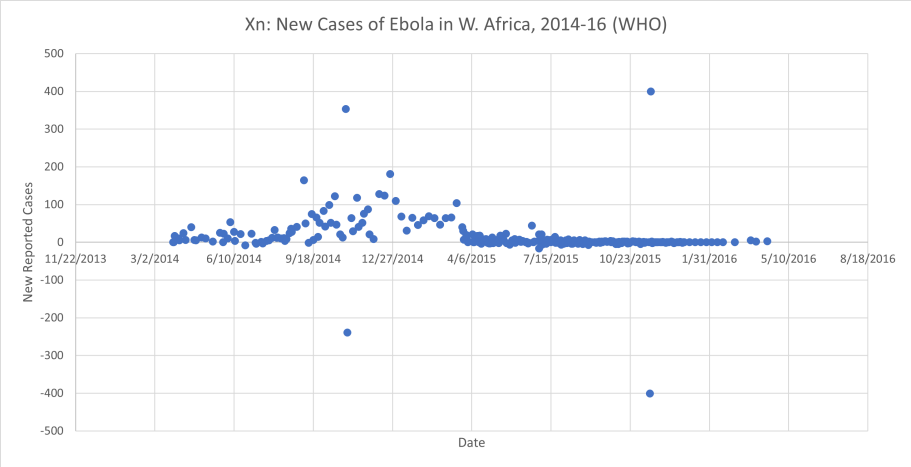


Figure 4: Newly Confirmed Ebola Cases in West Africa (2014-16), using the raw CDC data set.



2.5 H1N1 Data Set.

2.6 Assumptions & Methodology.

2.6.1 COVID-19.

So, we have our X_i data for each county, as a function of 1 day steps. This would imply, if $X_i = X_n$, that

$$n = \text{generation} = 1 \text{ day}$$

With the power of hindsight, we know that this is not the case. The current CDC guidance is that after symptoms occur, most people are infectious for the next 5 days [9]. So, we should instead have:

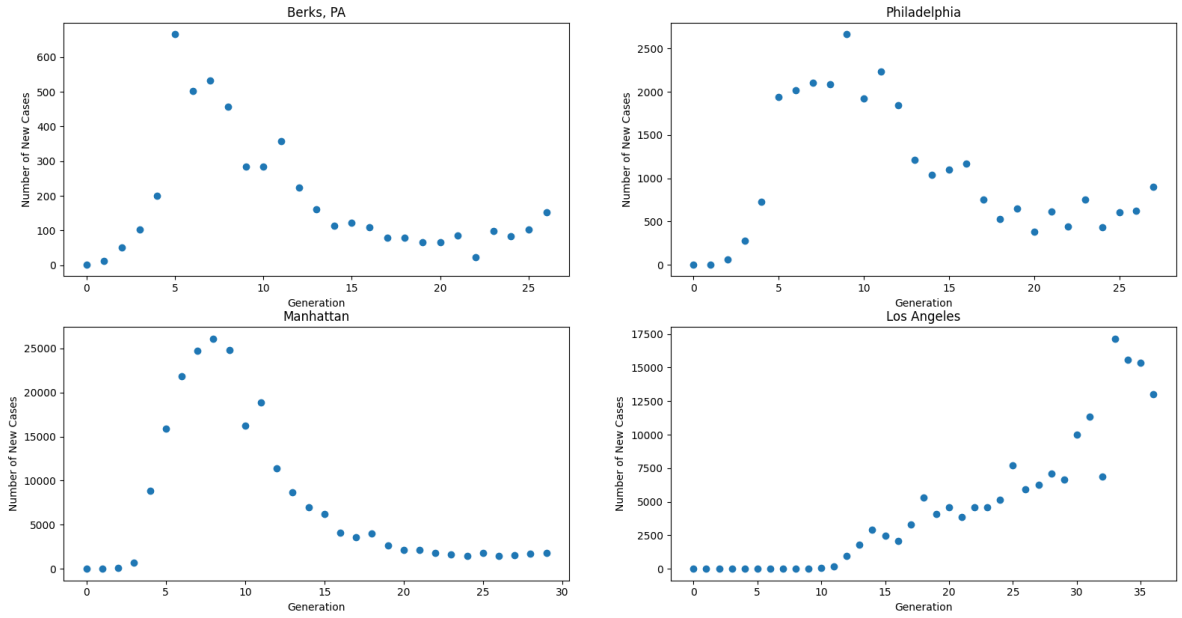
$$n = \text{generation} = 5 \text{ days}$$

This is an easy enough fix, and we can use the T_i data for each county to calculate the appropriate X_n, given the generation length. This is simply

$$X_n = T_i - T_{i-5}$$

where $n = 1$ for each county is with respect to the first detected case in that county. The data looks similar in shape, but note the peaks are higher: Recall form before that

Figure 5: X_n: New Cases Major US Counties, n = 5 days



$$EX_k = \lambda^k$$

for a branching process, and because we are assuming the number of infectious interaction to be $Y_k^{(j)} \in Po(\lambda)$, we therefore have

$$EY = \lambda = R_0$$

and, thus,

$$EX_k = R_0^k$$

We can use the data we have graphed above to estimate R_0 for each generation, k. We will denote the observed X_k as x_k, and because for each county we only have 1 data point at each generation, $x_n = \bar{X_n} \approx EX_n$ in this

case. The interesting thing is, that even though we only have 1 X_n for each $n \geq 1$ for each county, because we have several generations and we are assuming $Y_k^{(j)} \in Po(\lambda = R_0)$, we end up with a point of estimate of $\hat{R_0}$ per generation, per county. Also, let the estimators for λ– which we will call y_k– are simply derived by

$$y_k = (x_k)^{\frac{1}{k}}$$

for each observed generation k. (In other words let x_k be our observed newly infected at each generation of $n = 5$ days, and let y_k be the observed R_0 at that generation– this is probably a better way to say this...)

Those y_k come out to the following: Pretty neat! Now, we will let our prior for λ be

Figure 6: y_k: $\hat{\lambda}$ estimations by generation, by county, for n = 5 days

Generation (n)	Berks	Philadelphia	Manhattan	Los Angeles
1	3.606	1.732	3.162	0.000
2	3.708	3.979	4.380	0.000
3	3.186	4.072	5.178	0.000
4	2.888	3.738	6.156	0.000
5	2.955	3.530	5.016	0.000
6	2.431	2.966	4.168	0.000
7	2.191	2.602	3.541	0.000
8	1.975	2.338	3.096	1.318
9	1.759	2.201	2.751	1.302
10	1.672	1.989	2.414	1.466
11	1.632	1.902	2.272	1.554
12	1.516	1.783	2.052	1.693
13	1.438	1.661	1.911	1.707
14	1.371	1.589	1.805	1.703
15	1.350	1.549	1.727	1.630
16	1.319	1.515	1.631	1.567
17	1.274	1.445	1.577	1.569
18	1.259	1.391	1.547	1.571
19	1.233	1.382	1.483	1.516
20	1.222	1.328	1.442	1.494
21	1.224	1.339	1.417	1.456
22	1.144	1.304	1.385	1.443
23	1.211	1.318	1.361	1.421
24	1.193	1.275	1.339	1.407
25	1.195	1.279	1.334	1.411

$$\lambda \in \Gamma(1,1) \stackrel{d}{=} Exp(1)$$

and we can use our expression for the expectation of the posterior that we derived earlier

$$\hat{R_0} = E\lambda|y = p \cdot a = \frac{\alpha + \sum^{n} y_i}{\beta + n}$$

to estimate R_0 for each county interest, assuming $n = 5$, $\alpha = 1$, and $\beta = 1$. A summary of these results (including varied generation duration) is below:

Generation days	Berks	Philadelphia	Manhattan	Los Angeles
1	1.072	0.948	1.12	0.82
5	1.776	1.946	2.335	1.152
10	2.996	3.796	5.481	1.687

Table 1: $\hat{R_0}$ values for various generation lengths (i.e., infectious periods) in various US counties.

2.6.2 Ebola.

As for our assumptions, we will start by accounting for the peculiarities noted earlier with respect to the CDC data:

- **Newly infected Ebola patients was not recorded daily, so there are some several day gaps in the data:** This is handled by simply filling in those missing days using the last recorded number of infected.
- **At times, the total case count decreases by a few patients day-to-day:** This issue is handled identically to the first issue.

As for assumptions, we have already discussed that Ebola is not infectious until the onset of symptoms. It is also accepted that once symptoms emerge, and untreated Ebola victim will dies in about 10 days [19]. Therefore,

$$n = generation = 10 \text{ days}$$

and, recalling our notation that i represents days, while n represents generations,

$$X_n = T_i - T_{i-10}$$

Something else that is relevant in this case is that we are likely missing data regarding the earlier generations of this outbreak. As aforementioned, the outbreak was believed to have originated from bats in Guinea in December of 2013, but the data from the CDC begins on March 25, 2014– with an initial count of 86 Ebola cases. We have already established $n = 10$ days, this means that from December 31, 2013 to March 25, 2014 at least 5 generations have passed, as far as our Galton-Watson model is concerned. So, we will adjust accordingly.

Again, recall that we are assuming the number of offspring Ebola cases caused by an infected individual to be Poisson distributed, $Y_k^{(j)} \in Po(\lambda)$. This means,

$$EX_k = R_0^k$$

where, to account for the missing generations, $k \geq 5$.

Now, using the same prior and estimation of $\hat{R_0}$ we did in our exploration of the COVID-19 Data:

$$\lambda \in \Gamma(1,1) \stackrel{d}{=} Exp(1)$$

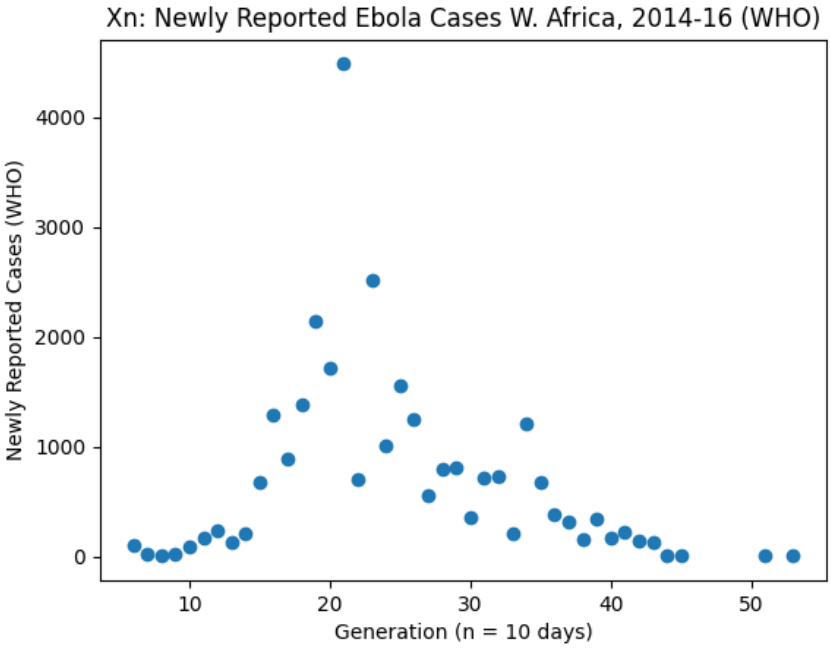
$$\hat{R_0} = E\lambda|y = p \cdot a = \frac{\alpha + \sum^{n} y_i}{\beta + n}$$

we can also make an estimate for R_0 of the Ebola outbreak in West Africa, varying some of the assumptions we have already made:

Generation (days)	**Initial generation**	$\hat{R_0}$
1	**1**	0.528
5	**1**	2.180
10	**1**	3.811
1	**3**	0.292
5	**3**	0.943
10	**3**	1.293
1	**5**	0.287
5	**5**	0.897
10	**5**	1.183

Table 2: $\hat{R_0}$ values for various generation lengths (i.e., infectious periods) in West Africa.

Figure 7: X_n, after making reasonable assumptions about the CDC data.



2.6.3 H1N1.

2.6.4 Discussion of Results.

Methodology Discussion. This method is only feasible if the infectious period for each individual is known– as we can see, the estimated $\hat{R}_0$ depends quite a bit on this interval. Too low, and we may underestimate R_0, and too high and we may overestimate R_0. Though, another way to think of the infectious period = 1 day case, is to think of it as the average number of people an infectious person infects in a day– though, they maybe infectious for longer.

Ebola Discussion. As previously mentioned in our discussion of R_0 versus R_e, the way the public attempts to mitigate the spread of disease has major implications on the resultant R_e value. An interesting note, in the case of Ebola, is that unlike swine flu or COVID-19, Ebola is spread only through contact with infected fluids. This means one can't contract the disease through coughing or other aerosolized methods of transmission. This means that all of the cases of Ebola in western Africa from 2014 to 2016 came directly from physical contact with an infected person's viscera.

The initial patient was believed to be a toddler from a village in Guinea that was infected by bats, during December of 2013 [17]. From there, it was able to spread into neighboring countries Sierra Leone and Liberia because of the increased mobility of people in West Africa compared to previous Ebola outbreaks, and the fact that their were no early detection systems for the disease and health care workers in the area were untrained on diagnosing the ancient illness [18]. While, while this all makes sense, the incubation period for Ebola is about 10 days [19], and people are not infectious until symptoms begin– at which point the disease is spread through infected bodily fluids.

It is reasonable to wonder how the disease spread so rapidly, given most people would avoid someone with Ebola-like symptoms. This is where the unique customs of western Africa shaped the effective reproduction, R_e, of Ebola in 2014. In the region, there is a strong "adherence to ancestral funeral and burial rites singled out as fuelling large explosions of new cases" [18]. Medical anthropologists in the region had even previously noted that the areas funeral practices were exceptionally high risk.For example, in Liberia and Sierra Leone "some mourners bathe in or anoint others with rinse water from the washing of corpses" [18]. Bathing in the viscera of a descendant of Ebola explains partially how Ebola was able spread so rapidly in the region. Other factors include severe shortage of medical workers, unfamiliarity with the disease, and the reliance of "traditional" healers over modern medical practitioners all being prevalent in the area.

As for the actual results, we can see why making reasonable assumptions is crucial if this method is to be used to determine R_0 for Ebola. Clearly, R_0 cannot be less than 1, otherwise an outbreak would not occur, as not enough people would be infected– this is why determining an appropriate generation length in days is needed. As we can see, if the generation duration is too low, in this case if we assume it to be 1 or 5 days, we get $R_0 < 1$, which is clearly false. Determining how far into outbreak we are is also paramount: we are assuming $X_0 = 1$, so if we underestimate how many generations of the disease have already propagated, then we will **overestimate** our R_0. For example, in the case of this Ebola data, the first data point we have is 86 confirmed cases. The assumption $X_0 = 1$, means if we assume $X_1 = 86$ as well, we will get

$$\hat{R}_0 = 9.27$$

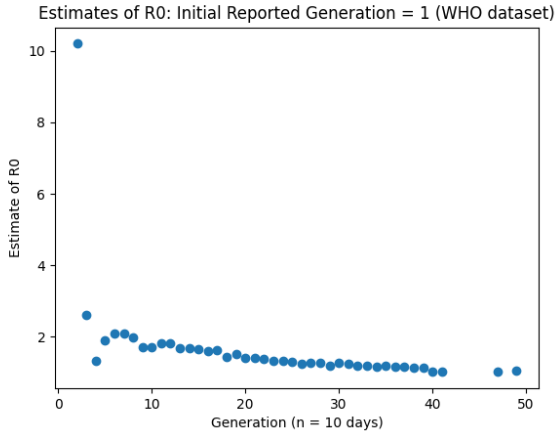
as our first estimate. But, correcting for the missing early generations, assume $X_{10} = 86$ will instead yield

$$EX_{10} = 86 = \hat{R}_0^{\,10}$$

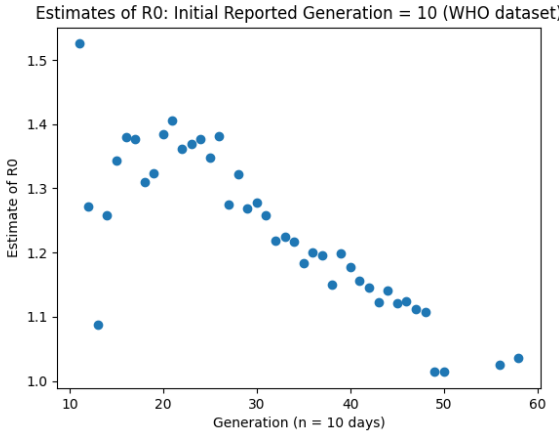
$$\hat{R}_0 = 1.57$$

Interestingly, the estimates of $\hat{R}_0$ do seem to level out over time, as seen in the charts below:

So, given our assumptions, assuming the first generation we have data for is actually $n = 5$ and that the generation length is 10 days, $\hat{R}_0 = 1.183$ is our estimate for R_0 of Ebola. The accepted value is between 1.4-2.0 [16][17].



(a)



(b)

3 Parameter Estimation Using Regression.

3.1 Estimating R_0 using Regression.

3.1.1 COVID-19

Perhaps a better way to estimate our parameter of interest, R_0, it's just simply take advantage of the fact that the expected value of Y is in fact λ. We also know from Theorem 7.2 in Gut's text [4] that the expected value of each generation in a Galton-Watson branching process is simply

$$EX(n) = (EY)^n$$

where n is the generation.

We can use the same data as in our previous method, but representing it differently, we can instead use **linear regression** to determine and estimate of R_0. So, we can rewrite:

$$EX(n) = (EY)^n$$

$$EX(n) = \lambda^n = R_0^n$$

In our data, let x_n be the observed value of X_n in each generation n, $n \geq 1$, and let $\hat{R}_0$ be the observed value of R_0. We therefore have

$$x_n = \hat{R}_0{}^n$$

We can easily linearize this by taking the natural logarithm of each side of our observed data for each county:

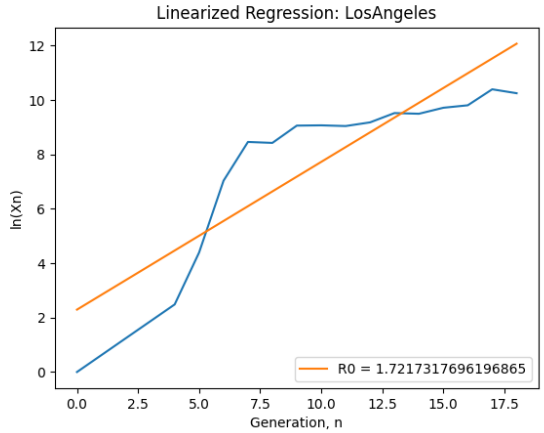
$$\ln(x_n) = n \cdot \ln(\hat{R}_0)$$

Notice that this is in the $y = mx + b$ format. Therefore, the slope, m, of our transformed data can be used to find $\hat{R}_0$ as follows

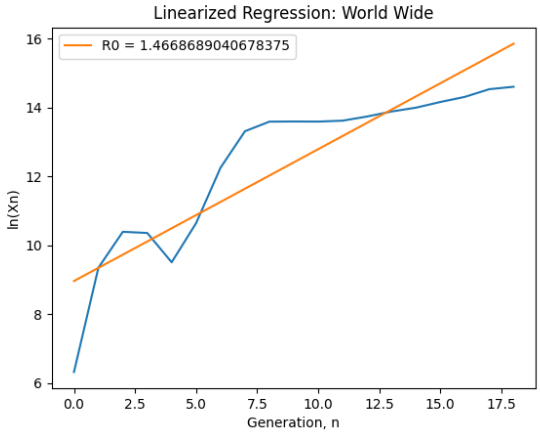
$$\hat{R}_0 = e^m$$

Interestingly, even though it is widely accepted across the literature that pandemic and epidemics are modelled well by branching processes were the offspring of each infected individual is assumed ot be Poisson distributed, many of the graphs from the various counties of interest do not appear to have the expected form of $R_0 = e^m$, as seen below:

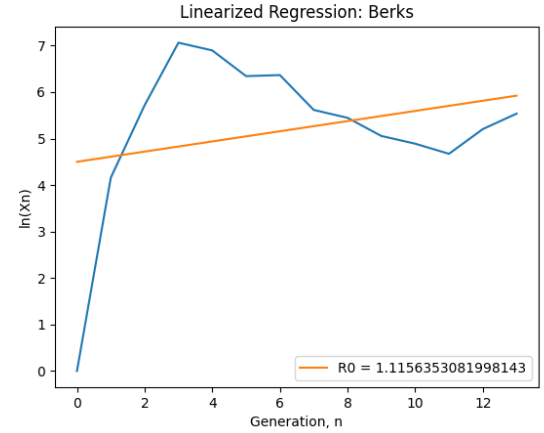
In fact, what it actually looks like is **two** linear functions, with different slopes. Around the third generation in both the cases of Philadelphia county and Berks county, something happens where the steep incline originally seen in cases suddenly drops off. Both of these counties are in Pennsylvania, and the first reported



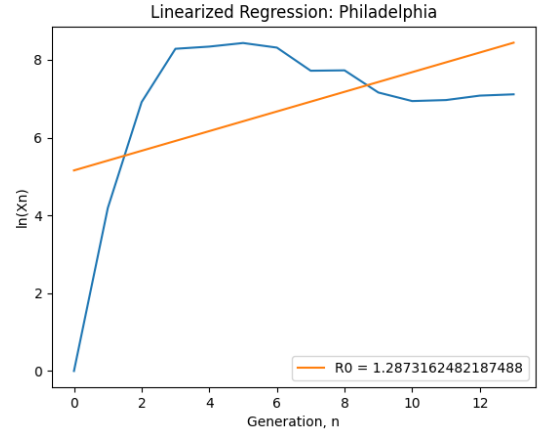
(a) Estimating R_0 in Los Angeles county, CA, from 1/22/20-7/27/20



(b) Estimating R_0 from all confirmed cases from around the world.



(a) Estimating R_0 in Berks county, PA, from 1/22/20-7/27/20



(b) Estimating R_0 in Philadelphia county, PA, from 1/22/20-7/27/20

cases of COVID in Berks and Philadelphia were March 18, 2020 and March 10, 2020, respectively [21]. Given that our generations are 5 days, the fourth generation (20 days after patient 0 is detected in each county) is early April– this is (roughly) where we start to see the slopes of each graph drop off. According to ABC27, the stay-at-home orders in Philadelphia county began **March 23, 2020**, and then Governor Tom Wolff placed all the state on stay-at-home orders on **April 1, 2020** [14]. This begs the questions: **did lock-downs decrease** R_0 **(really** R_e**)**? Based on the graphs it seems so– in fact, the slopes seem to be **negative** after generation 4, and a negative slope on these graphs implies an $R_e > 1$. We will explore this more in the next section, again using Baye's Analysis to see if λ before and after lock-downs is actually different.

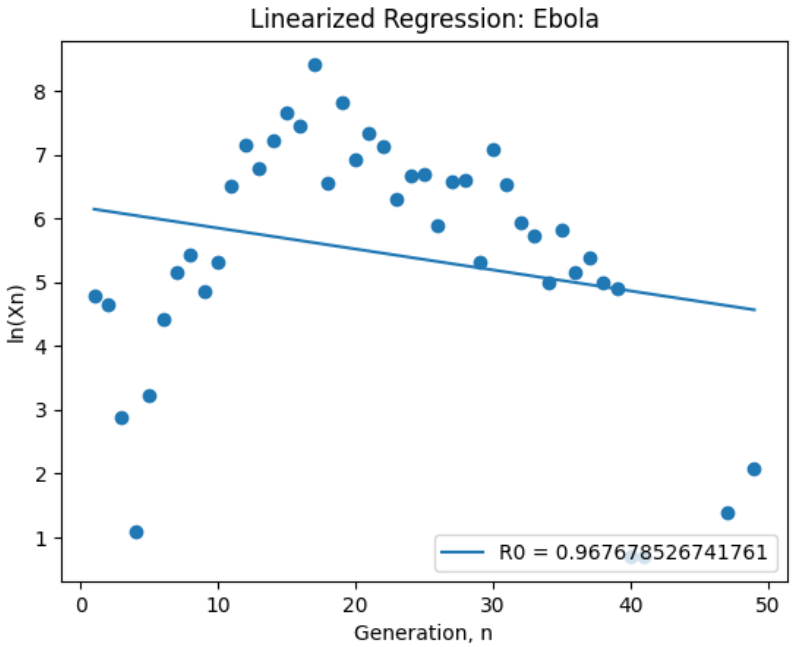
Stay tuned!

3.1.2 Ebola.

We will now use the same method for using linearization and regression to make estimates of $\hat{R_0}$ described in the case of COVID-19 to make estimates of $\hat{R_0}$ in the case of the Ebola outbreak in western Africa:

In the case of Ebola, the distinction between R_0 and R_e becomes even more relevant and visible. As we

Figure 8: $\hat{R_0}$ estimation for for all generations (3/25/2014-4/13/2016)



can see in the figure above, we can see a distinct drop in the slope somewhere around the 25th generation (about 250 days, so late November or early December of 2014 in real time). Recall, that the relationship between R_0 (or rather R_e) and the slope, m, is

$$m = \ln(\hat{R_0}) \iff \hat{R_0} = e^m$$

So, the sudden decrease in slope, m, reflects a decrease in $\hat{R_0}$. In fact,

$$m < 0 \iff \hat{R_0} < 1$$

In fact, if we split the data at the 25th generation and examine the resultant $\hat{R_0}$ values, we see that before the 25th generation there is a (fairly) **increasing** relationship in between $\ln(X_n)$ and n, with a calculated value of

$$\hat{R_0} = 1.20$$

Versus, after the 25th generation a fairly linear **decreasing**relationship in between $\ln(X_n)$ and n, with a calculated value of

$$\hat{R_0} = 0.739$$

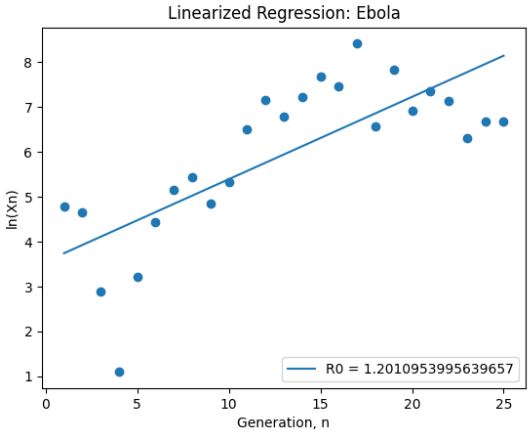
From Ahlberg's paper [32], we know that $\hat{R_0} < 1$ means the outbreak will die out. This naturally begs the question: what happened around November and December of 2014 to cause the sudden decrease in observed R_e?

4 Preventive Measures During the COVID-19 Pandemic.

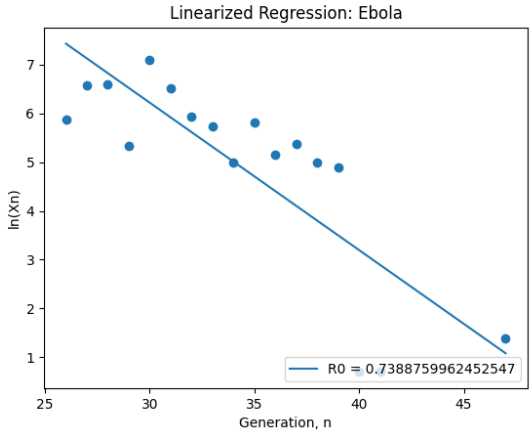
4.1 The Difference in λ Before & After.

4.1.1 Lockdowns.

During the COVID-19 pandemic, did instating a lockdown in a given county decrease the R_e?



(a) $\hat{R_0}$ estimation for first 25 generations (3/25/2014-11/30/2014)



(b) $\hat{R_0}$ estimation for 25th generation to end of outbreak (11/30/2014-4/13/2016)

5 Bayesian Machine Learning for Fatality Prediction.

6 Appendices

For the Python scripts, because of the varied formatting of the data used, some of these scripts are data-specific. Similar scripts exist for the other data, and are included in the attached zip file. These samples can be thought of as pseudocode.

6.1 Appendix A: SIR Model (Python)

```
s0 = 8*10**9
i0 = 1
r0 = 0

total_time = 250
time_step = 1

st = [s0/s0]
it = [i0/s0]
rt = [r0/s0]
b = 0.2
k = 0.04
## Traditional SIR Model

for time in range(0, total_time, time_step):
    t = int(time/time_step)

    ds = -b*st[t]*it[t]

    s1 = st[t] + ds
    st.append(s1)

    di = b*st[t]*it[t] - k*it[t]
    i1 = it[t] + di
    it.append(i1)
```

```
    dr = k*it[t]
    r1 = rt[t] + dr
    rt.append(r1)

## Quick Plot
times = [t for t in range(0, total_time+time_step, time_step)]

import matplotlib.pyplot as plt
import pandas as pd
df = pd.read_csv(r'E:\Research Project\Python Models\COVID19_Data_2020\World Wide.csv')

total_confirmed = list(df['Confirmed'])
Tn = list(df['Confirmed'])
generation_time = 5
Tn = Tn[0::generation_time]

Tn = [t/s0 for t in Tn]

# plot lines
plt.plot(times, st, label = "S(t)")
plt.plot(times, it, label = "I(t)")
plt.plot(times, rt, label = "R(t)")
plt.title("Example SIR Data")
plt.xlabel("Time (n)")
plt.ylabel("Proportion")
plt.legend()
plt.show()
```

6.2 Appendix B: Bayesian Parameter Estimation Model (Python)

```
import os
dir = os.path.dirname(os.path.realpath(__file__))
os.chdir(dir)

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import csv
from math import *

## Import COVID-19 Data (USA-- by county, beginning 1/22/2020)
import pandas as pd
df = pd.read_csv(r"E:\Research Project\Python Models\Ebola_2016_WestAfrica\case_counts_ebola2016.csv")

## Add a new column: New Confirmed Cases and New Deaths (these will represent X(1), X(2),..., X(n))
total_confirmed = list(df['Confirmed']) # The data from the CDC is strictly the total (Tn) number of co
new_confirmed = [total_confirmed[i] - total_confirmed[i-1] for i in range(1,len(total_confirmed))] # By

total_deaths = list(df['Deaths'])
new_deaths = [total_deaths[i] - total_deaths[i-1] for i in range(1,len(total_deaths))]

## Add the as columns on to our dataframe
new_confirmed.insert(0,0)
```

```
df['New Confirmed [X(n)]'] = new_confirmed
new_deaths.insert(0,0)
df['New Deaths'] = new_deaths

## We need to determine how many "offspring" (i.e., infectees) each infected person has from one genera
## Early in a pandemic, it is safe to assume that the infectees of each infector are independent.
generation_time = 10 # day
initial_generation = 5
generation_time = floor(generation_time)
## By independence, EX(n) = (EY(n-1))^X(n)
## Or, (new_cases gen. n+1) = (average number of infectees of gen. n)^(number of infectors in gen. n)

Tn = []
for i in range(len(df['Day'])):
    if i ==0 :
        cases = df['Confirmed'][i]
    if i !=0:
        day2 = df['Day'][i]
        day1 = df['Day'][i-1]

        cases = max(df['Confirmed'][i], cases)

        if day2 - day1 > 1:
            while day2 - day1 > 1:
                Tn.append(cases)
                day1 += 1
        elif day2 - day1 < 1:
            Tn.append(cases)

Tn.append(max(list(df['Confirmed'])))

Tn = Tn[0::generation_time]
Xn = [Tn[i]- Tn[i-1] for i in range(1,len(Tn))]
Xn.insert(0, Tn[0])
## Baye's Parameter Estimation
n = [i for i in range(initial_generation,len(Xn)+initial_generation-1)] ## the 4 and 3 represent the fa
n_copy = n.copy()
yn_1 = [Xn[i]**(1/n[i]) for i in range(len(n))]

sum_yi = sum(yn_1)
n = len(yn_1)

alpha = 1
beta = 1
print("Expected Value = R0:", (alpha+sum_yi)/(n+beta))

plot_y =  []
plot_n = []
plot_xn = []
for i in range(1, len(yn_1)):
    if yn_1[i] != 0:
        plot_y.append(yn_1[i])
        plot_n.append(n_copy[i])
        plot_xn.append(Xn[i])
```

```
plt.scatter(plot_n, plot_y)
# plt.plot(range(0,len(Tn)), list(Tn))
plt.xlabel("Generation (n = 10 days)")
plt.ylabel("Estimate of R0")
plt.title("Estimates of R0: Initial Reported Generation = 10 (WHO dataset)")
plt.show()

plt.scatter(plot_n,plot_xn)
plt.xlabel("Generation (n = 10 days)")
plt.ylabel("Newly Reported Cases (WHO)")
plt.title("Xn: Newly Reported Ebola Cases W. Africa, 2014-16 (WHO)")
plt.show()
```

6.3 Appendix C: Linearization & Regression Model (Python)

```
import os
dir = os.path.dirname(os.path.realpath(__file__))
os.chdir(dir)

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import csv
from math import *

## Import COVID-19 Data (USA-- by county, beginning 1/22/2020)
import pandas as pd
df = pd.read_csv(r"E:\Research Project\Python Models\Ebola_2016_WestAfrica\case_counts_ebola2016.csv")

## Add a new column: New Confirmed Cases and New Deaths (these will represent X(1), X(2),..., X(n))
total_confirmed = list(df['Confirmed']) # The data from the CDC is strictly the total (Tn) number of co
new_confirmed = [total_confirmed[i] - total_confirmed[i-1] for i in range(1,len(total_confirmed))] # By

total_deaths = list(df['Deaths'])
new_deaths = [total_deaths[i] - total_deaths[i-1] for i in range(1,len(total_deaths))]

## Add the as columns on to our dataframe
new_confirmed.insert(0,0)
df['New Confirmed [X(n)]'] = new_confirmed
new_deaths.insert(0,0)
df['New Deaths'] = new_deaths

## We need to determine how many "offspring" (i.e., infectees) each infected person has from one genera
## Early in a pandemic, it is safe to assume that the infectees of each infector are independent.
generation_time = 10 # day
initial_generation = 5
generation_time = floor(generation_time)
## By independence, EX(n) = (EY(n-1))^X(n)
## Or, (new_cases gen. n+1) = (average number of infectees of gen. n)^(number of infectors in gen. n)

Tn = []
for i in range(len(df['Day'])):
    if i ==0 :
```

```
        cases = df['Confirmed'][i]
    if i !=0:
        day2 = df['Day'][i]
        day1 = df['Day'][i-1]

        cases = max(df['Confirmed'][i], cases)

        if day2 - day1 > 1:
            while day2 - day1 > 1:
                Tn.append(cases)
                day1 += 1
        elif day2 - day1 < 1:
            Tn.append(cases)

Tn.append(max(list(df['Confirmed'])))

Tn = Tn[0::generation_time]
Xn = [Tn[i]- Tn[i-1] for i in range(1,len(Tn))]
Xn.insert(0, Tn[0])

### Regression
LnXn = list(np.log(Xn))
day = list(np.linspace(1, len(LnXn), num = len(LnXn), endpoint = True))

i = 0
while i < len(LnXn):
    if LnXn[i] == -inf:
        del LnXn[i]
        del day[i]
    else:
        i+=1


#Plot Xn vs n
gen1 = 25
gen2 = -1
day = day[gen1:gen2]
LnXn = LnXn[gen1:gen2]
plt.scatter(day, LnXn)

## Best Linear Predictor (Theorem 2.5.2 in GIP)
## L(X) = alpha + beta*X
x_bar = np.mean(day)
y_bar = np.mean(LnXn)
var_x = np.var(day)
var_y = np.var(LnXn)
cov_xy = np.cov(day, LnXn)[0][1]

alpha = y_bar - (cov_xy/var_x)*x_bar
beta = cov_xy/var_x

## and thus our BLP is
def BLP(x):
    return y_bar + (cov_xy/var_x)*(x - x_bar)
```

```
## Print the estimated R0
slope_m = cov_xy/var_x
R0_hat = exp(slope_m)

print(len(LnXn))
print("Estimated R0:", R0_hat)

# Plot the BLP line !
plt.plot(day, BLP(day), label = "R0 = %s" % R0_hat)
plt.xlabel("Generation, n")
plt.ylabel("ln(Xn)")
plt.title("Linearized Regression: %s" % 'Ebola')
plt.legend(loc = "lower right")
plt.show()
```

7 Bibliography.

References

Background Sources.

[1] Davidson-Pilon, C. (2016). *Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference*. Addison-Wesley.

[2] Andrieu, C., de Freitas, N., Doucet, A. *et al.* (2003). An Introduction to MCMC for Machine Learning. *Machine Learning* 50, 5–43. https://doi.org/10.1023/A:1020281327116

[3] Tse, K. (2014) Some Applications of the Poisson Process. *Applied Mathematics*,05,3011-3017. DOI: 10.4236/am.2014.519288.

[4] Gut, A. (2009). *An Intermediate Course in Probability*. Springer. [DOI] 10.1007/9781-4419-0162-0

Disease Mapping Specific Sources.

[5] Lupague, R. M.J.M., Mabborang, R. C., Bansil, A. G., & Lapague, M. M. (2023). Integrated Machine Learning Model For Comprehensive Heart Disease Risk Assessment Based On Multi-Dimensional Health Factors. *European Journal of Computer Science and Information Technology*, 11(3), 44-58. https://doi.org/10.37745/ejcsit.2013/vol11n34458

[6] Coly, S., Garrido, M., Abrial, D., & Yao, A. (2021). Bayesian hierarchical models for disease mapping applied to contagious pathologies. *PLoS ONE*, 16(1): e0222898. https://doi.org/10.1371/journal.pone.0222898

[7] Jarad Niemi, Ph.D. (2013, Jan 22). Bayesian inference for Poisson data [Video]. YouTube. [URL] https://www.youtube.com/watch?v=lNrpPNk6InU

Disease Specific Research.

COVID-19.

[8] Liu, Y., Gayle, A. A., Wilder-Smith, A. & Rocklov, J. (2020).The reproductive number of COVID-19 is higher compared to SARS coronavirus. Journal of Travel Medicine, 1-4. 10.1093/jtm/taaa021

[9] Centers for Disease Control and Prevention. (2023, May 11). Isolation and Precautions for People with COVID-19. COVID-19. [URL] https://www.cdc.gov/coronavirus/2019-ncov/your-health/isolation.html#:~:text=If%20you%20test%20positive%20for,at%20home%20and%20in%20public.

[10] Manathunga, S. S., Abeyagunawardena, I. A., & Dharmaratne, S. D. (2023). A comparison of transmissibility of SARS-CoV-2 variants of concern. Virology Journal, 20(59). https://doi.org/10.1186/s12985-023-02018-x

[11] Achaiah, N. C., Subbarajasetty, S. B., & Shetty, R. M. (2020). R0 and Re of COVID-19: Can We Predict When the Pandemic Outbreak will be Contained?. Indian Jounrla of Critical Care Medicine, 24(11), 1125-1127. 10.5005/jp-journals-10071-23649

[12] Netherlands Ministry of Health. (2023, June 11). Reproduction Number. Coronavirus Dashboard. [URL] https://coronadashboard.government.nl/landelijk/reproductiegetal

[13] Ryu, S., Kim, D., Ali, S. T., & Cowling, B. J. (2022). Serial Interval and Transmission Dynamics during SARS-CoV-2 Delta Variant Predominance, South Korea. Emergine Infectious Diseases (CDC), 28(2), 407-410. 10.3201/eid2802.211774

[14] A Year of COVID in Pennsylvania. (2021). ABC 27 WHTM. Retrieved November 06, 2023, from https://www.abc27.com/timeline-of-a-year-of-covid-19-in-pennsylvania/

[15] Procter, R. (2021, March 04). Remember when? Timeline marks key events in California's year-long pandemic grind. Cal Matters, [URL] https://calmatters.org/health/coronavirus/2021/03/timeline-california-pandemic-year-key-points/.

Ebola

[16] Kerkhove, M. D. V., Bento, A. I., Ferguson, N. M., & Donnelly, C. A. (2015). A review of epidemiological parameters from Ebola outbreaks to inform early public health decision-making. Scientific Data, 2. https://doi.org/10.1038/sdata.2015.19

[17] Centers for Disease Control and Prevention. (2019, March 18). 2014-2016 Ebola Outbreak in West Africa. CDC. [URL] https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak/index.html

[18] World Health Organization. (2015, January). Factors that contributed to undetected spread of the Ebola virus and impeded rapid containment. WHO. [URL] https://www.who.int/news-room/spotlight/one-year-into-the-ebola-epidemic/factors-that-contributed-to-undetected-spread-of-the-ebola-virus-and-impeded-rapid-containment

[19] Johns Hopkins Medicine. Ebola. Health. [URL] https://www.hopkinsmedicine.org/health/conditions-and-diseases/ebola#:~:text=If%20treatment%20is%20ineffective%20or,from%20the%20start%20of%20symptoms.

Datasets.

[20] Ministry of Health, Government of Mexico (2020). Information regarding COVID-19 cases in Mexico [Data set]. Salud. [URL] https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico

[21] Center for Systems Science and Engineering (CSSE) at Johns Hopkins (2023). Novel Coronavirus (COVID-19) Cases [Data set]. CSSE Johns Hopkins. [URL] https://github.com/CSSEGISandData/COVID-19

[22] Centers for Disease Control and Prevention (2021). Behavioral Risk Factor Analysis Surveillance System Data [Data set]. CDC. [URL] https://www.cdc.gov/brfss/annual_data/annual_2021.html

[23] Centers for Disease Control and Prevention (2019). 2014 Ebola Outbreak in West Africa Epidemic Curves [Data set]. CDC. [URL] https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak/cumulative-cases-graphs.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fvhf%2Febola%2Foutbreaks%2F2014-west-africa%2Fcumulative-cases-graphs.html.

[24] World Health Organization (2016). Ebola data and statistics [Data set]. WHO. [URL] https://apps.who.int/gho/data/node.ebola-sitrep.quick-downloads?lang=en

[25] Centers for Disease Control and Prevention (2020). Case Counts [Data set]. CDC. [URL] https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak/case-counts.html.

[26] Centers for Disease Control and Prevention (2014). CDC Estimates of 2009 H1N1 Influenza Cases, Hospitalizations and Deaths in the United States [Data set]. CDC. [URL] https://www.cdc.gov/h1n1flu/estimates_2009_h1n1.htm.

Epidemiology.

SIR Models.

[27] Smith, D. & Moore, L. (2004, December). *"The SIR Model for Spread of Disease - The Differential Equation Model"*. The Mathematical Association of America. https://maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model

[28] Cooper, I., Mondal, A., & Antonopoulos, C. (2020). A SIR model assumption for the spread of COVID-19 in different communities. *Chaos, Solitons, & Fractals*, 139. DOI: https://doi.org/10.1016/j.chaos.2020.110057. URL: https://www.sciencedirect.com/science/article/pii/S0960077920304549?via%3Dihub

Branching Models.

[29] Jacob, C. (2010). Branching Proceseses: Their Role in Epidemiology. *International Journal of Environmental Research and Public Health*, 7(3), 1186-1204. DOI: 10.3390/ijerph7031204

[30] Bartoszynski, R. (1965). Branching Processes and the Theory of Epidemics. *Berkeley Symposium on Mathematical Statistics & Probability*, 4, 259-269. [URL] https://digicoll.lib.berkeley.edu/record/113134

[31] Laha A. K., & Majumdar, S. (2022). A multi-type branching process model for epidemics with application to COVID-19. *Stochastic Environmental Research and Risk Assessment*, 4, 259-269. DOI: https://doi.org/10.1007/s00477-022-02298-9

[32] Ahlberg, D. (2021, May 20). Epidemics and branching processes [Lecture notes]. Stockholm University. URL: https://staff.math.su.se/daniel.ahlberg/notes-epidemics.pdf

[33] Cooper, I., Mondal, A., & Antonopoulos, C. G. (2020). Title of article. A SIR model assumption for the spread of COVID-19 in different communities, 139. 10.1016/j.chaos.2020.110057