A comparison of models for count data as applied to COVID-19 deaths and air pollution

Lauren Hund

May 11, 2020

1 Introduction

I decided to look at the study data linking air pollution and COVID-19 deaths [3] after reading about the results in the Washington Post. From a link in the article, I saw that the data were publicly available, so I downloaded them, with the goal of looking at whether the study results were sensitive to the form of the linear predictor. My computer crashed when I tried to fit the zero-inflated, negative binomial, mixed effects model. So, I decided to fit a simpler model - a quasi-poisson model - as a quick approximation. A description of my analysis and the results follow.

2 Methods

I attempted to reproduce the analysis from April 22, 2020 and fit some additional models.

Dataset. I could not reproduce the dataset exactly. To create the dataset, I used the preprocessing code from May 6, 2020, which has been changed since the April 22, 2020 version. I tweaked the code to try to create an April 22 dataset. My dataset is close to the April 22 data, but I cannot match the population numbers, death rate, or number of counties. I have 3,089 counties in my analysis, instead of 3,087 (I used the version of the code that combines the boroughs of New York City into one group). There are 45,856 COVID-19 deaths in my dataset, which is slightly higher than the April 22 number (45,817), presumably because either the data source has been updated or because I have a few extra counties.

Linear predictor I attempted to match the main model in the analysis. Specifically, I fit generalized linear models using a log-link function with an offset for the log-population. I included the following covariates: PM 2.5 (ug/m3), quintiles of population density, percent poverty, log-median house value, log-median household income, percent owner occupied housing, percent less than high school education, percent black, percent hispanic, 3 age groups, days since stay at home order (I think I found a typo in the code for this variable - Sys.date should be date_of_study), days since first case, rate of hospital beds, percent obese, percent smokers, average summer temperature, average winter temperature, average summer relative humidity, average winter relative humidity.

Likelihoods. I compare two different 'likelihoods' for count data - the negative binomial and Poisson. The negative binomial is convenient because you can include both overdispersion and zero-inflation at the same time. With the Poisson distribution, you are stuck with one or the other (as far as I know) - either overdispersion or zero-inflation. So, I compared two different sets of models: (1) zero-inflated negative binomial and zero-inflated Poisson; and (2) negative binomial and quasi-Poisson. For the zero-inflated models, I only look at the count portion of the model, not the systematic zeros model.

Sensitivity analyses. To address the fact that the Poisson model is limited to either overdispersion (quasi-Poisson) or zero-inflation, I looked at different subsets of the data. (1) I omitted New York state; (2) I omitted counties with 0 deaths; (3) I omitted both counties with 0 deaths and counties in New York State; (4) I omitted counties with fewer than 10 deaths and New York state. I hypothesize that a quasi-Poisson model fit to the data without counties with 0 counts should produce reasonably similar results to a zero-inflated model.

Random effects. I do not include any models with state-specific random effects for four reasons: (1) my laptop crashed when I tried to fit the mixed effects model; (2) if the study results are robust to spatial confounding, then the study results should not substantively change without random effects, [1]; (3) the point estimate from a model without random effects should correspond to a GEE estimate; and (4) my results without random effects are quite close to the April 22 study results, suggesting state specific random effects do not change the results too much.

Uncertainty. I do not estimate uncertainty in mortality rate ratios, because I am only looking at sensitivity of the point estimate, not uncertainty.

3 Results

My attempt to reproduce the demographics table is in Table 1. My population numbers clearly do not match, which are causing some discrepancies in the demographics.

Using a Poisson model (either quasi-poisson or zero-inflated Poisson), I find that the direction of the effect flips. Regression coefficients for the zero-inflated Poisson and negative binomial models (compared to the April 22 analysis) are presented in Table 2. My coefficients in Table 2 are based on the data with New York excluded. The estimated mortality rate ratio is consistent across the different models and subsets of data that I considered (Table 3). I also looked at using quintiles of pollution as a predictor (results not shown), and there is no evidence of a clear trend in mortality rates with pollution quintiles.

In terms of model selection, I do not see an obvious choice between the Poisson and negative binomial (aside from the convenience of the negative binomial likelihood, since it can accommodate both overdispersion and zero inflation). In all models that were fit, the Poisson had smaller residuals, on average, than its negative binomial counterpart (results not shown); but this result is not interesting, since this result is essentially a consequence of choosing the Poisson model which places more weights on areas with higher counts, as described in [2]. I compared the quasi-Poisson and negative binomial distributions using the model fit diagnostics proposed in [2]. Specifically, I examined model fit by plotting the residuals (Figure 1), the predicted variance of the residuals as a function of the mean (Figure 2), and the estimated weights as a function of the fitted mean (Figure 3). To make these plots, I again used the dataset without New York and fit models without zero-inflation.

Ultimately, I do not find any evidence in the data for picking one model over the other. The negative binomial model places rather little weight on counties with large counts. It is unclear if this property of the negative binomial is desirable in this application.

4 Conclusion

This analysis suggests the April 22 study results are potentially sensitive to selection of the model likelihood. The analysis is limited by two factors: (1) I cannot reproduce the April 22 data exactly and (2) there is not a Poisson model that can handle both overdispersion and zero-inflation.

References

[1] James S Hodges and Brian J Reich. Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4):325–334, 2010.

[2] Jay M Ver Hoef and Peter L Boveng. Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772, 2007.

[3] Xiao Wu, Rachel C Nethery, Benjamin M Sabath, Danielle Braun, and Francesca Dominici. Exposure to air pollution and covid-19 mortality in the united states. *medRxiv*, 2020.

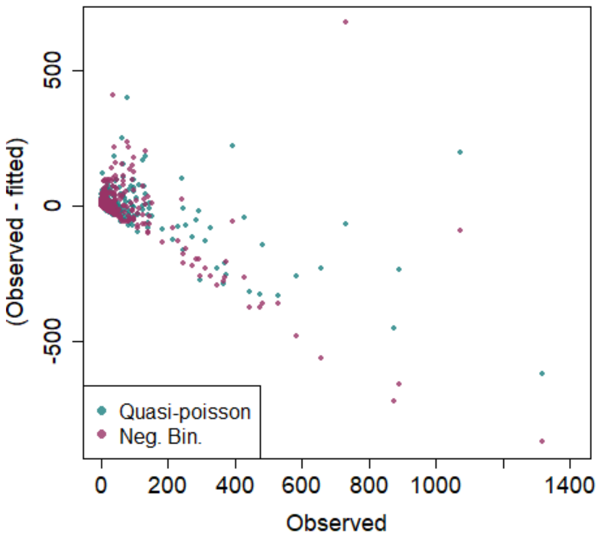

Figure 1: Observed deaths versus model residuals (observed − fitted).

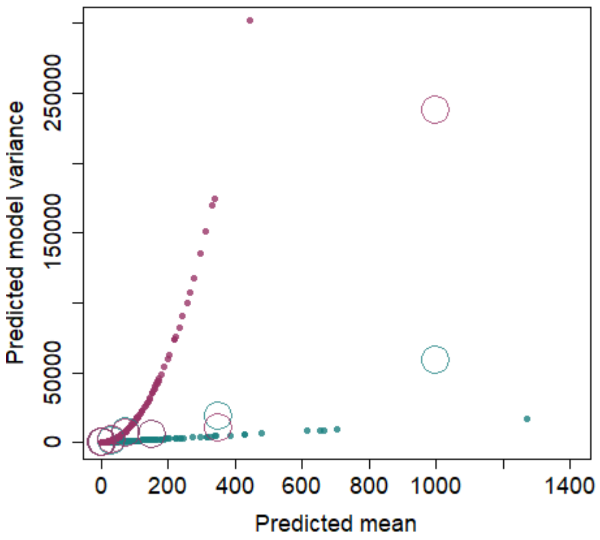

Figure 2: Estimated variance of outcome as a function of fitted mean (solid points, blue is quasi-Poisson, red is negative binomial). Hollow circles are empirical variances of the model residuals, binned into different groups.

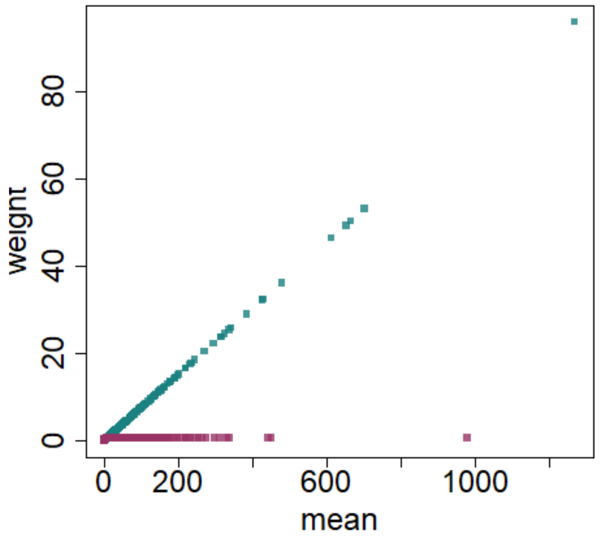

Figure 3: Estimated weights as a function of fitted mean (blue is quasi-Poisson, red is negative binomial).

	Mean	SD	Mean	SD
COVID-19 death rate (per 100,000)	3.40	10.60	4.50	12.80
Average PM2.5 (ug/m3)	8.40	2.50	8.40	2.50
Rate of hospital beds (per 100,000)	242.00	391.90	300.80	428.40
Days since first case	23.60	10.70	23.60	10.70
Days since stay-at-home order	18.30	12.40	17.60	12.00
Pct Smokers	17.40	3.50	17.40	3.50
Pct Obese	32.90	5.40	32.90	5.40
Pct In poverty	10.50	5.70	10.50	5.90
Pct Less than high school education	21.20	10.40	21.30	10.70
Pct Owner-occupied housing	74.20	8.80	75.00	8.30
Pct Hispanic	7.60	12.30	7.50	12.30
Pct Black	8.20	14.20	8.00	14.10
Pct >=65 years of age	16.00	4.10	16.00	4.10
Pct 45-64 years of age	26.40	3.00	26.40	3.00
Pct 15-44 years of age	37.60	6.50	37.60	6.50
Population density (person/sq. mi.)	406.70	1732.60	387.20	1810.90
Median household income ($1,000)	49.00	13.10	49.30	13.40
Median house value ($1,000)	136.00	89.40	135.70	89.90
Average summer temperature (F)	86.00	5.70	86.00	5.70
Average winter temperature (F)	45.10	11.90	45.10	11.90
Average summer relative humidity (Pct)	89.00	9.60	89.00	9.60
Average winter relative humidity (Pct)	87.50	4.80	87.50	4.80

Table 1: Mean and standard deviation of variables in analysis for original data (left) and my attempt to recreate the data (right).

	Poisson	Negative Bin	Original
PM 2.5 (ug/m3)	0.92	1.07	1.08
Population density (Q2)	1.21	0.97	0.86
Population density (Q3)	0.84	0.62	0.58
Population density (Q4)	0.75	0.49	0.47
Population density (Q5)	1.01	0.63	0.52
Pct Poverty	1.28	1.24	1.02
log(Median house value)	1.07	0.80	1.17
log(Median household income)	1.84	1.70	1.28
Pct Owner-occupied housing	0.91	1.04	1.12
Pct Less than high school education	1.58	1.05	1.36
Pct Black	1.56	1.46	1.45
Pct Hispanic	1.10	1.07	1.00
Pct =65 years of age	1.30	1.04	1.15
Pct 15-44 years of age	0.56	0.86	0.93
Pct 45-64 years of age	0.94	1.01	0.96
Days since stay-at-home order	1.26	1.27	1.28
Days since first case	2.75	3.39	2.96
Rate of hospital beds	0.72	0.97	1.12
Pct Obese	0.83	0.86	0.94
Pct Smokers	1.51	1.30	1.08
Average summer temperature (F)	1.02	0.83	0.96
Average winter temperature (F)	0.53	1.07	1.18
Average summer relative humidity (Pct)	1.24	1.02	0.84
Average winter relative humidity (Pct)	0.84	0.92	1.00

Table 2: Mortality rate ratios for all variables in the main analysis; my zero-inflated models using the Poisson and negative binomial are in the first two columns (fit to data excluding NY state); the original results from April 22 are in the far right column.

Analysis	Poisson	NegBin
Zero: all	0.92	1.07
Zero: no NY	0.92	1.07
All	0.92	1.09
No NY	0.93	1.08
> 0	0.92	1.01
>0 & no NY	0.92	1.00
>10, no NY	0.88	0.86

Table 3: Sensitivity analysis comparing models on different subsets of data.