COVID19 Cases Introductory Data Review

Micheal Wilson

2023-03-04

Preface

This exploration into COVID-19 case count will leverage the John Hopkins University GitHub data. The intent will be to focus on two different locales. Jefferson County, New York, and Faulkner County, Arkansas, and compare them in terms of the case counts over time. Both counties are largely rural with respect to their states, and generally have only one key population center. It is also two locations where I've lived a significant number of years of my life. Having the experience of living there provides me the privilege of having some degree of 'subject matter expertise' as far some differences are concerned between the areas. While the treatment here will be largely superficial, it's mostly intended to demonstrate R markdown functionality, knitting, or other basics of data science as a field.

Goal

The fundamental goal, having no epidemiology or medical experience whatsoever, is to see if there's any insights from 'just' the macroscopic motion of the COVID19 pandemic in terms of it's case counts and simple modeling. Then, conjecture of the various forces acting on the pandemic will be discussed to describe the data.

The states' initial response to the pandemic greatly affected their numbers on the onset, and this is likely reflected down to the county level. In general, state government differences in response, the local cultures, the increased population density of Faulkner county compared to Jefferson, and seasonal differences between the two locales create numerous effects notable in the data. All of those factors likely have an impact on the total case count, or spread, of the disease.

Packages

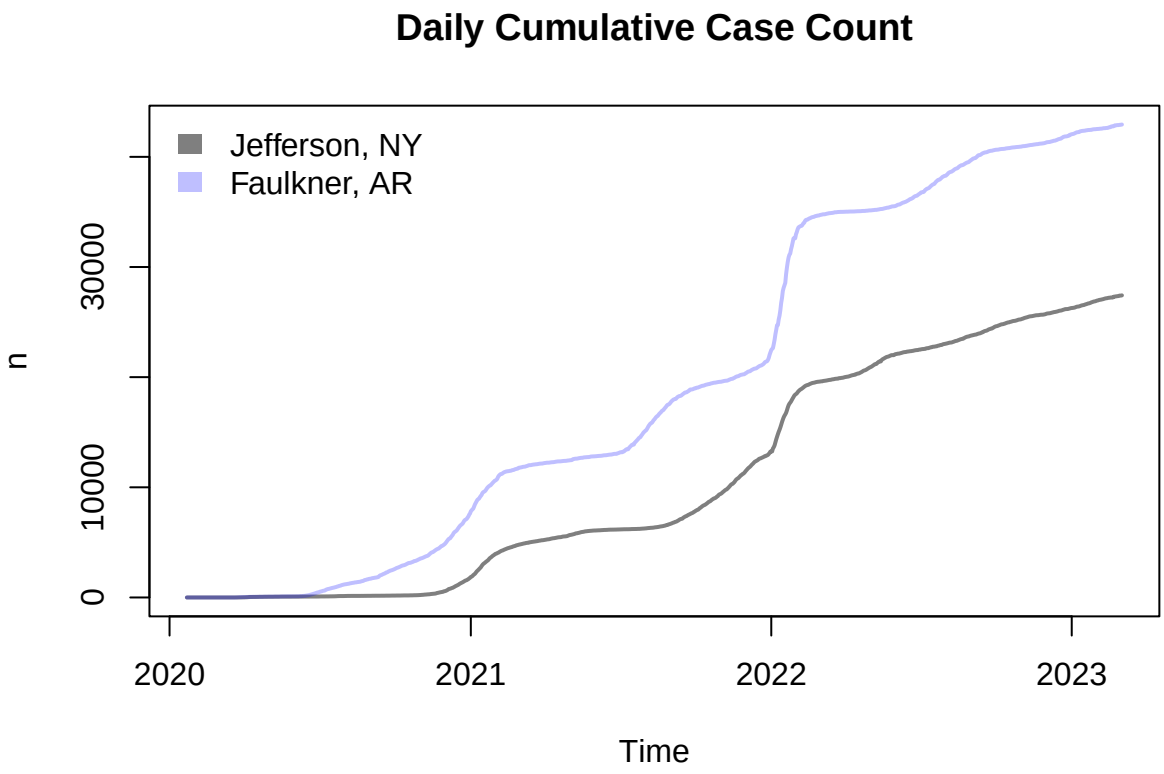
Most of this project will focus on using base R for compatibility for others who may use this work. 'Tidyverse' and 'lubridate' are loaded in case certain analytical tasks need to be performed using those packages. A more important package, 'usdata', is loaded to grab additional county-level information, and it's particularly important for extensions of this data exploration by having access to demographic data associated to the Johns Hopkins University data through the county FIPS keys.

Data Extract, Load, Transform (ETL)

Data extraction and tidying is performed largely in base R for compatibility. Full code and markdown are available. Processes for tidying and transforming could be automated for this dataset for rapid report generation and model development, but those methods are not employed here. Static or fixed callouts, references, and definitions, are leveraged specifically for this narrow analysis of a considerable dataset.

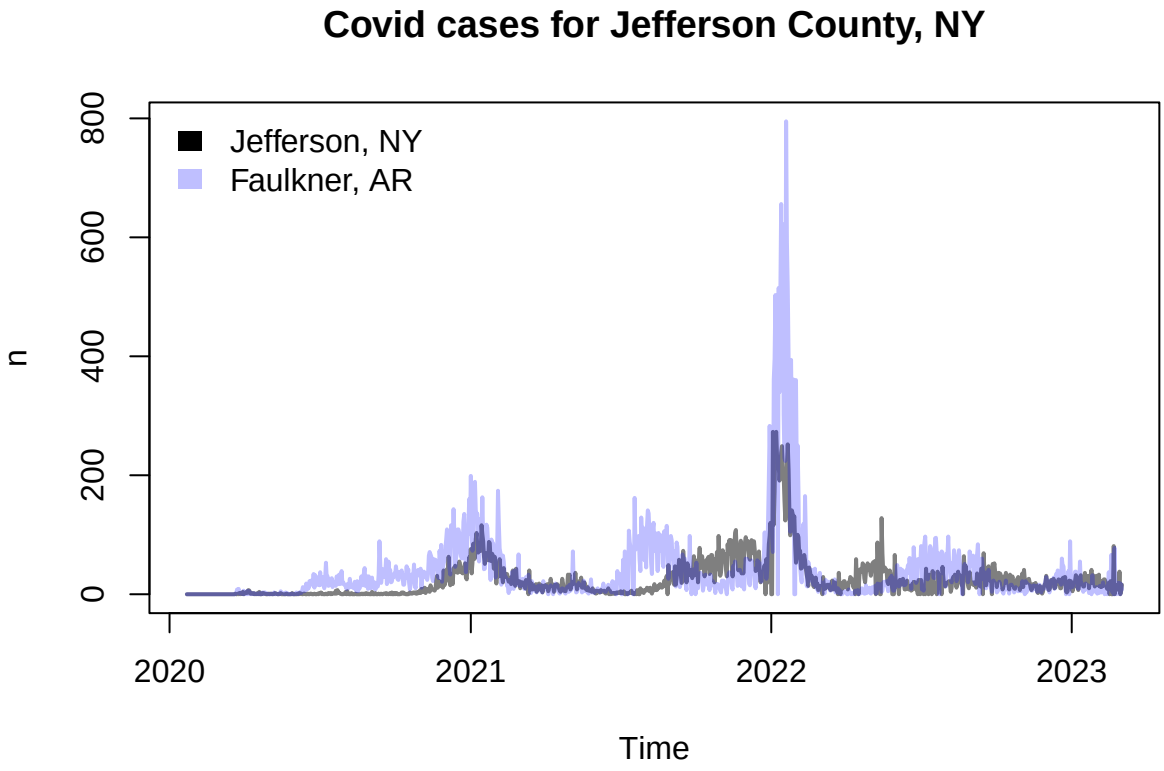
Data Exploration through Visuals

The daily, cumulative sum, over time is plotted immediately below. It is interesting to note the generally 'sigmoidal' shape of the curves. This is characteristic of epidemics' studies as well as any count, growth, or spread process being modeled.



A few key differences exist between the curves. Firstly, Faulkner County, AR, shown with the blue line, has significantly more cases throughout the entirety of the pandemic. This is of interest given that the populations of both counties are largely the same. Jefferson County, NY, and Faulkner County, AR, have populations of 116,721 and 123,498 respectively (according to 2020 US Census data.) With only a ~5% difference in populations, it's remarkable that there is a significantly larger amount of cases in Faulkner County than in Jefferson. This observation warrants further exploration.

The next graphic depicts the daily counts, and is effectively what generates the cumulative summation chart shown previously. The next chart shows the daily rate of spread, or using a physics phrasing, it shows the velocity magnitude of the pandemic's spread in the locale.



Some key points of interest are the timings of peaks of both counties. In Faulkner, the peaks coincide with warmer times of the year. Both counties have peaks around the holidays. Both counties appear to have delayed, minor (localized), peaks in case counts in the spring months after the holiday season. The magnitude of those peaks is largely similar in the exception of the holiday season for 2022.

The additional seasonal peaks during the summers in Faulkner County are a surprising occurrence on the surface. Arkansas is notably quite warmer than New York in the summer. The difference in the states' cultures with regards to outdoor activities in the summer are largely moot. Both counties feature a wealth of outdoor activities and in states that pride themselves and capitalize on their availability of outdoor activities. The key difference that seasonality would seem to impose is that in Arkansas summers, it's so much more hot than what people are reasonably comfortable with, so they tend to stay indoors. Whereas, in NY, the summers are nearly ~20deg (F) cooler on average, and provide a much better opportunity to be outside.

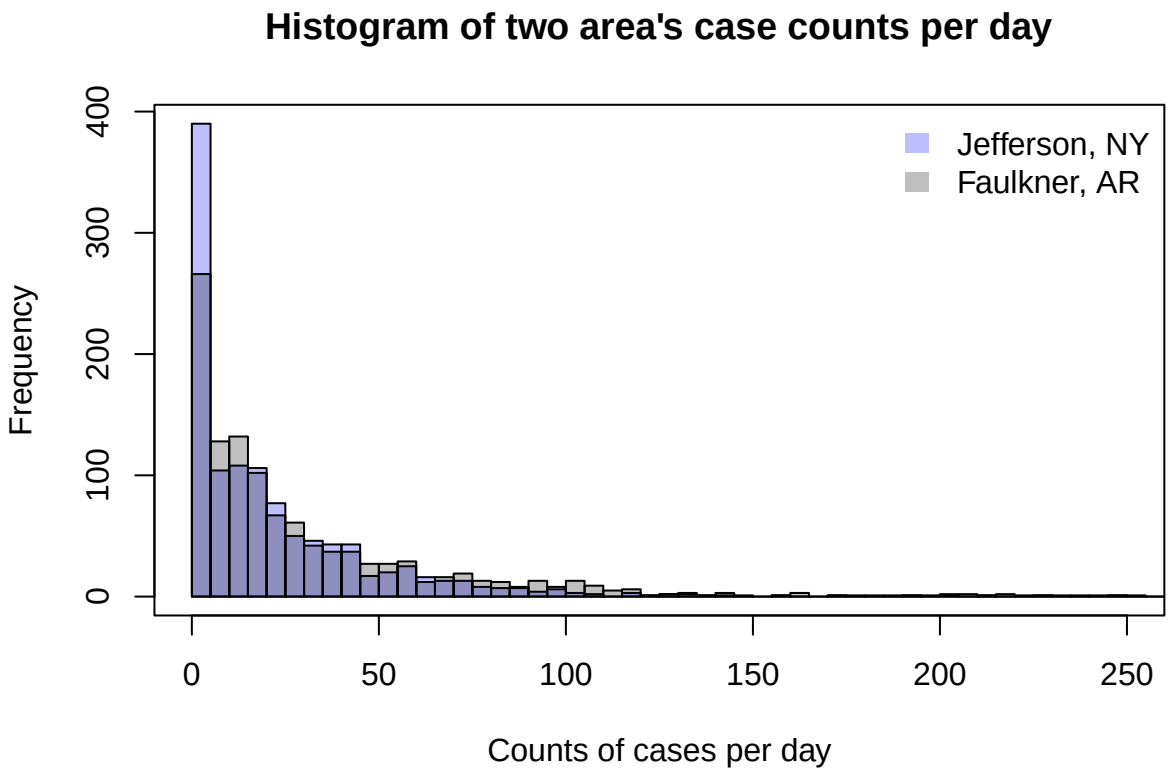
The peak near the holiday season of 2022 is most easily, and likely, attributed to the introduction and outbreak of the highly contagious 'Omicron' variant in the wake of the holiday season. Every holiday season has marked peaks due to peoples' tendencies to want to spend time with eachother around that time of year– a cultural artifact true for both counties.

The one truly unclear difference is the larger quantity of case counts generated early on in 2020. There's no immediately plausible reason why the county begin accruing cases at a much larger rate than Jefferson County. It could be attributed to the summer seasonal conditions noted previously, but the case counts are much lower than the following year summer peaks–indicating other forces may be at play.

This next chart examines the distribution of case counts over time via a layered histogram. Since the numbers of cases reported each day is largely independent[1] from the cases reported on any other day, we can assert

[1] *Note: Independence in the data points here specifically refers to the numbers reported on one day are independent from the next. It is understood that cultural events, or seasonal changes as demonstrated above, can create localized changes in time*

that the distribution of counts is generally a 'Poisson Point Process'.



This histogram definitely reveals a distribution that appears to be that of the "Gamma family" of distributions. That is the family of distributions commonly used to characterize a quantity of interest in an interval. At this time, the connections between the distributions of cases per day and the factors or forces involved in a pandemic will not be explored. It is likely that there is a distribution with variables that vary with respect to time that could model expected case counts on any one day, but that is beyond the scope of this analysis.

What is gained from looking at cases by day is statistics about the rate of spread of the epidemic.

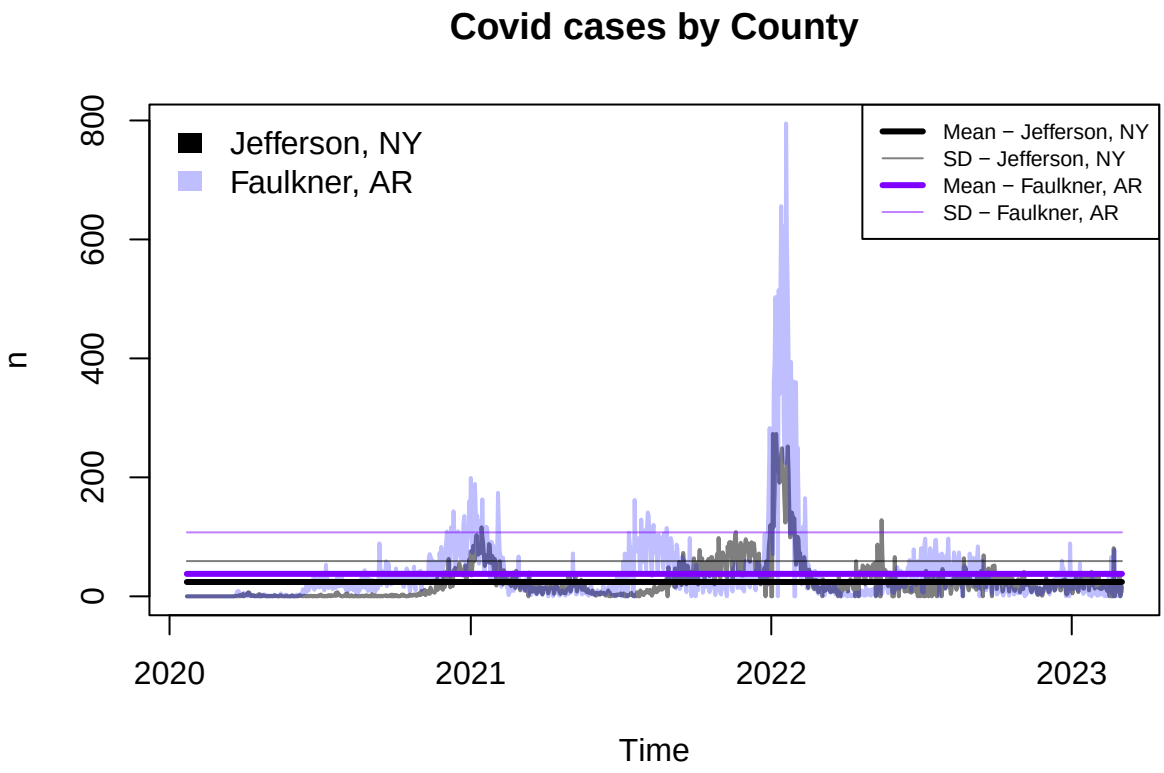
Data Modeling

Doing simple statistics between the daily case counts by county to start things off:

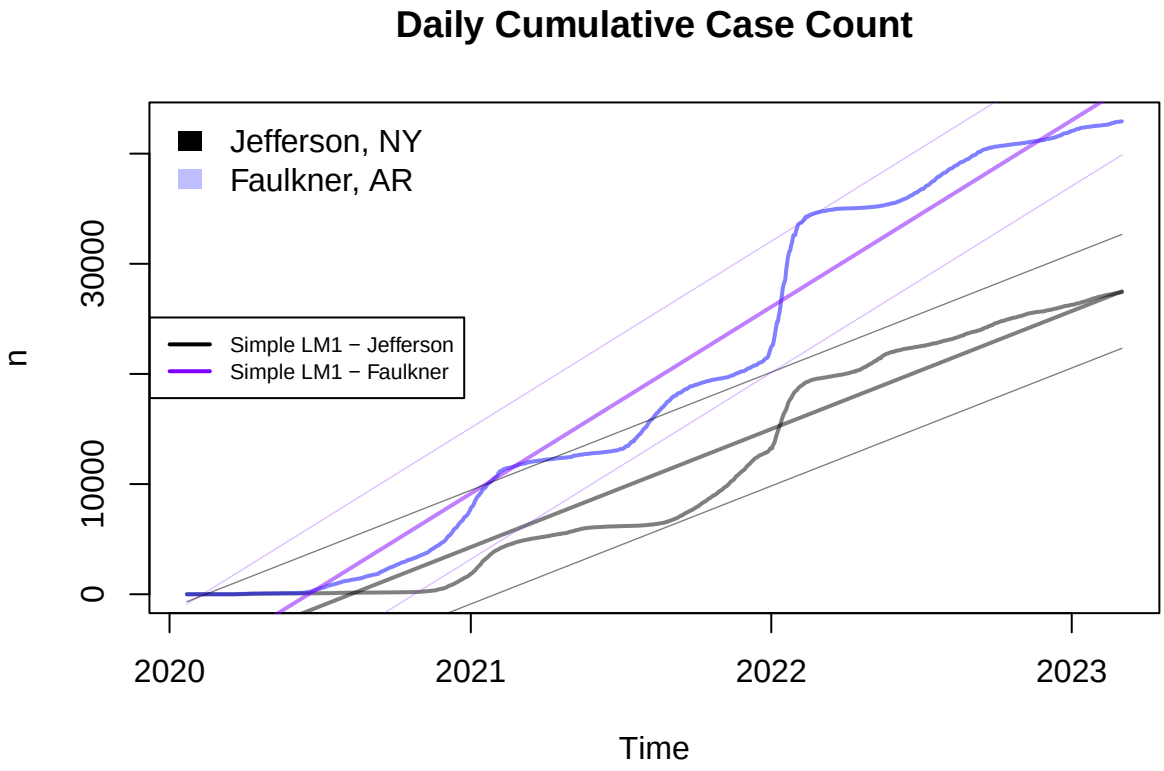
```
## [1] "Jefferson County Daily Case Avg:  24.12"

## [1] "Jefferson County Daily Case Standard Deviation:  35.11"

## [1] "Faulkner County Daily Case Avg:  37.76"

## [1] "Faulkner County Daily Case Standard Deviation:  69.76"
```

to the case counts created for any given day. That is not the same as saying that "because we had x cases yesterday, today we'll have y", and considering that there's an incredible number of factors that go into recording a case count that introduce statistical noise, characterizing case counts as a Poisson Point Process should be permissible.

With the first and second moments of the data known, basic modeling can begin. This following approach is rudimentory for demonstration purposes only, but it can serve as a valuable insight to characterizing different time points in the data depending on the goal.



This shows that the mean and mean plus one standard deviation seem to adequately capture the majority of daily case count data. This provides indicators in the data for where there may be outliers, but a better use might be for using those time periods for investigation cross correlating events or auto correlations in time series modeling.




```
##
## Call:
## lm(formula = df[, "Jefferson, New York, US"] ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4721.5 -2211.6   372.2  2191.2  5847.2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5876.5666   156.1177  -37.64   <2e-16 ***
## x              29.3543     0.2377  123.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2630 on 1135 degrees of freedom
## Multiple R-squared:  0.9308, Adjusted R-squared:  0.9307
## F-statistic: 1.525e+04 on 1 and 1135 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = df[, "Faulkner, Arkansas, US"] ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4559.9 -2623.7  -502.5  2345.6  6864.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6910.7206   180.4151   -38.3   <2e-16 ***
## x              46.4109     0.2747   169.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3040 on 1135 degrees of freedom
## Multiple R-squared:  0.9618, Adjusted R-squared:  0.9617
## F-statistic: 2.855e+04 on 1 and 1135 DF,  p-value: < 2.2e-16
```

Model Review and Bias

Given the model's predictive performance, the p-values of the model parameters, and the R-squared being considerably favorable, it would seem this is a fine model to describe the cumulative counts over time for these respective counties. For the intent of this assignment, that being to simply create a model of interest, the objective is met. Yet, if one were to naively accept this model as a factual epidemiological behavior of COVID-19, the very next logical leap could be to also assert "this is how COVID-19 pandemics spread." On the other hand, there is much more at play than a simple linear relationship between "here's a type of epidemic" and "here's the cases over time." This begins the review, and the discussion on the implicit biases for this model.

The model only illuminates the general function used to map a time variable to it's co-domain as a cumulative sum of cases. This model does not explore relationships or factors that contributed to the spread over the last 3 years. Subscribing to these approaches on modeling by "just using the linear model", "just using the package", or now, "just let the AI handle it" presents a host of issues. This is tantamount, and arguably just as dangerous and potentially amoral, as businesses using simple linear regression to do financial forecasts and earnings projections to appease the shareholders if it makes the earnings report look better. This model does not demonstrate true causality, only correlation to time, at best. The bias here is implicit, but it's very plainly one that dismisses the complicated nature of epidemics.

Going Further

Given that the intent of this data review is primarily to demonstrate the capabilities of R markdown documentation, a more appropriate approach will be proposed (but not demonstrated) for sake of completeness. The spread of a disease is something that starts off very slow and with a small proportion of the population, but at some point in time it begins to rapidly climax, and eventually it will slow down due to (mostly) population saturation. Since it is well known that s-shaped curves is a family of curves associated to growth processes in finite spaces, proposing a model that produces a sigmoid shape would be a much better handling than simple linear equation. In the context of R language, using the 'glm' or 'nls' functions to fit the COVID-19 to "Richard's curve", or a generalized logistic function, would be a better approach. Since those curves are smooth and generally stop after their first plateau, the cumulative case counts presented would need several of those curves summed in different phases time in order to properly account for all of the individual contributory processes that accounts for all of the covid cases in the respective counties.

More often, since there's auto-correlation, or dependency, between the data points in cumulative sums (each successive point on the x axis depends on the point behind it), looking at the total case counts violates alot of what's really sought after in regression modeling. (For example, residuals will not be gaussian-distributed.) That being said, time series modeling techniques exist to address auto-correlative data in time. Yet, another

approach would be to consider the daily case counts. Figuring out which factors go into generating a number of cases per day might provide better results.