Comparison of models for under reporting in Covid-19 and lung cancer cases

In the US by State

April 11, 2024

Nathen Byford

Baylor University

Project Overview

- Compare how models perform on potentially under reported data.
  - Under reported data is when the observed value is not the true value due to some form of measurement error.
- Under reported data can bias analyses and affect the following decisions.
- Project will compare 3 models on two responses, Covid-19 and lung cancer[1] counts

Covid Count Data

Covid-19 count data is likely under reported due to many reasons:

- It can manifest in different severity levels,
- Diagnostic challenges,
- Stigma or social implications,
- Some people are apprehensive to get tested,
- etc.

Under reported data can bias the estimates to be lower than they really are, can be thought of as unintentional missing data.

lung cancer count data

lung cancer as a more serious disease may not be as under reported and might not benefit from using an under reported model.

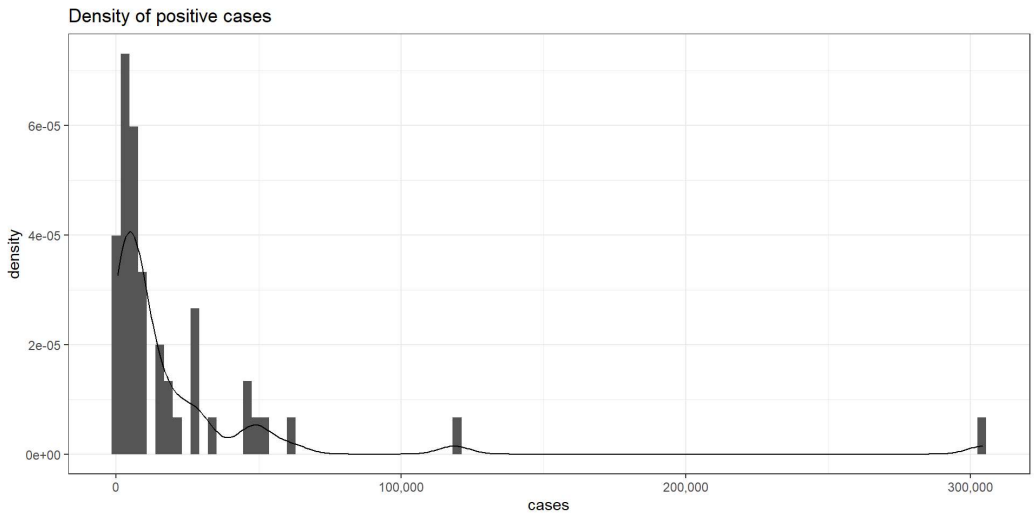
It is reasonable to assume there is low under reporting of lung cancer in the US;

- Serious illness,
- Robust disease surveillance infrastructure, and
- Test maturity.

Covid Data

- Covid-19 counts for Lower 48 & DC from April 2020
- 23 variables
- Response: Positive cases (count)
- Spatial Component: State (lattice)

Distribution of response



Summary statistics for variables

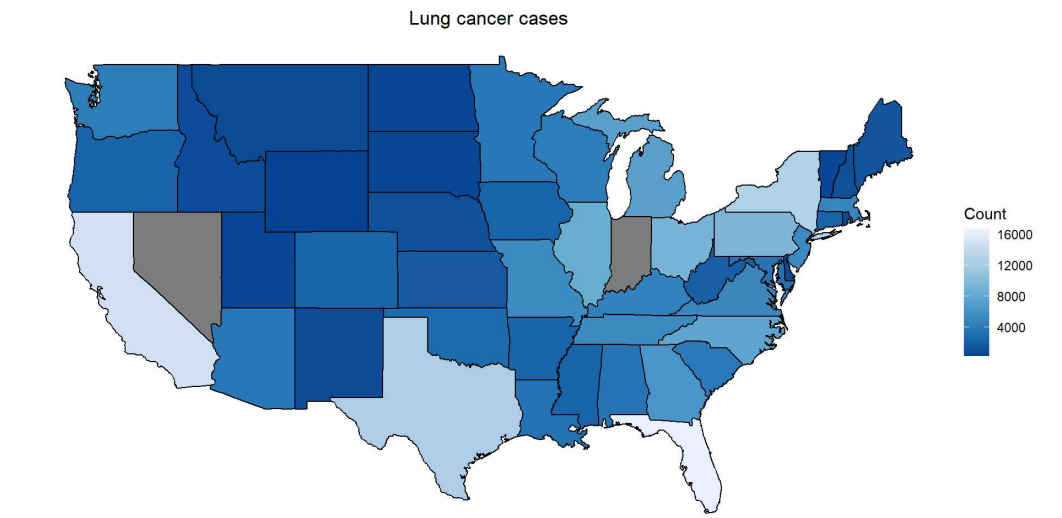
Characteristic	N = 49[1]
Positive tests	7,562 (3,618, 21,742)
Total tests	81,465 (42,667, 161,181)
Testing Rate	0.018 (0.015, 0.027)
Population Density	106 (52, 231)
Air Pollution	7.40 (6.80, 8.20)
Obesity	30.9 (28.7, 34.4)
Smoking	16.10 (14.50, 19.00)
Excessive Drinking	18.20 (16.40, 19.40)

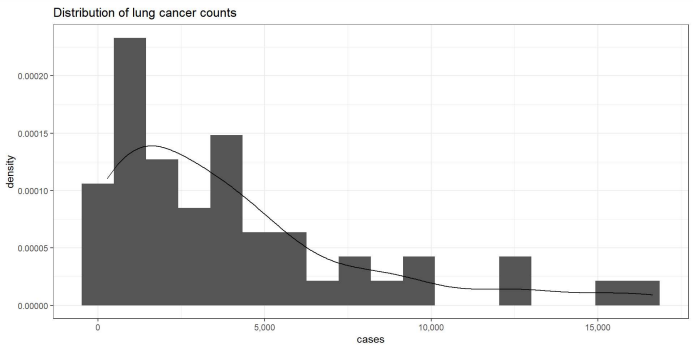
[1] Median (IQR)

Not exhaustive list of variables and summary statistics

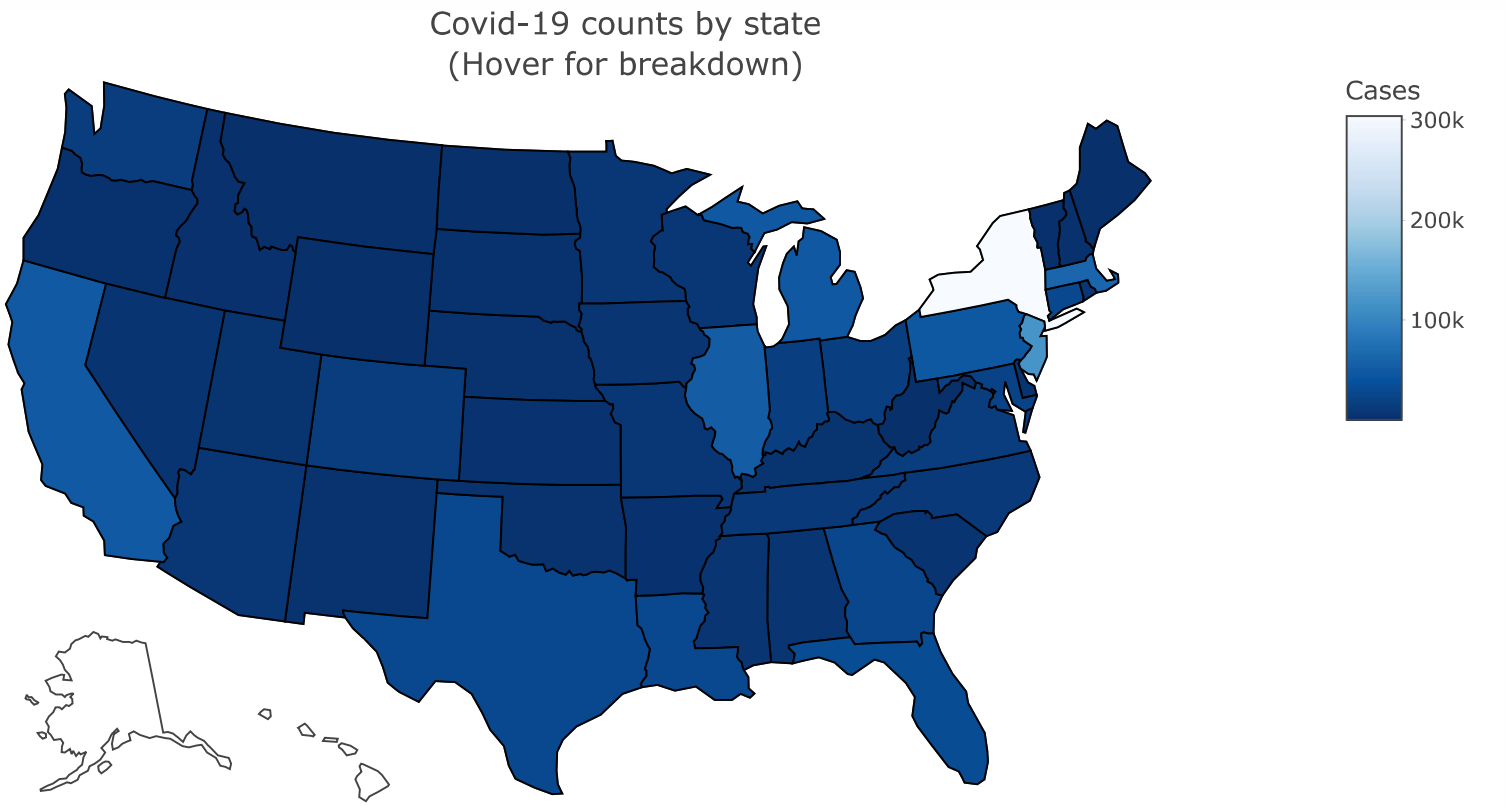
Lung cancer data

- Same covariates, new response variable.
- Nevada and Indiana did not meet USCS[1] publication criteria

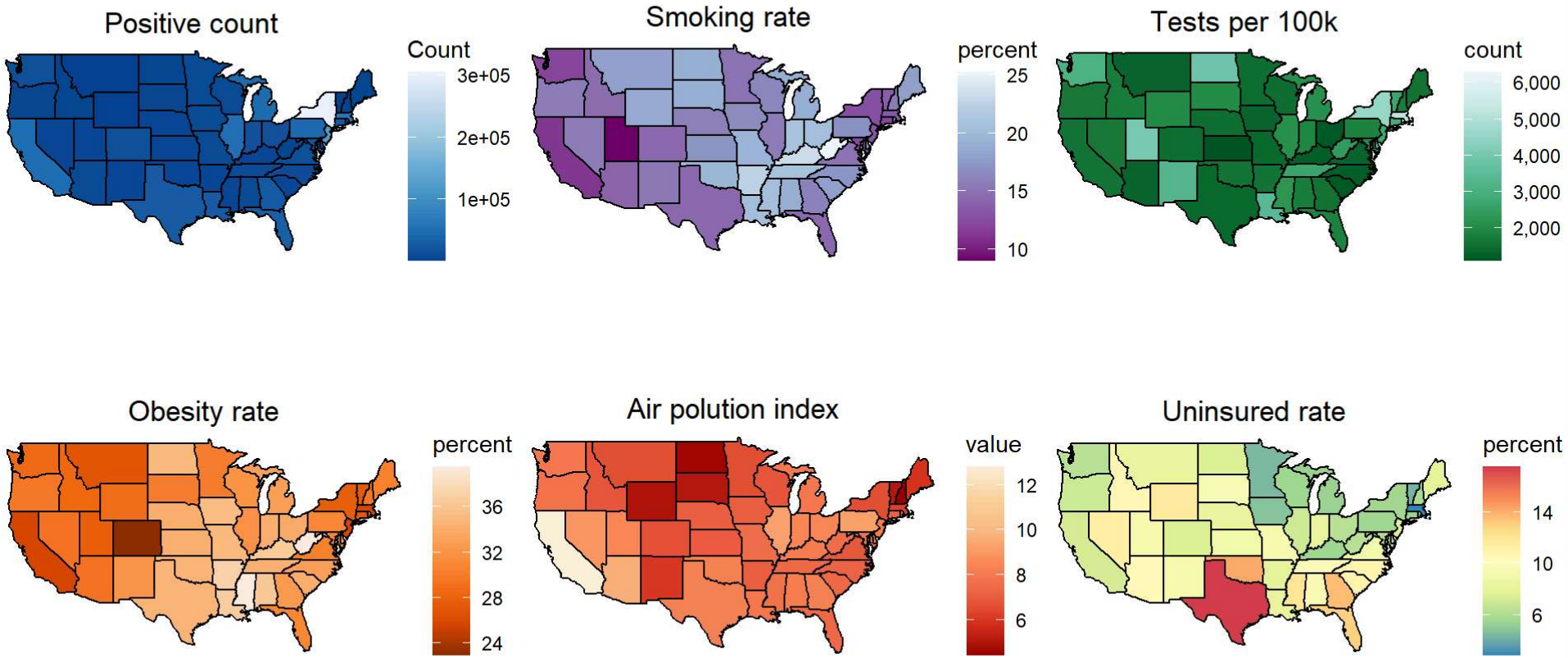




EDA


Covid-19 counts by state
(Hover for breakdown)
Cases
300k
200k
100k

Variables


Positive count
Count
3e+05
2e+05
1e+05
Smoking rate
percent
25
20
15
10
Tests per 100k
count
6,000
5,000
4,000
3,000
2,000
Obesity rate
percent
36
32
28
24
Air polution index
value
12
10
8
6
Uninsured rate
percent
14
10
6

Methods

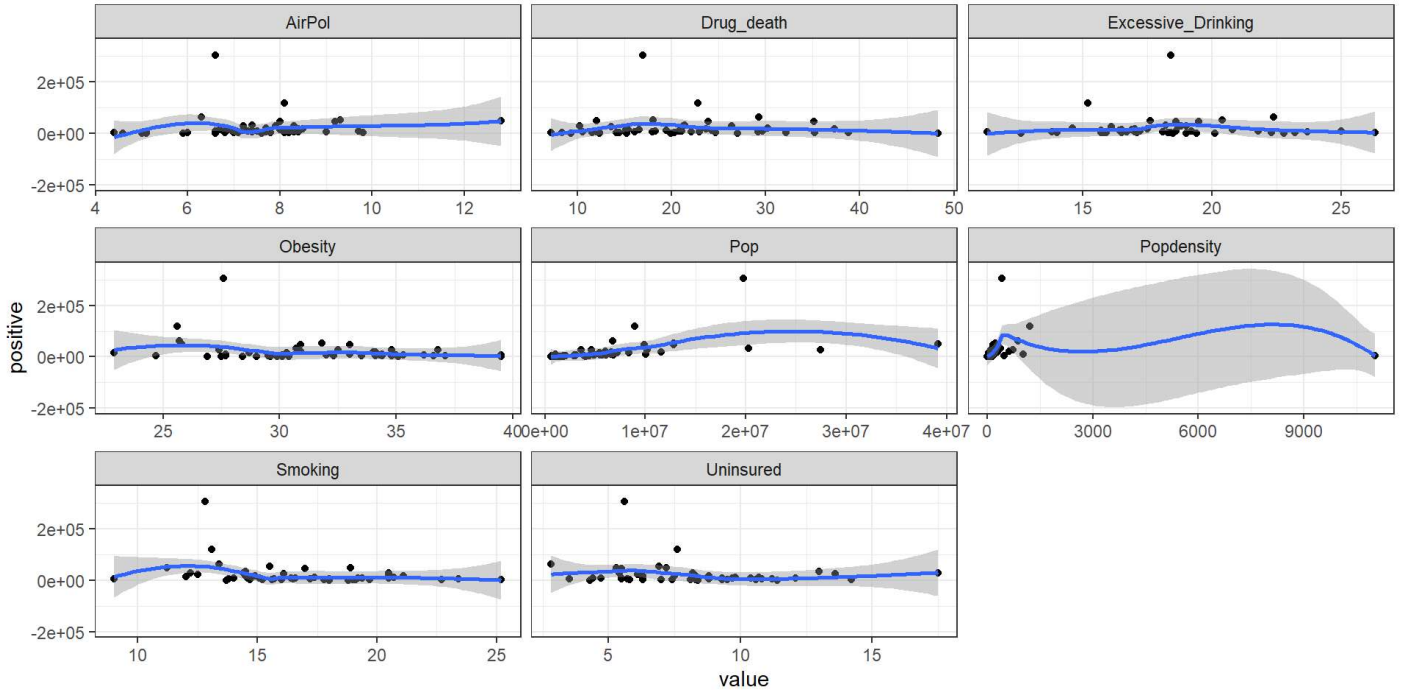
- Non-spatial naive model
- Spatial model
- Under reported spatial hierarchical model

Model comparison methods

- Deviance information criterion (DIC), and
- Watanabe-Akaike information criterion (WAIC).

Non-spatial model (Naive)

- Regression model ignoring spatial component
- multivariate regression model
- Model selection



Spatial model

- Poisson regression
  - Using a log link on the Poisson mean[1]
  - Can compare frequentest and Bayesian model
- Poisson Kriging

Under reporting hiearchical model

- Idea comes from paper for correcting under reported lung cancer counts in Brazil (Stoner, Economou, and Drummond Marques da Silva 2019)
- Extension of Poisson-logistic regression model
  - Adds an under reporting component
- Have code for model from paper as a starting point

Hiearchical model

let z_s be the observed (under reported) counts, y_t be the true unknown counts, π_s be the under reporting rate, and λ_s be the Poisson mean.

The hierarchical model can be written as,

$$z_s|y_s \sim \text{Binomial}(\pi_s, y_s)$$
$$\downarrow$$
$$y_s \sim \text{Poisson}(\lambda_s)$$

where π_s uses a logit link function and λ_s uses a log link function to determine values for the parameters.

Thank You!

References

Stoner, Oliver, Theo Economou, and Gabriela Drummond Marques da Silva. 2019. “A Hierarchical Framework for Correcting Under-Reporting in Count Data.” *Journal of the American Statistical Association* 114 (528): 1481–92. https://doi.org/10.1080/01621459.2019.1573732.