Data Precision During the Pandemic: Evaluating Normalization Approaches for Reliable COVID-19 Quantification in Wastewater

2023-04-17

Authors list: Marlin Lee, Dagmara Antkiewicz, Adélaïde Roguet, Kayley Janssen, Martin Shafer, Abe Megahed, Kyllan Wunder

Abstract

The COVID-19 pandemic has presented unprecedented challenges to global public health, requiring rapid and accurate monitoring strategies. Wastewater-based epidemiology (WBE) has emerged as a promising approach for early community detection of SARS-CoV-2, the virus responsible for COVID-19. However, challenges remain in using the measured wastewater concentrations to reliably predict new outbreaks or call significant community viral load changes. The intrinsic complexity and variability of wastewater surveillance data necessitates the use of normalization approaches including viral markers, laboratory controls, and general collection source factors to establish and improve the connection between wastewater concentration and community health / trends. In this study, we investigated various characteristics of the collected samples, including human fecal marker (PMMoV) concentration, reported measurement LOD (limit of detection), and laboratory viral recovery control (BCoV), in an effort to reduce variance in the SARS-CoV-2 measurement and to enhance its connection to collected case data. Additionally, we explored the information about the site's population, location, and current dominant SARS-CoV-2 variant as predictors of the relationship between wastewater concentration and case positivity. Our findings revealed that the linear relationship between the log of sewage concentration of SARS-CoV-2 and the log of reported cases was significantly improved when controlling for the site's population and LOD. These findings contribute to our understanding of the factors that can impact the accuracy and precision of wastewater-based epidemiology for COVID-19 detection. By identifying and incorporating appropriate normalizers, we can enhance the accuracy and precision of wastewater-based epidemiology, thus facilitating early detection and response to potential outbreaks, and supporting effective public health interventions in the ongoing fight against the pandemic. Further research in this area is warranted to optimize the use of normalizers and improve the precision of wastewater-based COVID-19 detection methods.

Background

As epidemiologists continue to study the Covid-19 pandemic, caused by the novel coronavirus SARS-CoV-2, there is growing interest in exploring unconventional surveillance methods to track the spread of the virus in communities. One such approach gaining attention is WBE, also known as wastewater surveillance or sewage surveillance. WBE involves monitoring and analyzing wastewater samples to detect the presence of SARS-CoV-2, which has been found in human feces.

The concept behind WBE is that infected individuals shed the virus in their feces, which then enters the wastewater system and can be detected in sewage. By analyzing wastewater samples for viral RNA, epidemiologists can estimate the prevalence of Covid-19 in a community, even among asymptomatic individuals who may not have been tested through traditional clinical testing methods.

One of the key advantages of wastewater surveillance is its ability to provide population-level data, as wastewater samples can be collected from sewage treatment plants or other collection points that serve large populations. This makes it a cost-effective and non-invasive method for monitoring the overall prevalence of Covid-19 in a community, especially in areas with limited testing resources or where asymptomatic cases may be missed through clinical testing alone. However, it's important to note that there are issues to using wastewater surveillance for Covid-19 monitoring.The accuracy and sensitivity of wastewater surveillance can also be influenced by various factors such as sampling protocols, wastewater treatment processes, and environmental factors, which may introduce variability in the results.

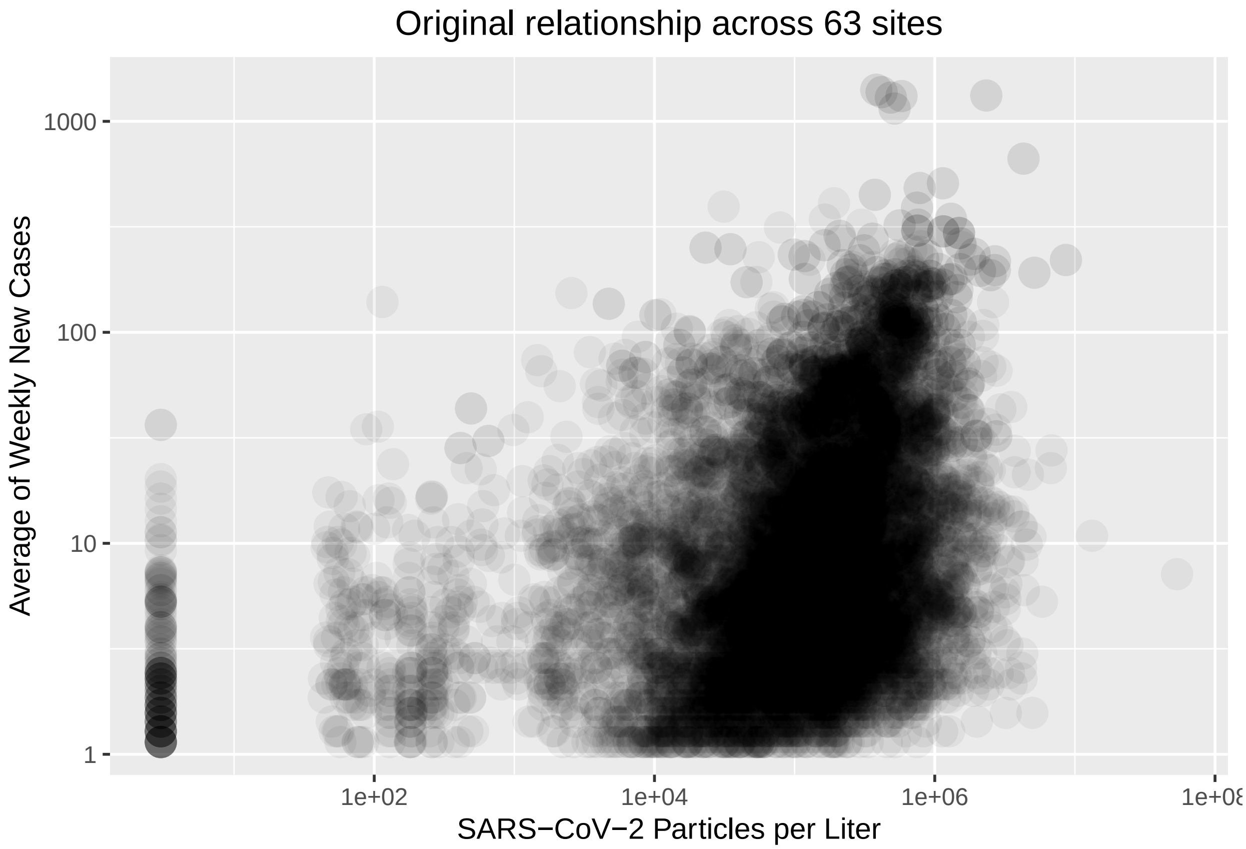
Data Collection:

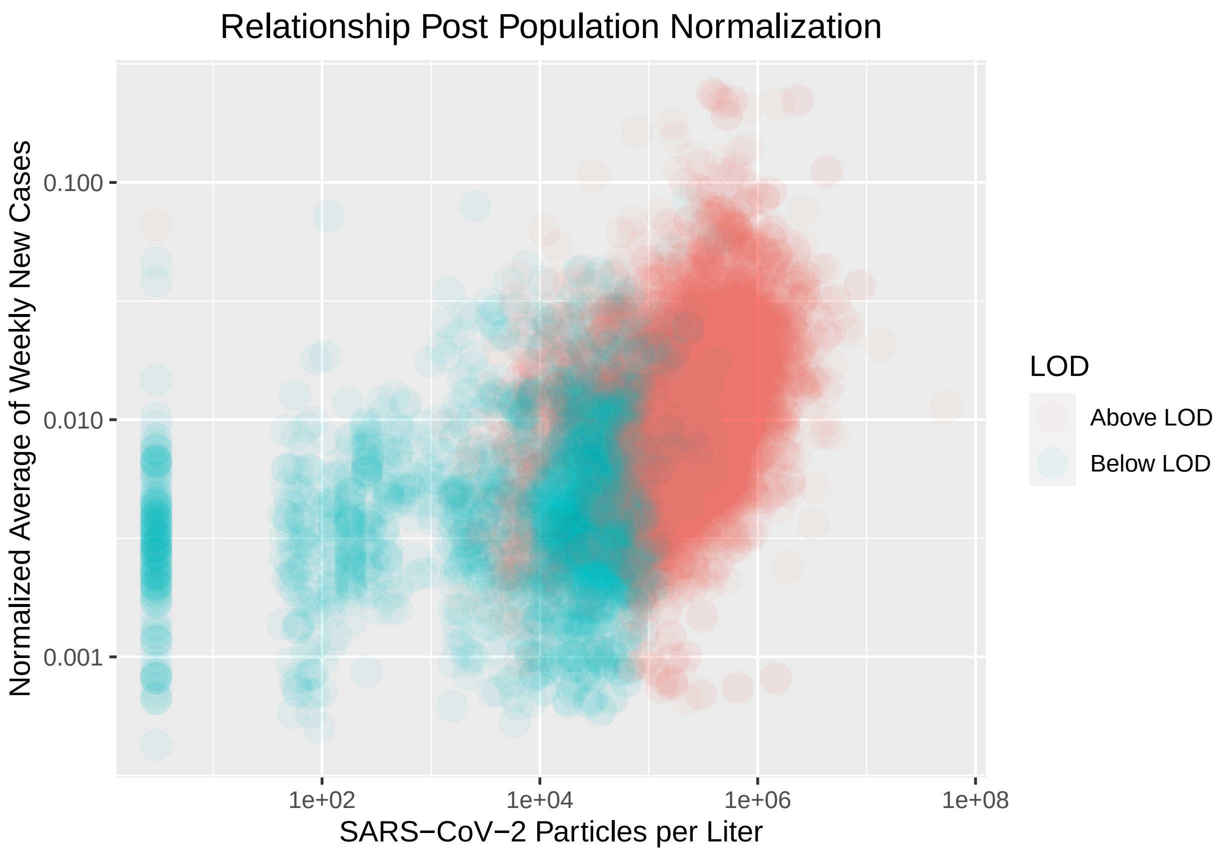
Wastewater samples were given to us by the Department of Health Services(DHS) and curated by the State Lab of Hygiene (SLH). These samples were collected from sewage treatment plants in multiple locations across Wisconsin over a period of 2 years from September of 200 to December of 2022. Samples were collected at regular intervals and processed according to established protocols for WBE. Specifically, samples were analyzed for the concentration of the SARS-CoV-2 gene's N1 and N2. The sample was also measured for a human fecal marker, PMMoV. The Level of Detection (LOD) was determined by the SLH and reported for each concentration. Laboratory viral recovery control was performed using bovine coronavirus (BCoV) as a surrogate virus, added to the wastewater samples at a known concentration during the processing step, to assess the efficiency of viral recovery and potential inhibitory effects in the assay. Clinical data on reported Covid-19 cases in the corresponding geographic locations were obtained from DHS.

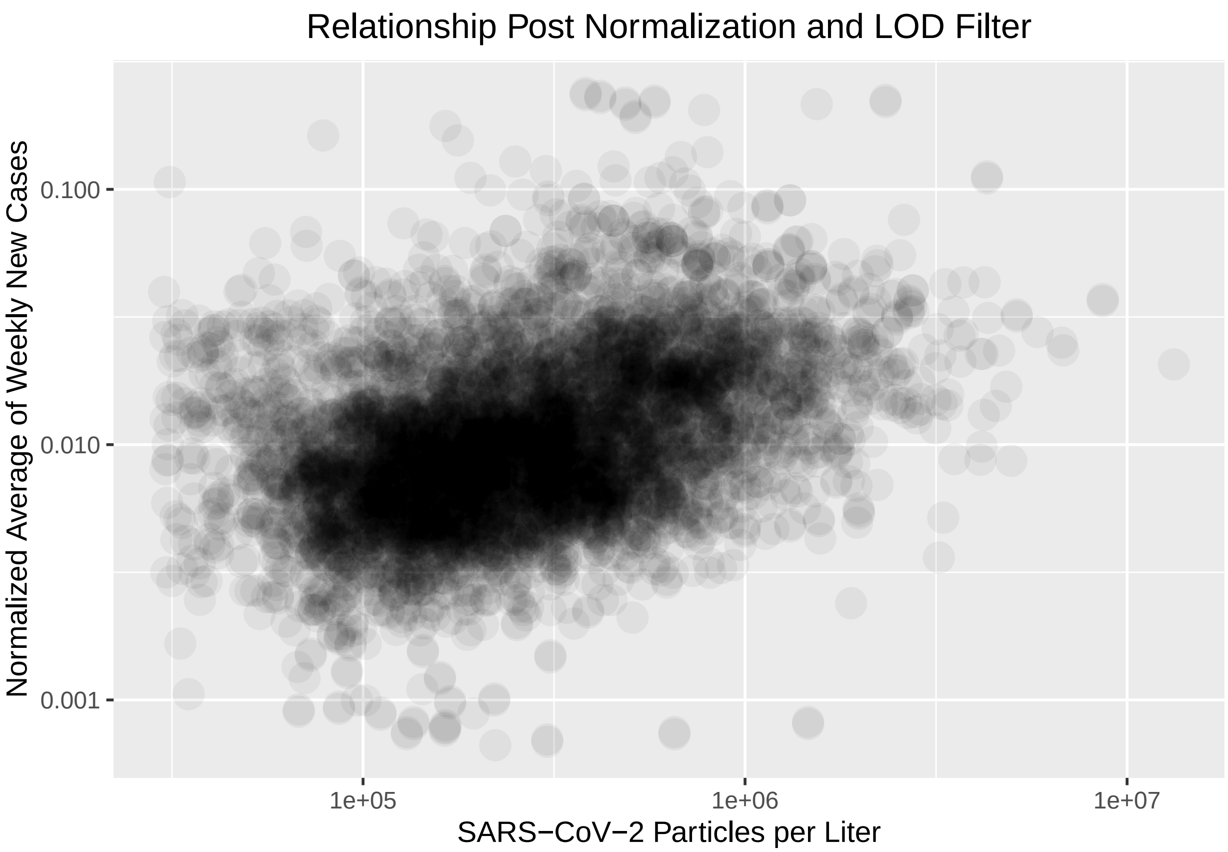
Data Analysis

Our aim is to comprehensively capture the linear relationship between the two signals under investigation. To assess the strength of the linear trend after data modification, we employed linear regression and evaluated the impact of normalization and filtering on the relationship. We utilized the R^2 metric to gauge the quality of the model fit. Additionally, we employed random forest to capture the potential role of covariates on the reported errors to the fullest extent possible. The impact of these covariates on the relationship was visualized through partial plots generated by the models.

Results


Original relationship across 63 sites
Average of Weekly New Cases
1000
100
10
1
SARS−CoV−2 Particles per Liter
1e+02
1e+04
1e+06
1e+08


Relationship Post Population Normalization
Normalized Average of Weekly New Cases
0.100
0.010
0.001
1e+02
1e+04
1e+06
1e+08
SARS−CoV−2 Particles per Liter
LOD
Above LOD
Below LOD



```
##
## Call:
## lm(formula = case_rate ~ avg_sars_cov2_conc, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4786 -1.0003 -0.1675  0.8605  4.7966
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.079002   0.065870   1.199     0.23
## avg_sars_cov2_conc 0.184926   0.005678  32.571   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.233 on 9012 degrees of freedom
## Multiple R-squared:  0.1053, Adjusted R-squared:  0.1052
## F-statistic:  1061 on 1 and 9012 DF,  p-value: < 2.2e-16
```

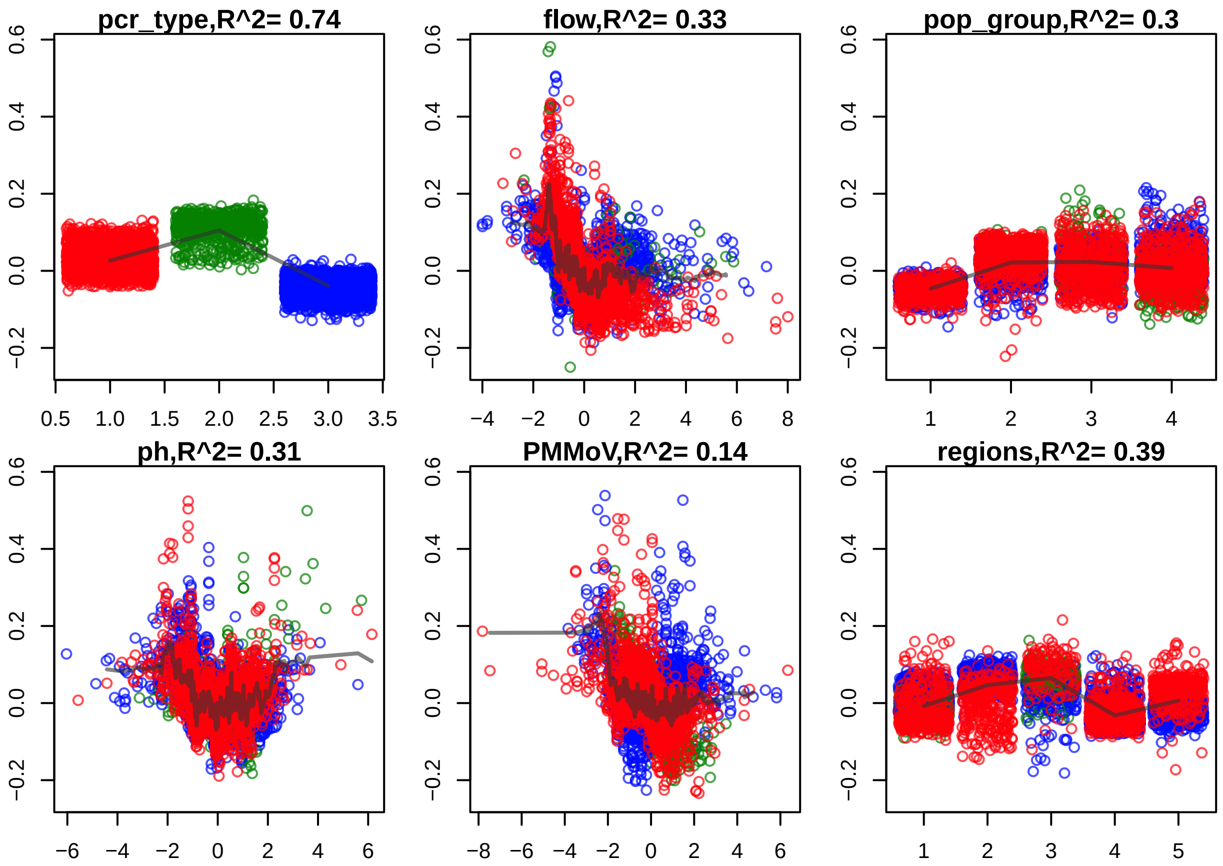
```
##
## Call:
## lm(formula = case_rate ~ avg_sars_cov2_conc + log(pop), data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

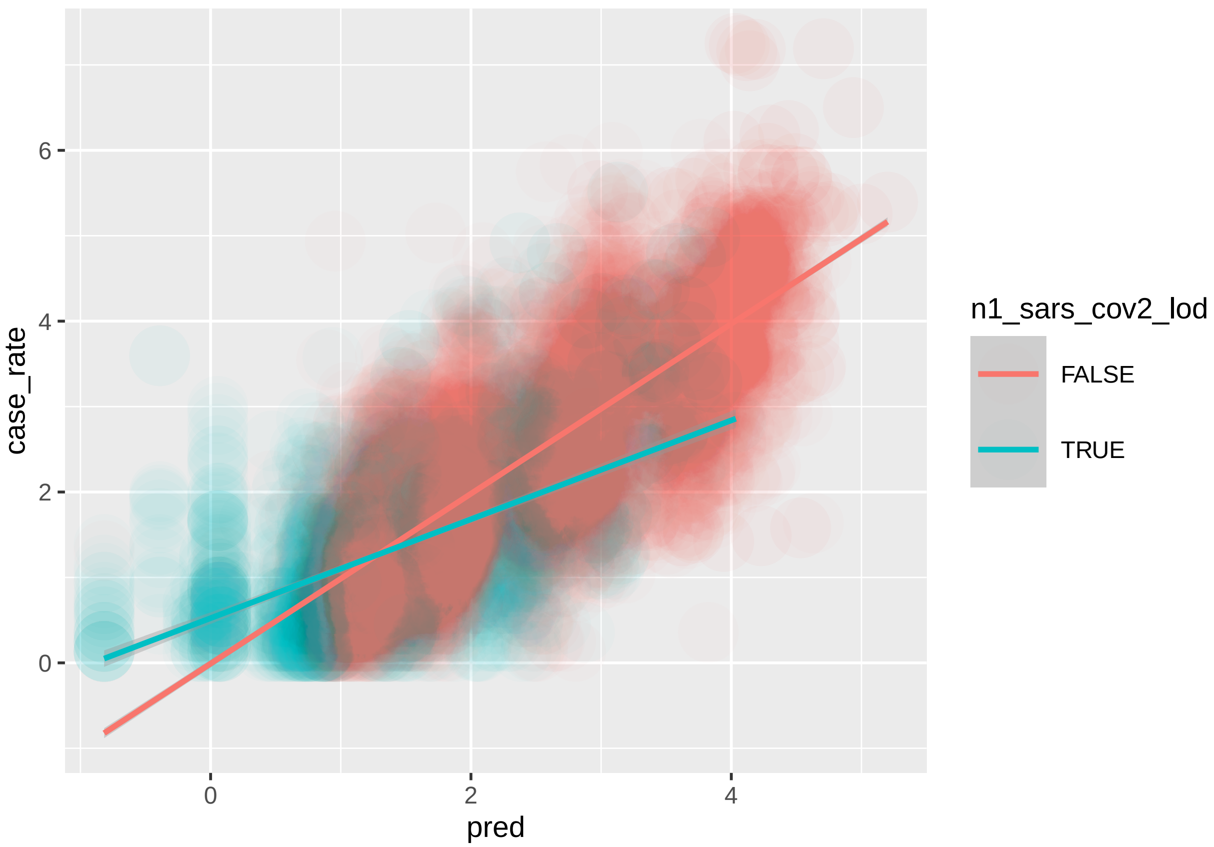
	%IncMSE	IncNodePurity
regions	63.10622	73.42110
PMMoV	68.07717	274.73034
ph	72.08750	230.92665
pop_group	75.92171	70.63900
flow	80.73994	272.57386
pcr_type	80.96885	51.43222

```
## -2.9737 -0.4623 -0.0368  0.3910  3.2504
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -9.434328   0.179832  -52.46   <2e-16 ***
## avg_sars_cov2_conc  0.393083   0.013802   28.48   <2e-16 ***
## log(pop)            0.670324   0.006122  109.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6965 on 5283 degrees of freedom
## Multiple R-squared:  0.7137, Adjusted R-squared:  0.7136
## F-statistic:  6586 on 2 and 5283 DF,  p-value: < 2.2e-16
```

```
##           %IncMSE IncNodePurity
## pcr_type  80.96885      51.43222
## pop_group 75.92171      70.63900
## regions   63.10622      73.42110
## ph        72.08751     230.92665
## flow      80.73994     272.57386
## PMMoV     68.07717     274.73034
```

```
## [1] "compute goodness-of-fit with leave-one-out k-nearest neighbor(guassian weighting), kknn package
```


pcr_type,R^2= 0.74
flow,R^2= 0.33
pop_group,R^2= 0.3
ph,R^2= 0.31
PMMoV,R^2= 0.14
regions,R^2= 0.39
0.6
0.4
0.2
0.0
−0.2
0.5
1.0
1.5
2.0
2.5
3.0
3.5
−4
−2
0
2
4
6
8
1
2
3
4
−6
−4
−2
0
2
4
6
−8
−6
−4
−2
0
2
4
6
1
2
3
4
5



```
##
## Call:
## lm(formula = case_rate ~ avg_sars_cov2_conc:pop_group + pop_group,
##     data = df_LOD)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4671 -0.5307 -0.0177  0.4799  3.2242
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1.01763    0.24530  -4.148 3.39e-05 ***
## pop_group2                    0.92297    0.31456   2.934  0.00336 **
## pop_group3                    0.79942    0.32483   2.461  0.01388 *
## pop_group4                    0.21299    0.30760   0.692  0.48869
## avg_sars_cov2_conc:pop_group1  0.18161    0.01949   9.316  < 2e-16 ***
## avg_sars_cov2_conc:pop_group2  0.15769    0.01605   9.826  < 2e-16 ***
## avg_sars_cov2_conc:pop_group3  0.24816    0.01740  14.265  < 2e-16 ***
## avg_sars_cov2_conc:pop_group4  0.37599    0.01504  24.993  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7728 on 6321 degrees of freedom
## Multiple R-squared:  0.6442, Adjusted R-squared:  0.6438
## F-statistic:  1635 on 7 and 6321 DF,  p-value: < 2.2e-16
```

	adjusted R^2	mse	num_param
All categorical interaction model	0.771	0.383	145
All categorical indirect model	0.756	0.413	49
Sub data all interaction model	0.767	0.386	73
Sub data indirect interaction model	0.755	0.410	24
Base relationship	0.139	1.464	3