COVID-19 Cases vs. Mortality

Art Steinmetz

12/4/2020

Introduction

I have a macabre fascination with tracking the course of the COVID-19 pandemic. I suspect there are two reasons for this. One, by delving into the numbers I imagine I have some control over this thing. Second, it feels like lighting a candle to show that science can reveal truth at a time when the darkness of anti-science is creeping across the land.

The purpose of this project is, as usual, twofold. First, to explore an interesting data science question and, second, to explore some techniques and packages in the R universe. We will be looking at the relationship of COVID-19 cases to mortality. What is the lag between a positive case and a death? How does that vary among states? How has it varied as the pandemic has progressed? This is an interesting project because is combines elements of time series forecasting and dependent variable prediction.

I have been thinking about how to measure mortality lags for a while now. What prompted to do a write-up was discovering a new function in Matt Dancho's `timetk` package, `tk_augment_lags`, which makes short work of building multiple lags. Not too long ago, managing models for multiple lags and multiple states would have been a bit messy. The emerging "tidy models" framework from RStudio using "list columns" is immensely powerful for this sort of thing. It's great to reduce so much analysis into so few lines of code.

This was an exciting project because I got some validation of my approach. I am NOT an epidemiologist or a professional data scientist. None of the results I show here should be considered authoritative. Still, while I was working on this project I saw this article in the "Wall Street Journal" which referenced the work by Dr. Trevor Bedford, an epidemiologist at the University of Washington. He took the same approach I did and got about the same result.

Aquire and Clean Data

There is no shortage of data to work with. Here we will use the NY Times COVID tracking data set which is updated daily. The package `covid19nytimes` lets us refresh the data on demand.

date	location	location_type	location_code	location_code_type	data_type	value
2020-11-28	Alabama	state	01	fips_code	cases_total	244993
2020-11-28	Alabama	state	01	fips_code	deaths_total	3572
2020-11-28	Alaska	state	02	fips_code	cases_total	31279
2020-11-28	Alaska	state	02	fips_code	deaths_total	115
2020-11-28	Arizona	state	04	fips_code	cases_total	322774
2020-11-28	Arizona	state	04	fips_code	deaths_total	6624

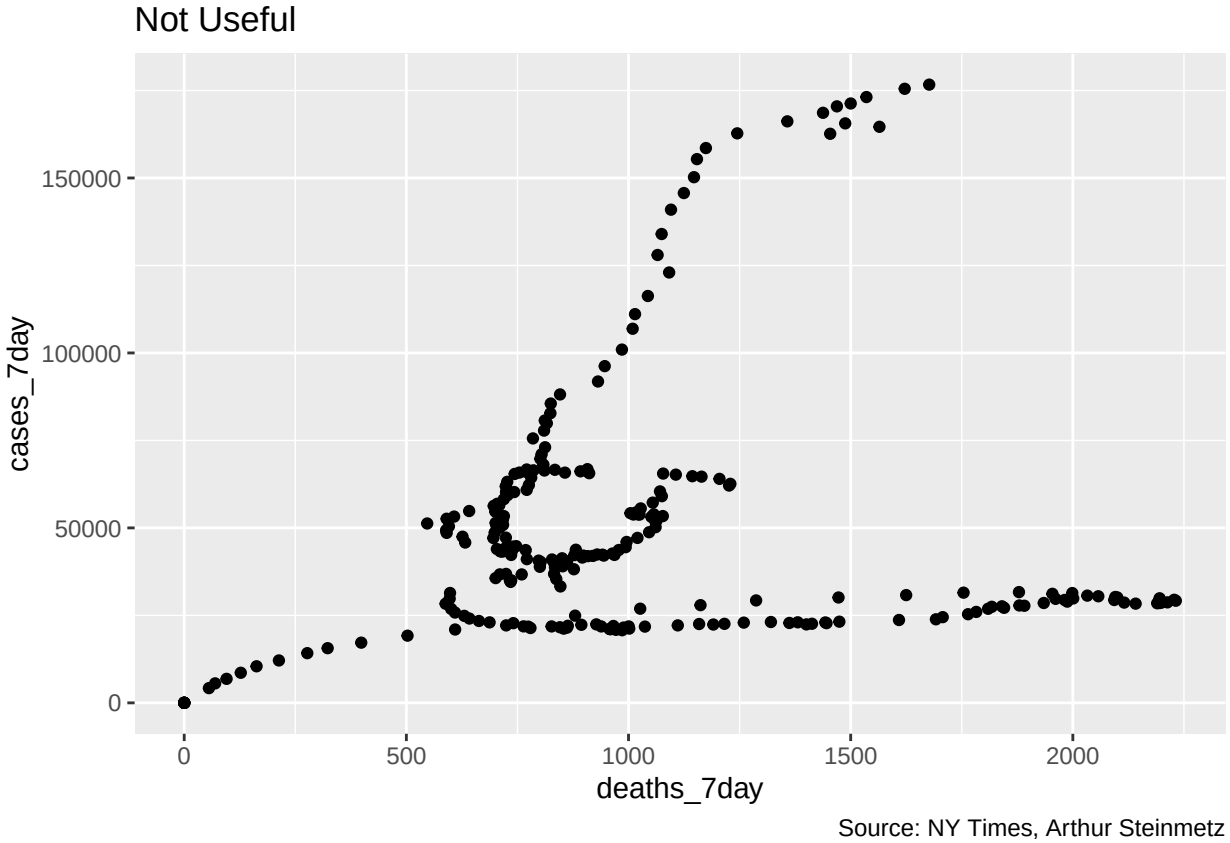
The NY Times data is presented in a "long" format. When we start modeling, long will suit us well but first

we have to add features to help us and that will require `pivot`ing to wide, adding features and then back to long. The daily data is so irregular the first features we will add are 7-day moving averages to smooth the series. We'll also do a nation-level analysis first so we aggregate the state data as well.

date	cases_total	deaths_total	cases_7day	deaths_7day
2020-03-24	53906	784	6857.571	95.28571
2020-03-25	68540	1053	8599.714	127.28571
2020-03-26	85521	1352	10448.571	162.85714
2020-03-27	102847	1769	12121.286	213.14286
2020-03-28	123907	2299	14199.143	277.00000
2020-03-29	142426	2717	15625.714	322.85714
2020-03-30	163893	3367	17202.429	398.42857
2020-03-31	188320	4302	19202.000	502.57143
2020-04-01	215238	5321	20956.857	609.71429
2020-04-02	244948	6537	22775.286	740.71429
2020-04-03	277264	7927	24916.714	879.71429

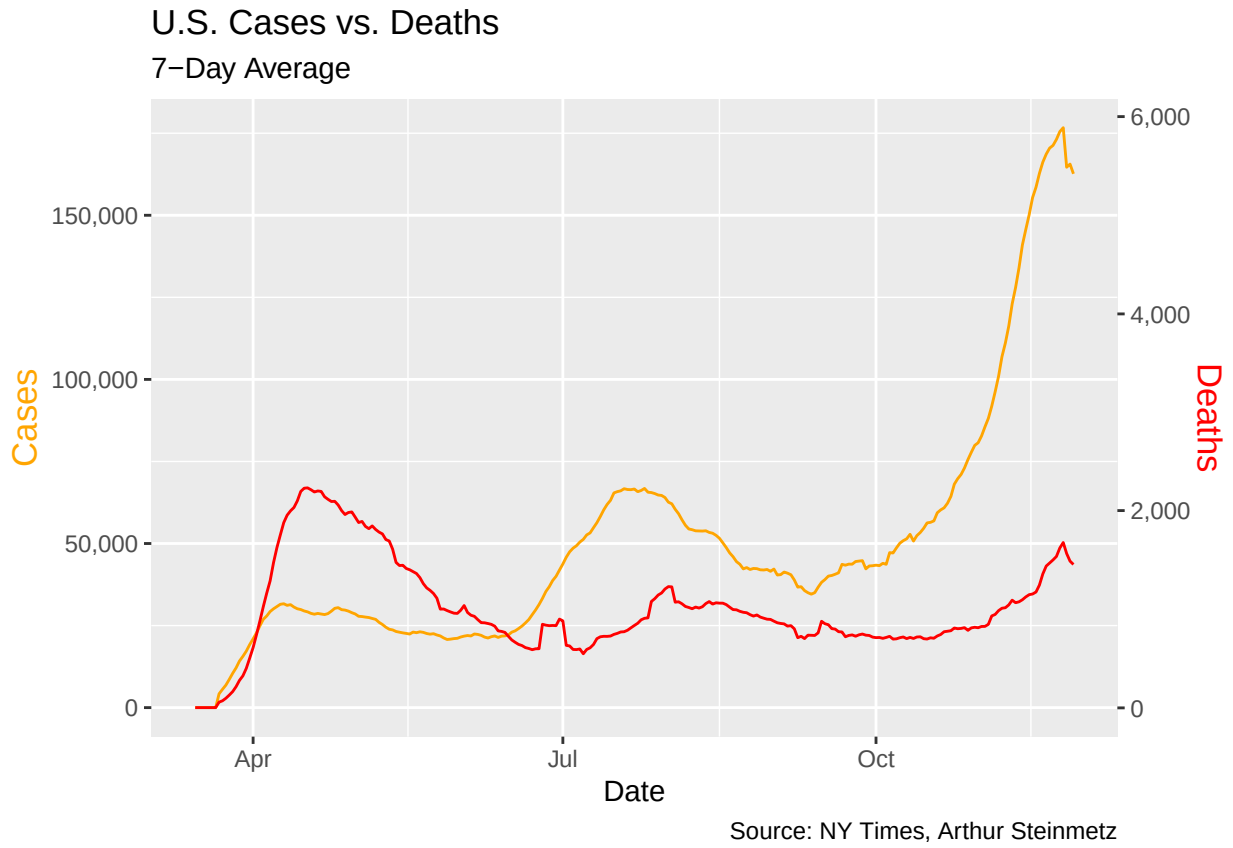
Exploratory Data Analysis

We might be tempted to simply regress deaths vs. cases but a scatter plot shows us that would not be satisfactory. As it turns out, the relationship of cases and deaths is strongly conditioned on date. This reflects the declining mortality rate as we have come to better understand the disease.



We can get much more insight plotting smoothed deaths and cases over time. It is generally bad form to use

two different y axes on a single plot but but this example adds insight. A couple of observations are obvious. First when cases start to rise, deaths follow with a lag. Second, we have had three spikes in cases so far and in each successive instance the mortality has risen by a smaller amount. This suggests that, thankfully, we are getting better at treating this disease. It is NOT a function of increased testing because positivity rates have not been falling.



This illustrates a problem for any modeling we might do.It looks like the more cases surge, the less the impact on deaths. This is NOT a valid conclusion. A simple regression of deaths vs. cases and time shows the passage of time has more explanatory power than cases in predicting deaths so we have to take that into account.

```
## # A tibble: 3 x 5
##   term           estimate   std.error statistic  p.value
##   <chr>             <dbl>       <dbl>     <dbl>    <dbl>
## 1 (Intercept) 76542.      10054.           7.61 5.15e-13
## 2 cases_7day      0.00828     0.00112      7.41 1.86e-12
## 3 date           -4.11        0.547       -7.52 9.11e-13
```

Build Some Models

We'll approach this by running regression models of deaths and varying lags (actually leads) of cases. We chose to lead deaths as opposed to lagging cases because it will allow us to make predictions about the future of deaths given cases today. We include the date as a variable as well. Once we've run regressions against each lead period, we'll chose the lead period that has the best fit (R-Squared) to the data.

The requires a lot of leads and a lot of models. Fortunately, R provides the tools to make this work very

simple and well organized. First we add new columns for each lead period using `timetk::tk_augment_lags`. This one function call does all the work but it only does lags so we have to futz with it a bit to get leads.

I chose to add forty days of leads. I don't really think that long a lead is realistic and, given the pandemic has been around only nine months, there aren't as many data points forty days ahead. Still, I want to see the behavior of the models. Once we have created the leads we remove any dates for which we don't have led deaths.

date	cases_total	deaths_total	cases_7day	deaths_7day	deaths_7day_lead0	deaths_7day_lead1
2020-03-15	3597	68	0.000	0.00000	0.00000	0.00000
2020-03-16	4504	91	0.000	0.00000	0.00000	0.00000
2020-03-17	5903	117	0.000	0.00000	0.00000	0.00000
2020-03-18	8342	162	0.000	0.00000	0.00000	0.00000
2020-03-19	12381	212	0.000	0.00000	0.00000	0.00000
2020-03-20	17998	277	0.000	0.00000	0.00000	0.00000
2020-03-21	24513	360	0.000	0.00000	0.00000	55.57143
2020-03-22	33046	457	4204.714	55.57143	55.57143	69.57143
2020-03-23	43476	578	5565.143	69.57143	69.57143	95.28571
2020-03-24	53906	784	6857.571	95.28571	95.28571	127.28571

. . . etc up to 40

Now we start the job of actually building the linear models and seeing the real power of the tidy modeling framework. Since we have our lead days in columns we revert back to long-form data. For each date we have a case count and 40 lead days with the corresponding death count. As will be seen below, the decline in the fatality rate has been non-linear, so we use a second-order polynomial to regress the `date` variable.

Our workflow looks like this:

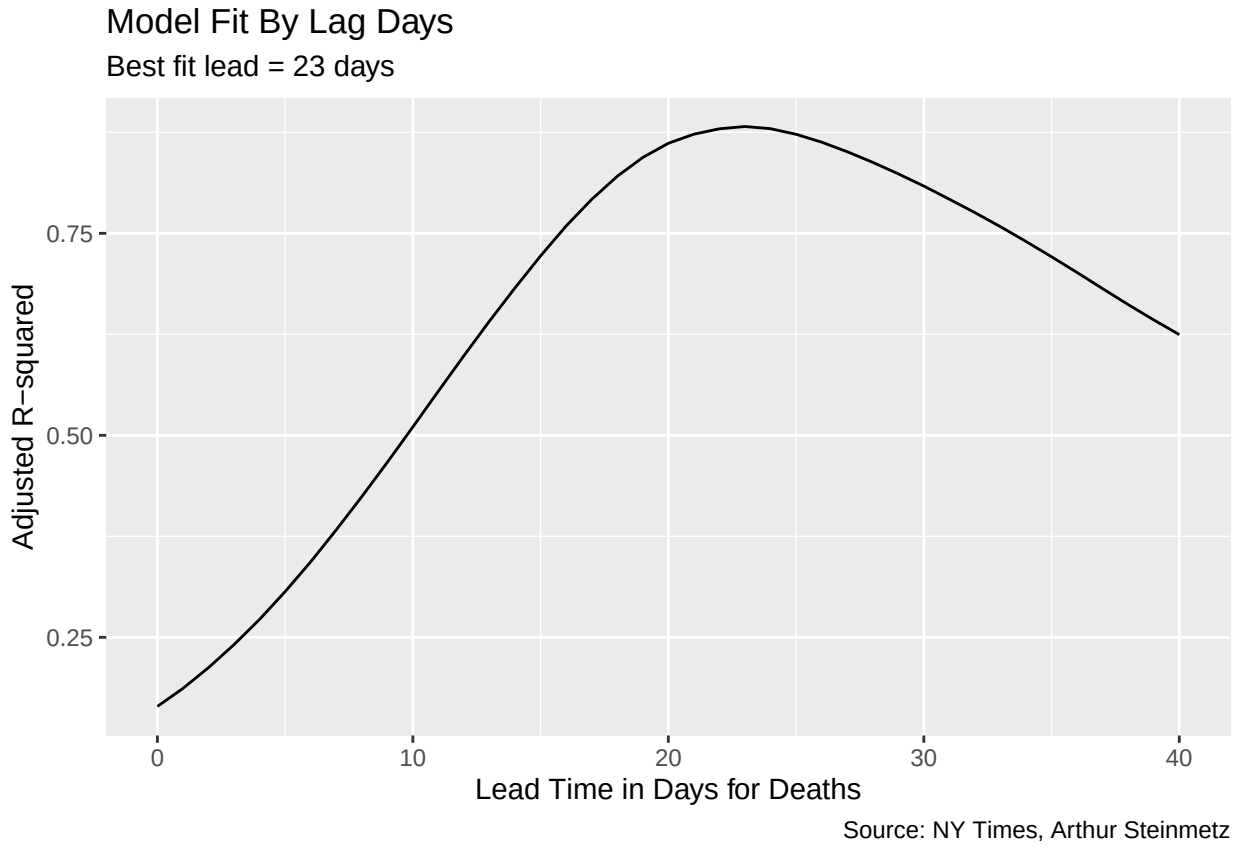
1. Create the lags using `tk_augment_lag` (above).
2. `pivot` to long form.
3. `nest` the data by lead day and state.
4. `map` the data set for each lead day to a regression model.
5. Pull out the adjusted R-Squared using `glance` for each model to determine the best fit lead time.

The result is a data frame with our lead times, the nested raw data, model and R-squared for each lead time.

```
## # A tibble: 41 x 4
##     lead data              model  adj_r
##    <dbl> <list>            <list> <dbl>
##  1     0 <tibble [218 x 3]> <lm>   0.164
##  2     1 <tibble [218 x 3]> <lm>   0.187
##  3     2 <tibble [218 x 3]> <lm>   0.212
```

```
##  4     3 <tibble [218 x 3]> <lm>   0.241
##  5     4 <tibble [218 x 3]> <lm>   0.272
##  6     5 <tibble [218 x 3]> <lm>   0.307
##  7     6 <tibble [218 x 3]> <lm>   0.343
##  8     7 <tibble [218 x 3]> <lm>   0.383
##  9     8 <tibble [218 x 3]> <lm>   0.424
## 10     9 <tibble [218 x 3]> <lm>   0.467
## # ... with 31 more rows
```

To decide the best-fit lead time we choose the model with the highest R-squared.

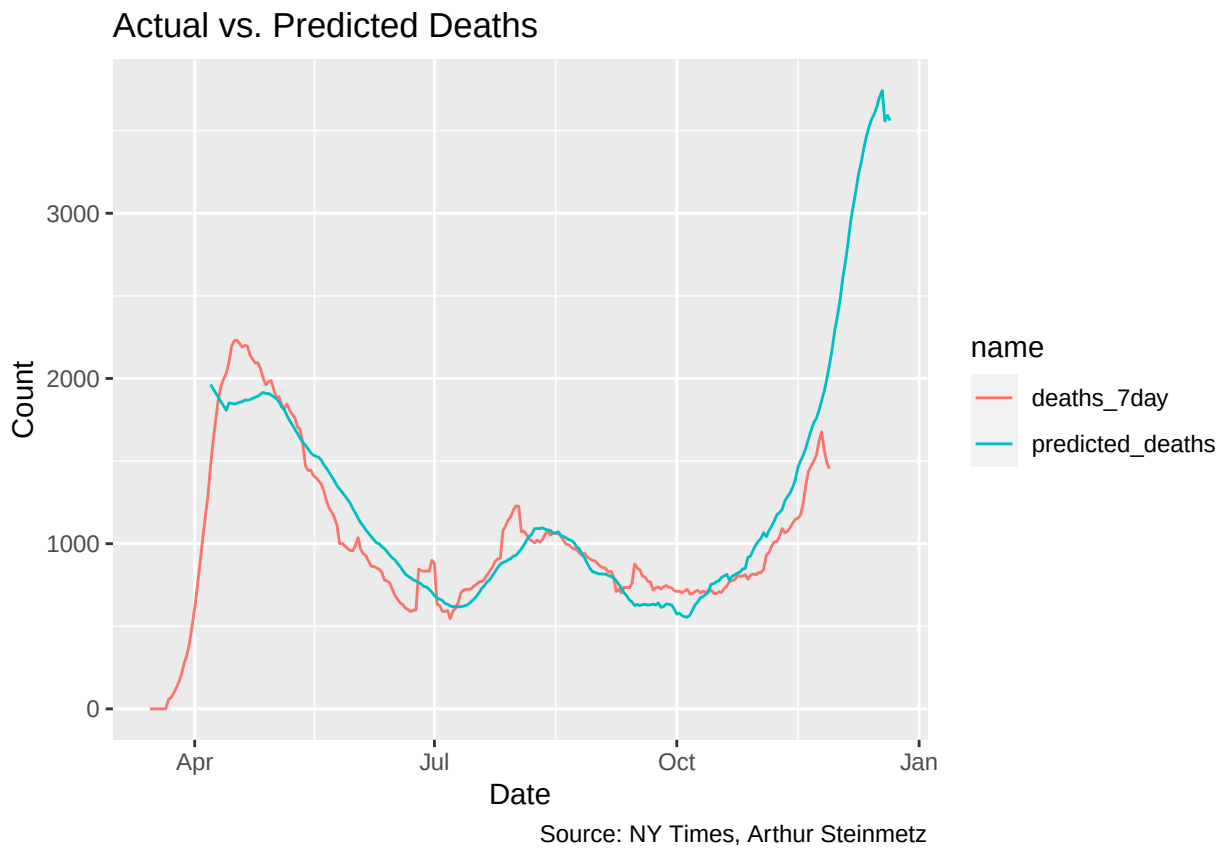


We can have some confidence that we are not overfitting the `date` variable because the significance of the case count remains. With a high enough degree polynomial on the `date` variable, cases would vanish in importance.

```
## # A tibble: 4 x 5
##   term             estimate  std.error statistic  p.value
##   <chr>               <dbl>      <dbl>     <dbl>    <dbl>
## 1 (Intercept)      436.      38.0           11.5 4.21e-24
## 2 cases_7day         0.0167   0.000993      16.8 5.45e-41
## 3 poly(date, 2)1 -7306.     227.           -32.2 5.87e-84
## 4 poly(date, 2)2  4511.     167.            26.9 1.02e-70
```

Make Predictions

The best-fit lead time is 23 days but let's use `predict` to see how well our model fits to the actual deaths.



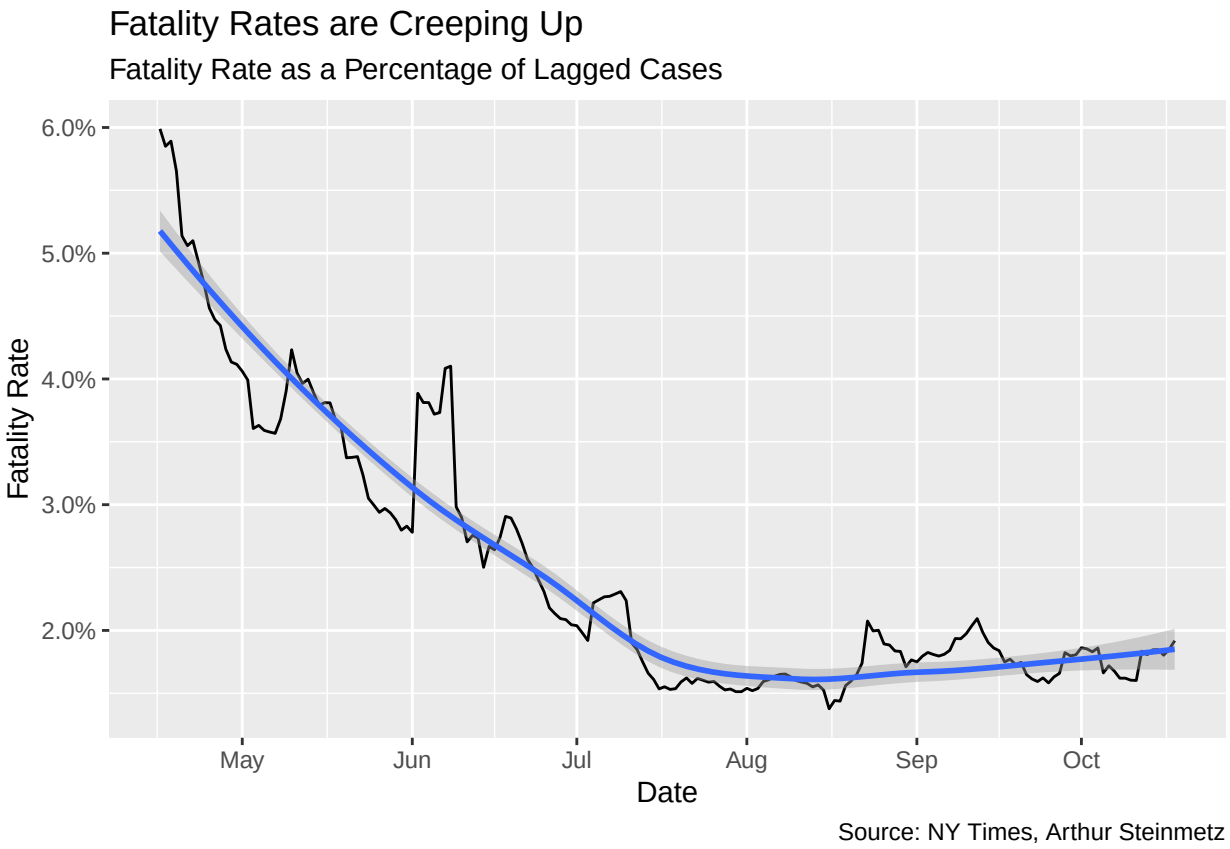
This is a satisfying result, but sadly shows deaths about to spike. This is despite accounting for the improvements in treatment outcomes we've accomplished over the past several months. The 23-day lead time model shows a 1.7% mortality rate over the whole length of observations but conditioned on deaths falling steadily over time.

Declining Mortality Rate

Once we've settled on the appropriate lag time, we can look at the fatality rate per identified case. This is but one possible measure of fatality rate, certainly not THE fatality rate. Testing rate, positivity rate and others variables will affect this measure. We also assume our best-fit lag is stable over time so take the result with a grain of salt. The takeaway should be how it is declining, not exactly what it is.

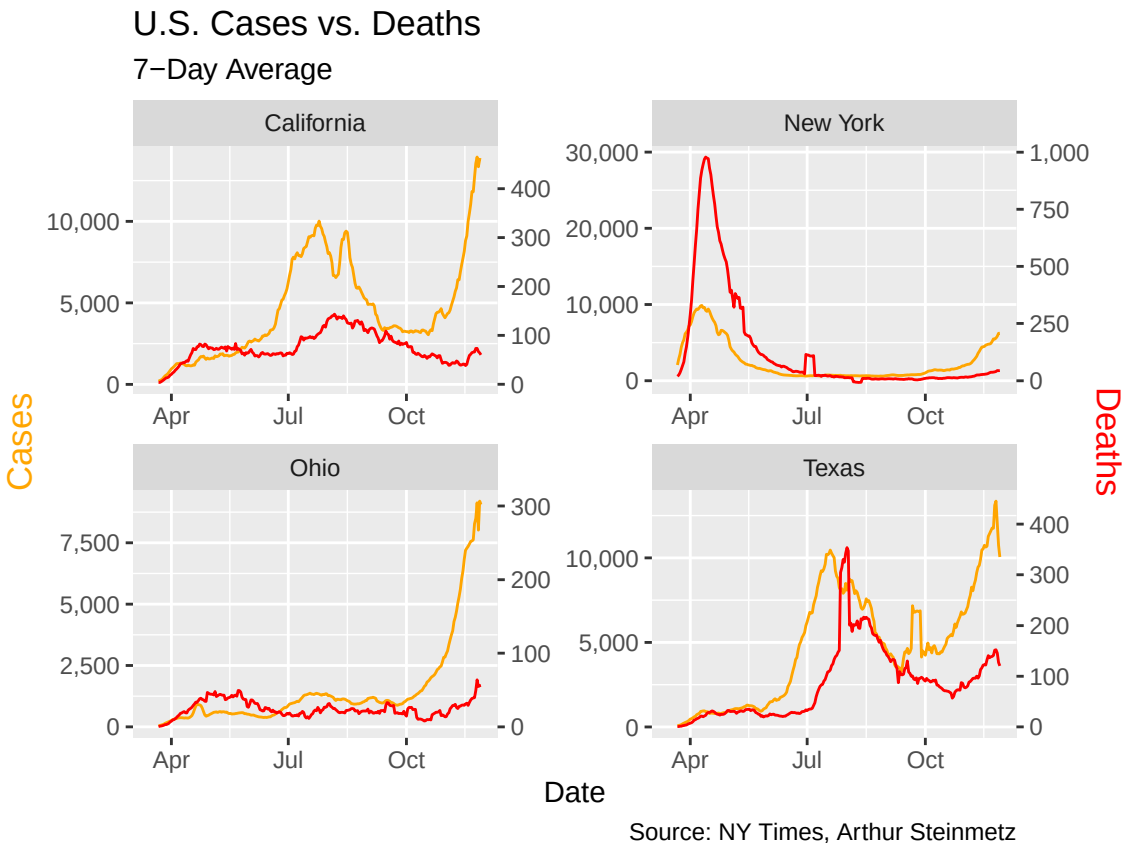
Early on, only people who were very sick or met strict criteria were tested so, of course, fatality rates (on this metric) were much, much higher. To minimize this we start our measure at the middle of April.

Sadly, we see that fatality rates are creeping up again.



State-Level Analysis

One problem with the national model is each state saw the arrival of the virus at different times, which suggests there might also be different relationships between cases and deaths. Looking at a few selected states illustrates this.



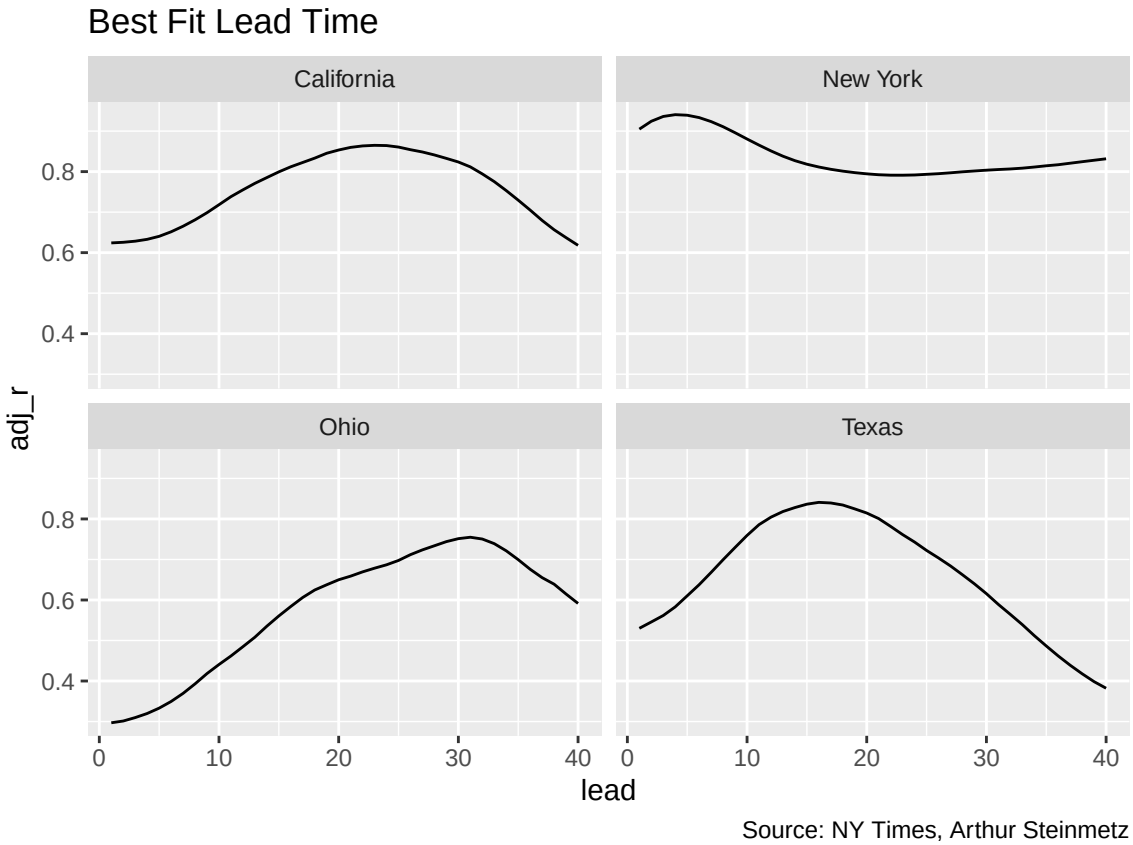
In particular we note New York, where the virus arrived early and circulated undetected for weeks. Testing was rare and we did not know much about the course of the disease so the death toll was much worse. Tests were often not conducted until the disease was in advanced stages so we would expect the lag to be shorter.

In Texas, the virus arrived later. There it looks like the consequences of the first wave were less dire and the lag was longer.

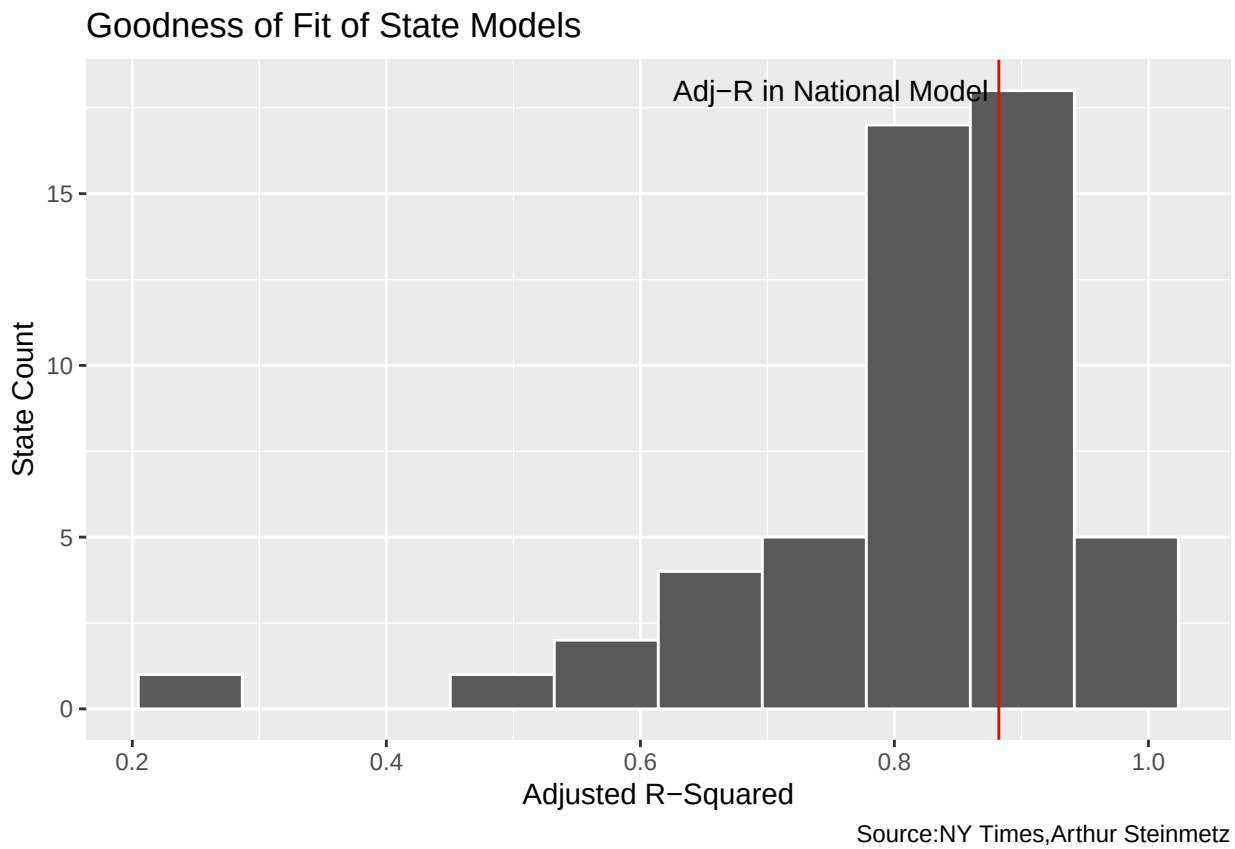
Run State Models

Now we can run the same workflow we used above over the state-by-state data. Our data set is much larger because we have a full set of lags for each state but building our data frame of list columns is just as easy.

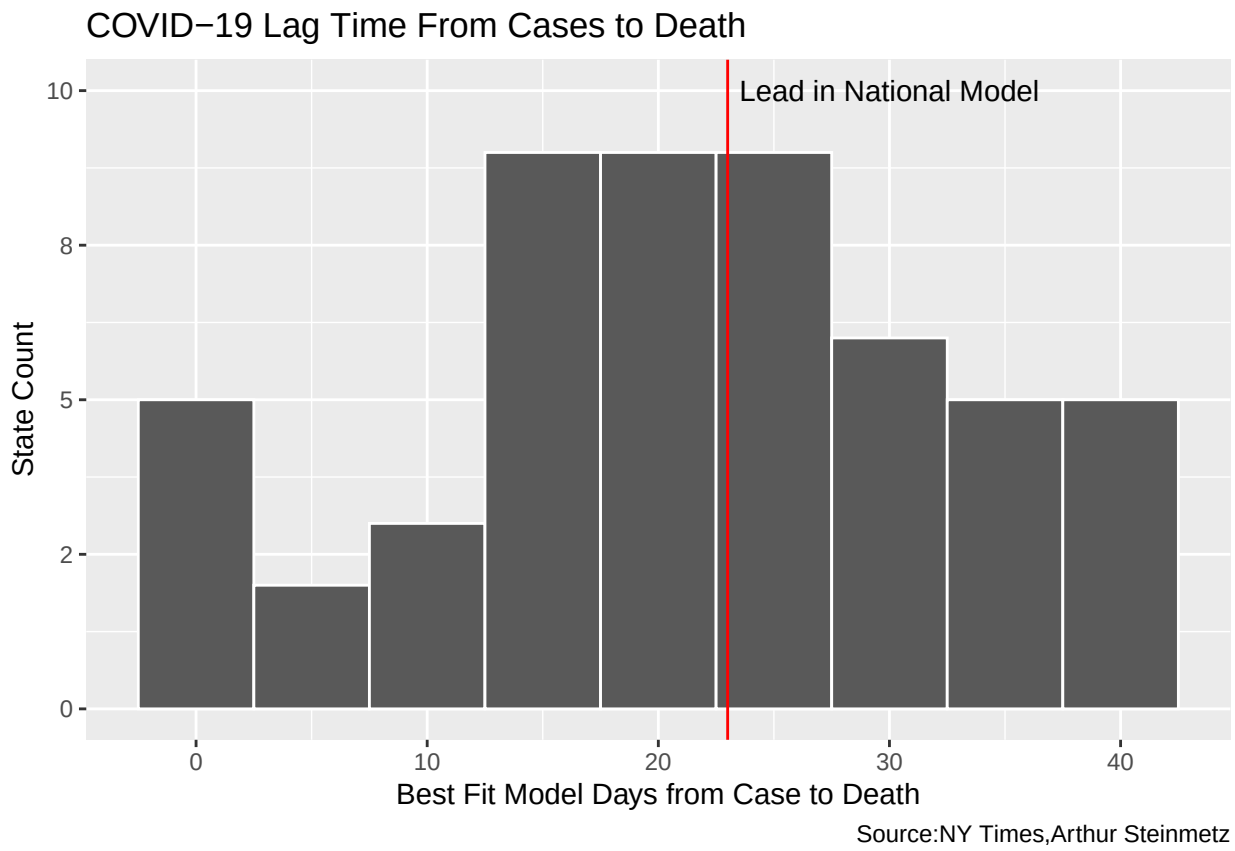
Looking at the lags by state shows similar results to the national model, on average, as we assume, but the dispersion is large. Early in the pandemic, in New York, cases were diagnosed only for people who were already sick so the lead time before death was much shorter.



To see how the fit looks for the data set as a whole we look at a histogram of all the state R-squareds. We see many of the state models have a worse accuracy than the national model.



There are vast differences in the best-fit lead times across the states but the distribution is in agreement with our national model.



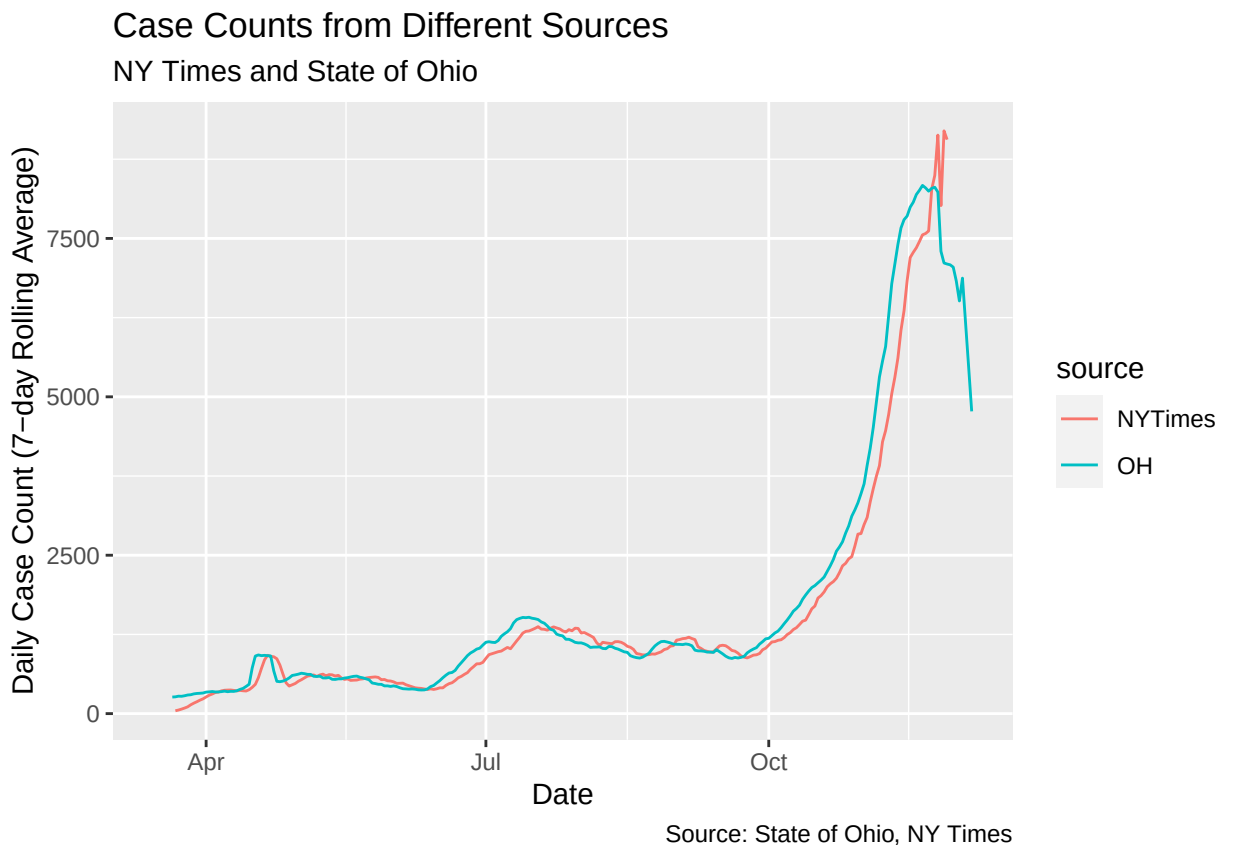
Validate with Individual Case Data from Ohio

This whole exercise has involved proxying deaths by time and quantity of positive tests. Ideally, we should look at longitudinal data which follows each individual. The state of Ohio provides that so we'll look at just this one state to provide a reality check on the foregoing analysis. In our proxy model, Ohio shows a best-fit lead time of 31 days, which is much longer than our national-level model.

state	adj_r	lead
Ohio	0.7548416	31

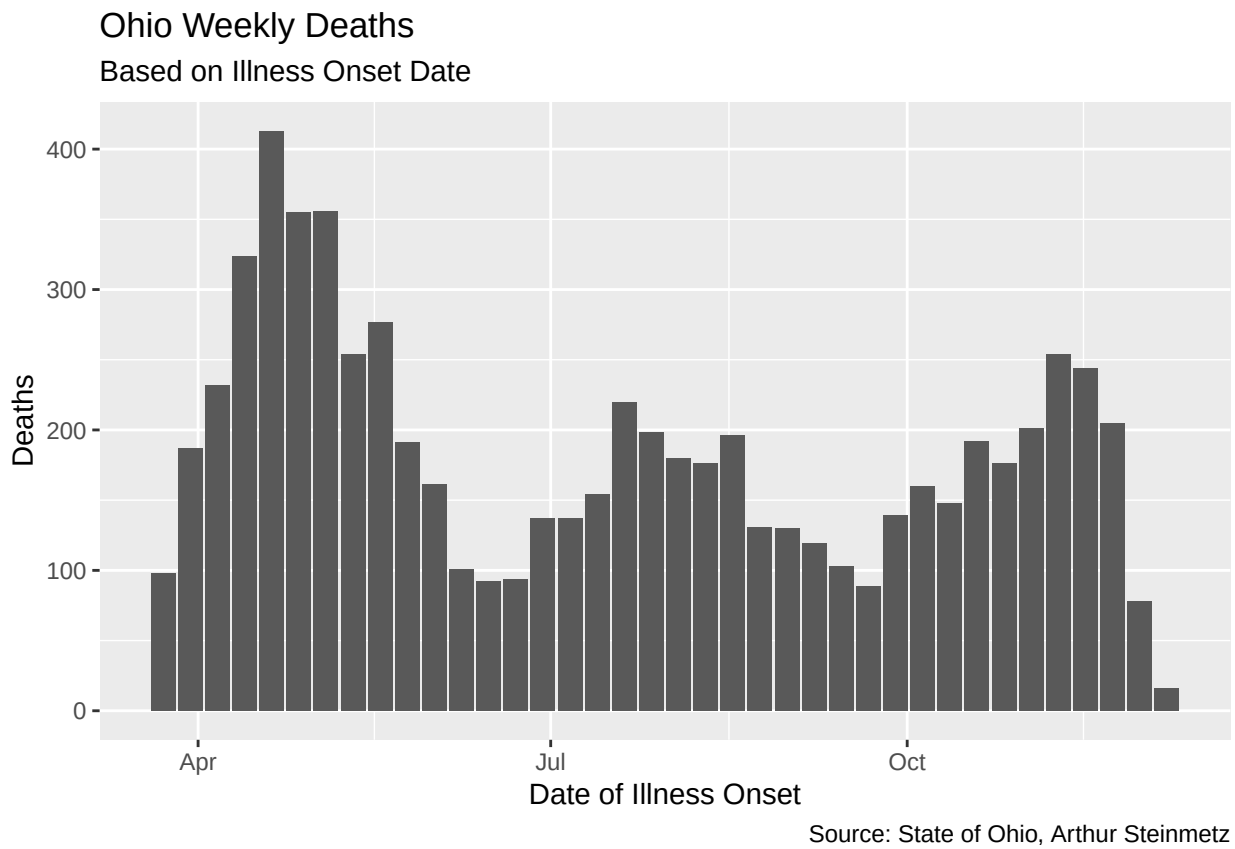
The caveat here is the NY Times data uses the "case" date which is presumably the date a positive test is recorded. The Ohio data uses "onset" date, which is the date the "illness began." That is not necessarily the same as the test date.

How comparable are these data sets? Let's compare the NY Times case count and dates to the Ohio "Illness Onset" dates.

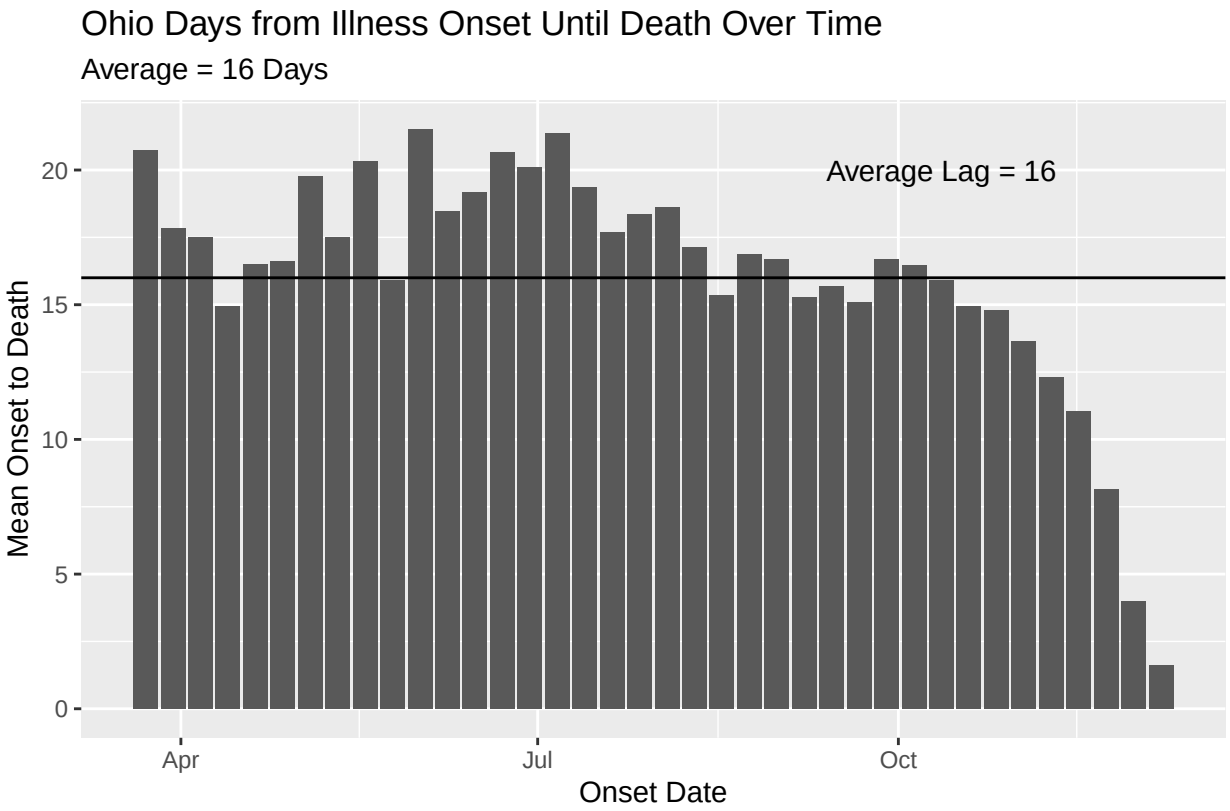


We clearly see the numbers line up almost exactly but the Ohio data runs about 4 days ahead of the NY Times data.

For each individual death, we subtract the onset date from the death date. Then we aggregate the county-level data to statewide and daily data to weekly. Then take the weekly mean of deaths.

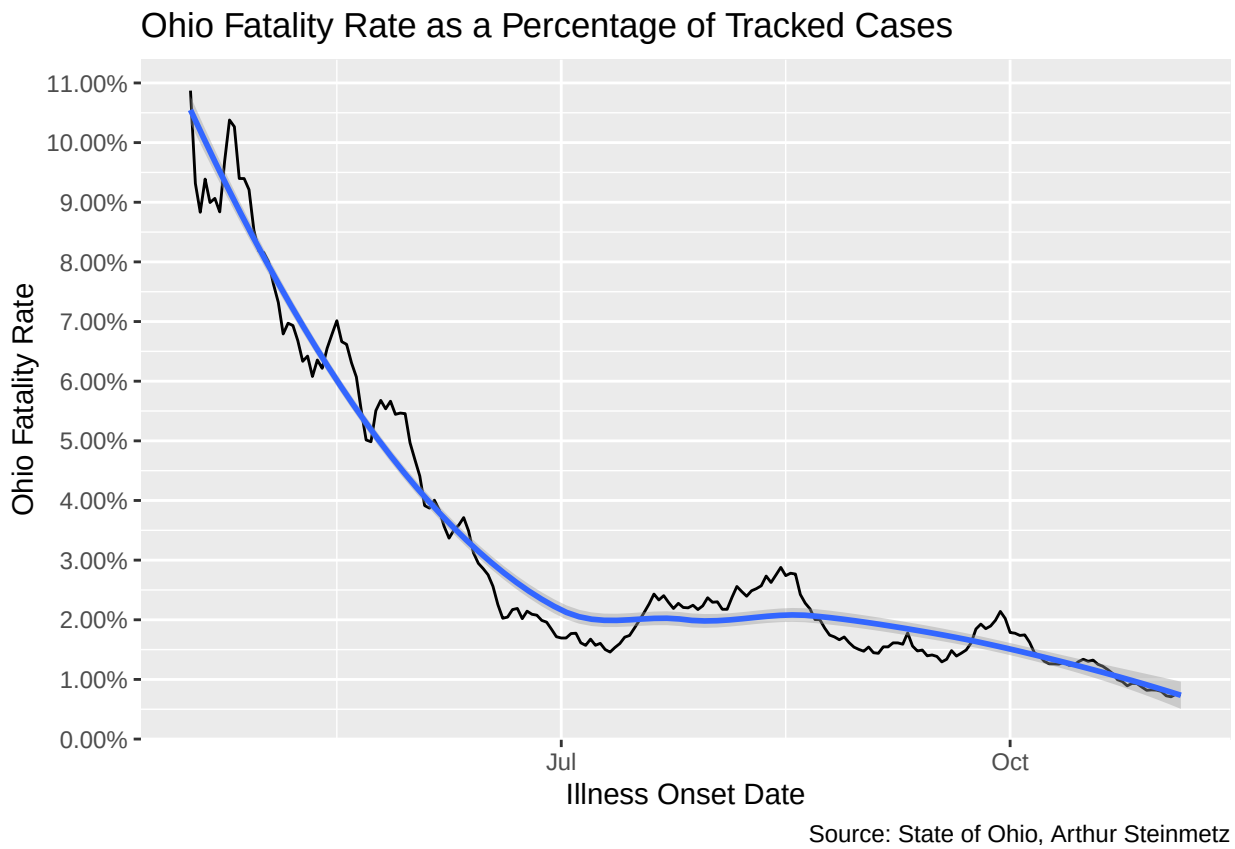


When we measure the average lag, we find that it has been fairly stable over time in Ohio. Unfortunately, it differs substantially from our proxy model using untracked cases.



Note the drop off at the end of the date range. This is because we don't yet know the outcome of the most recently recorded cases. Generally, while we have been successful in lowering the fatality rate of this disease, the duration from onset to death for those cases which are fatal has not changed much, at least in Ohio.

Since we have the actual number of deaths associated with every onset date we can calculate the "true" fatality rate. As mentioned, the fatality rate of the more recent cases is not yet known. Also the data is too sparse at the front of the series so we cut off the head and the tail of the data.



The fatality rate in Ohio seems to have been worse than our national model but it is coming down. Again, this result comes from a different methodology than our proxy model.

Conclusion

Among the vexing aspects of this terrible pandemic is that we don't know what the gold standard is for treatment and prevention. We are learning as we go. The good news is we ARE learning. For a data analyst the challenge is the evolving relationship of of all of the disparate data. Here we have gotten some insight into the duration between a positive test and mortality. We can't have high confidence that our proxy model using aggregate cases is strictly accurate because the longitudinal data from Ohio shows a different lag. We have clearly seen that mortality has been declining but our model suggests that death will nonetheless surge along with the autumn surge in cases.

What are the further avenues for modeling? There is a wealth of data around behavior and demographics with this disease that we don't fully understand yet. On the analytics side, we might get more sophisticated with our modeling. We have only scratched the surface of the `tidymodels` framework and we might apply fancier predictive models than linear regression. Is the drop in the fatality rate we saw early in the pandemic real? Only people who were actually sick got tested in the early days. Now, many positive tests are from asymptomatic people. Finally, the disagreement between the case proxy model and the longitudinal data in Ohio shows there is more work to be done.