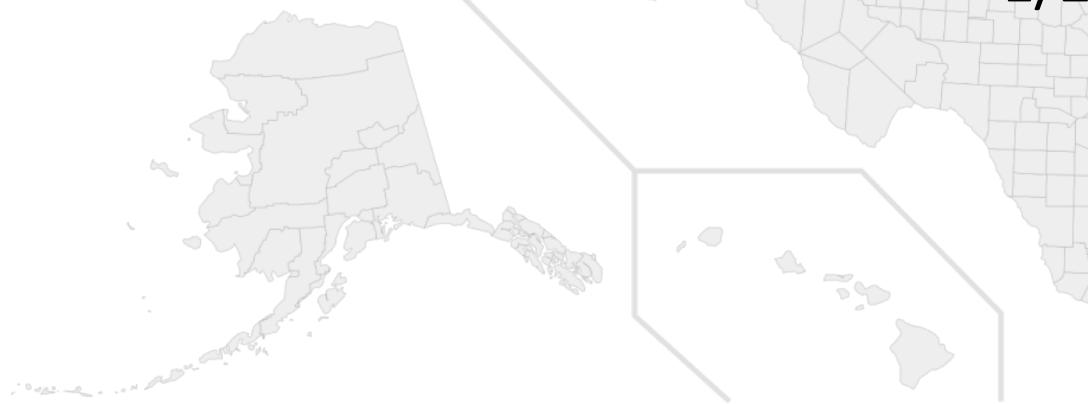


Predicting Covid-19 Deaths at the US County Level Using Census Data

Nick Kinnaird

1/22/2021



Idea: Knowing which counties are more at risk means resources can be deployed more efficiently.

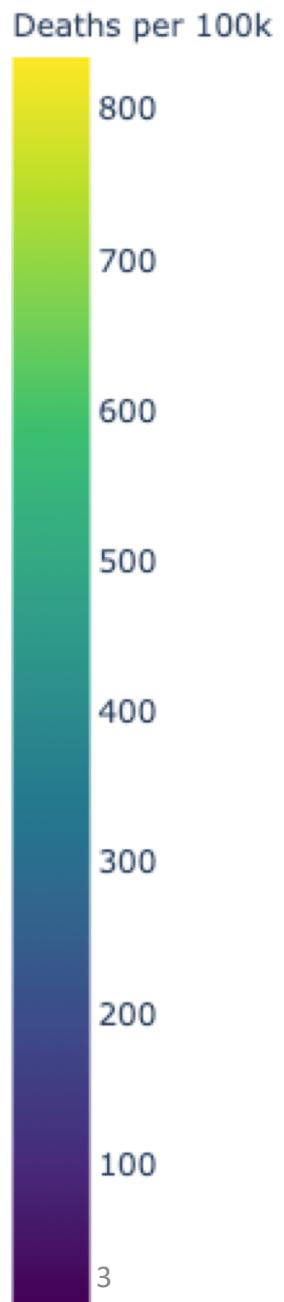
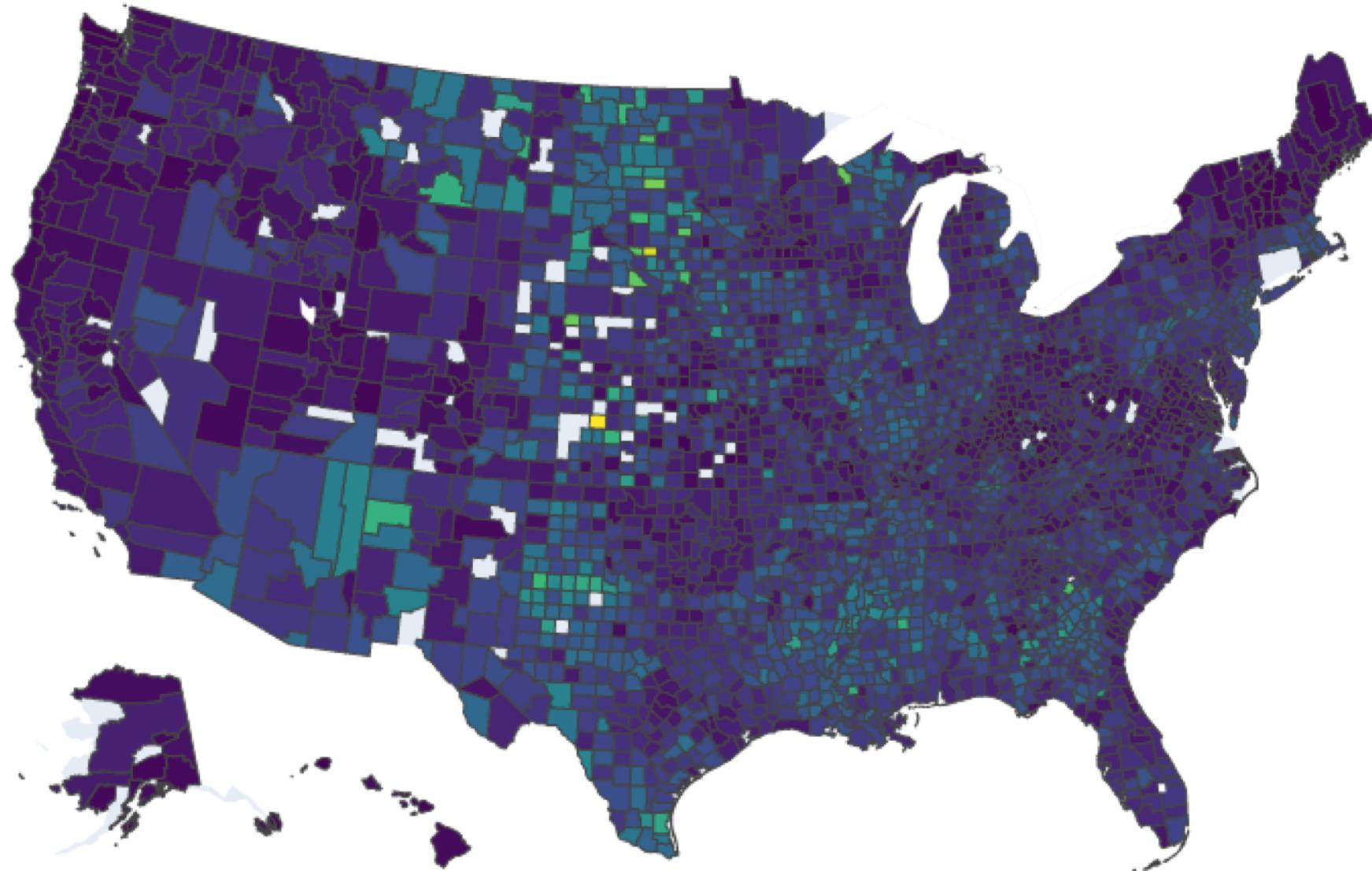
Target variable: **Cumulative** Covid-19 deaths per county **scaled by population**

- Data acquired from the NYT Covid-19 Github repository
- Dated 1/18/2021
- ~3000 counties

Features: US Census data scraped from the Census Bureau

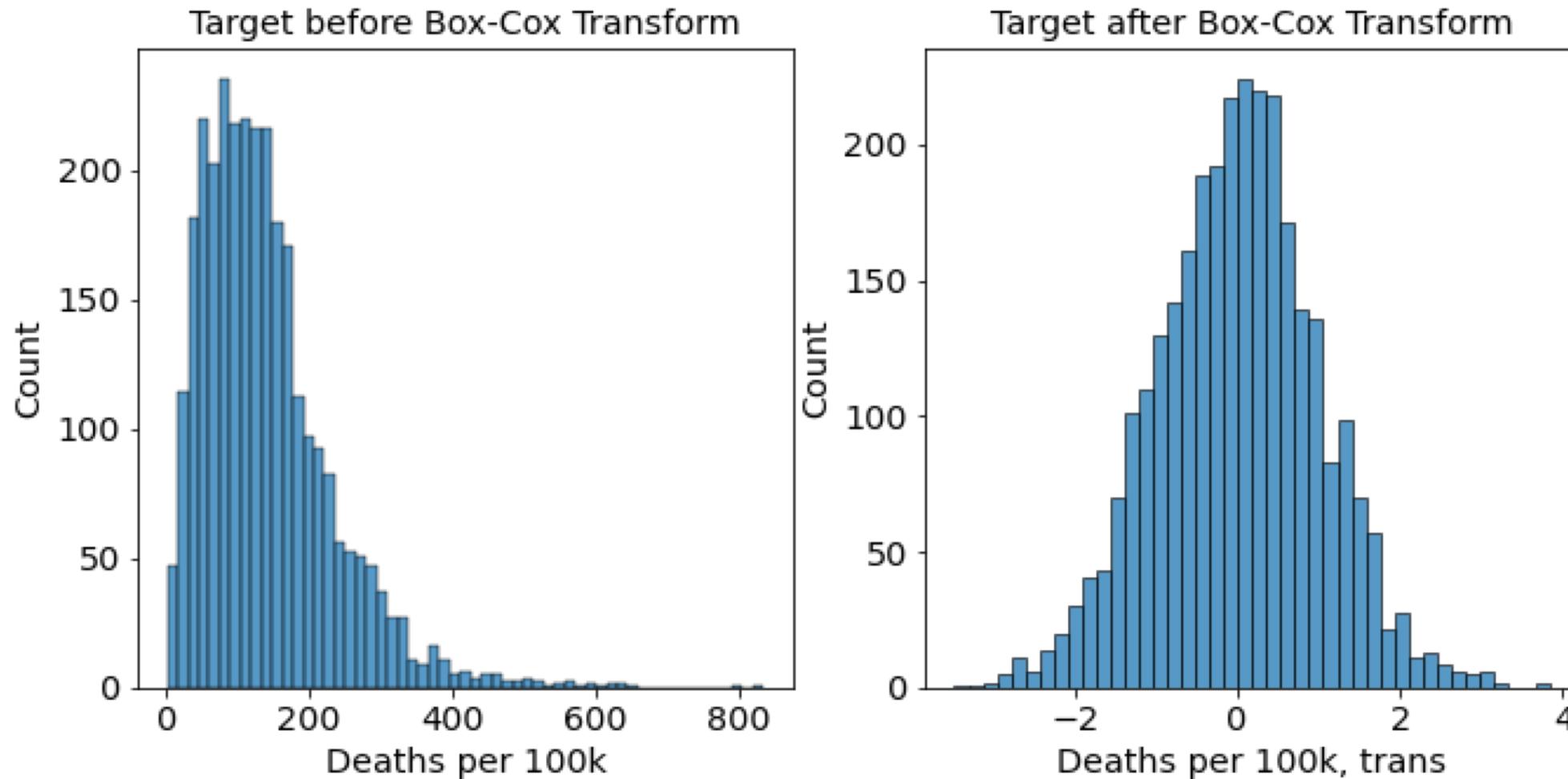
- 62 features consisting of various demographic and economic data

US County Cumulative Deaths per 100k People



Made using Plotly Choropleth Map. Blank counties are places where either my web-scraping failed, or I dropped outliers/entries with issues.

Target Distribution with and without Box-Cox Transform



- Transforming the target distribution to a normal shape was necessary for my regression to work consistently.

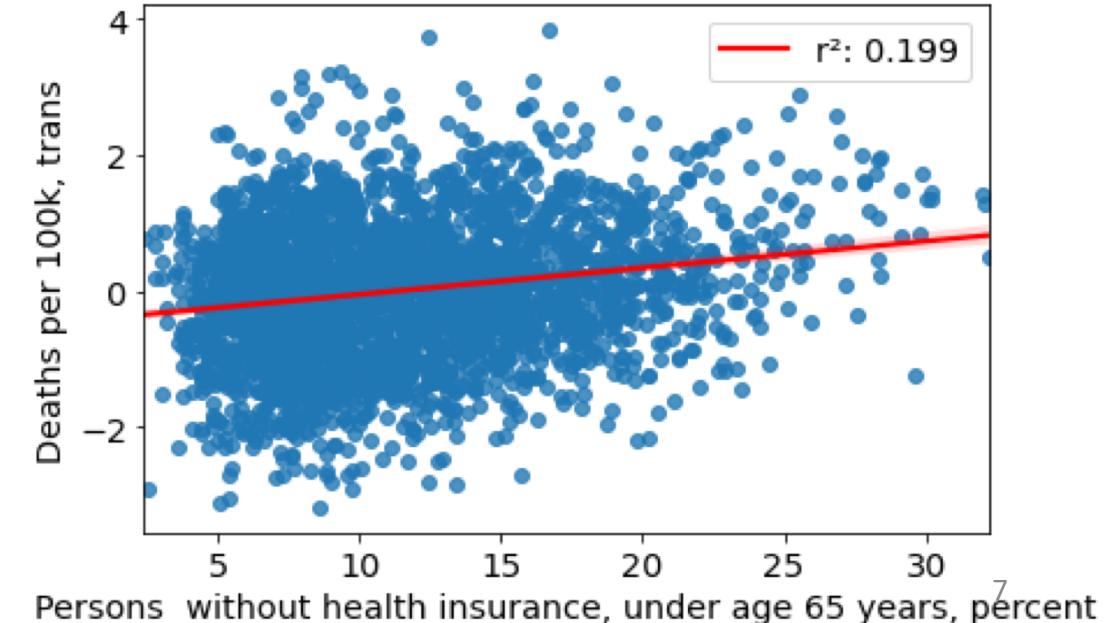
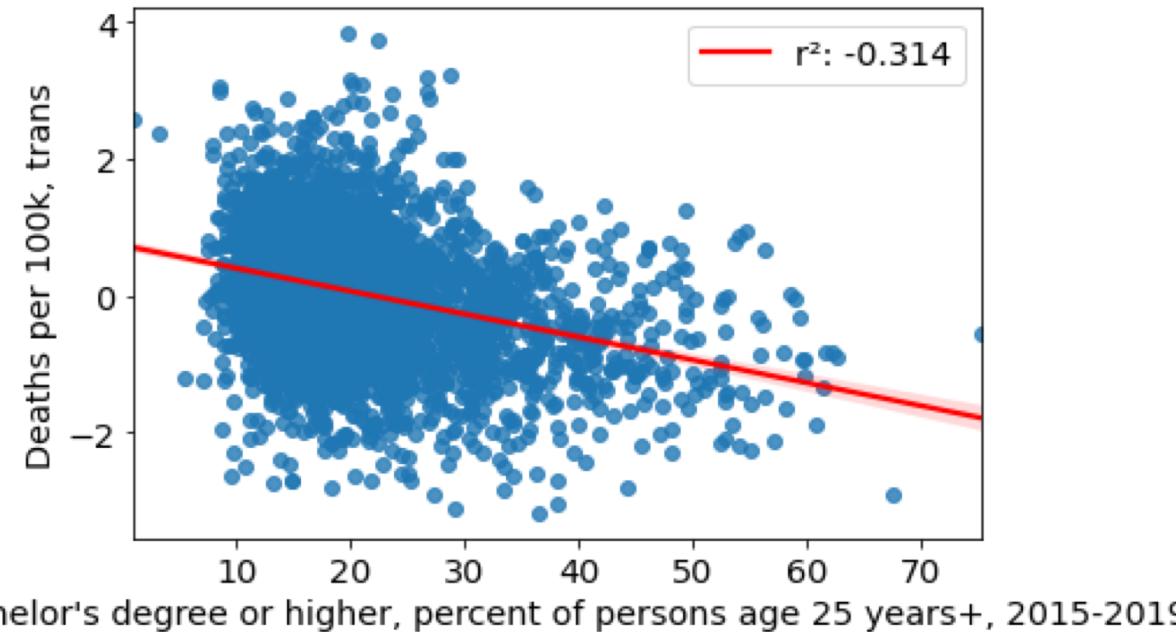
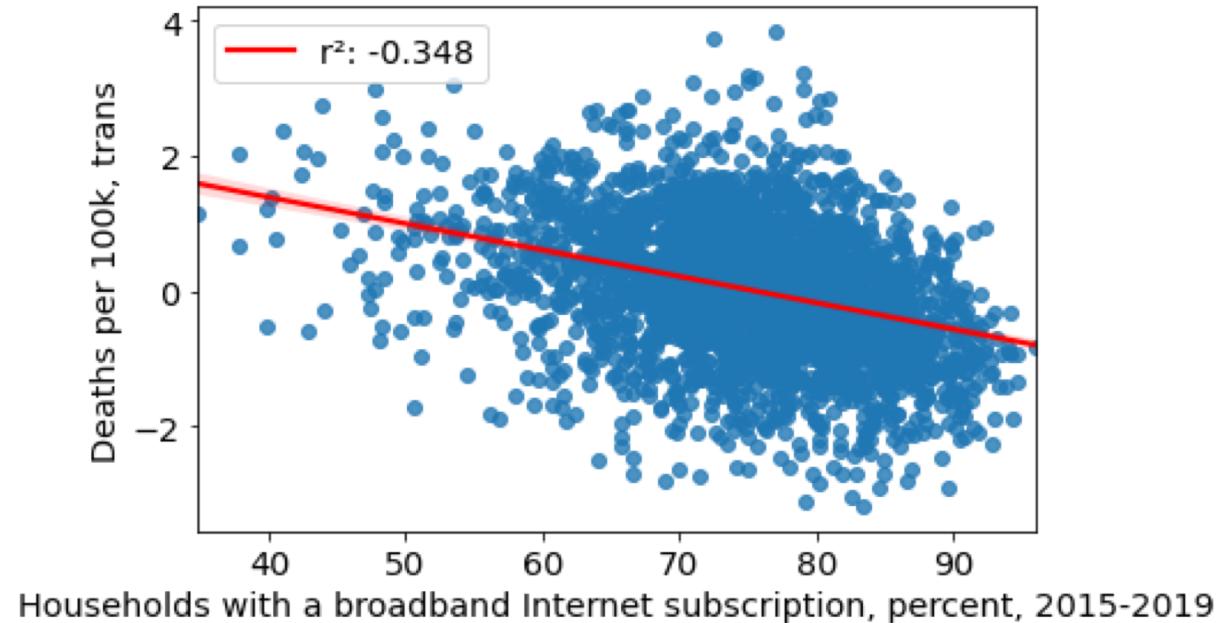
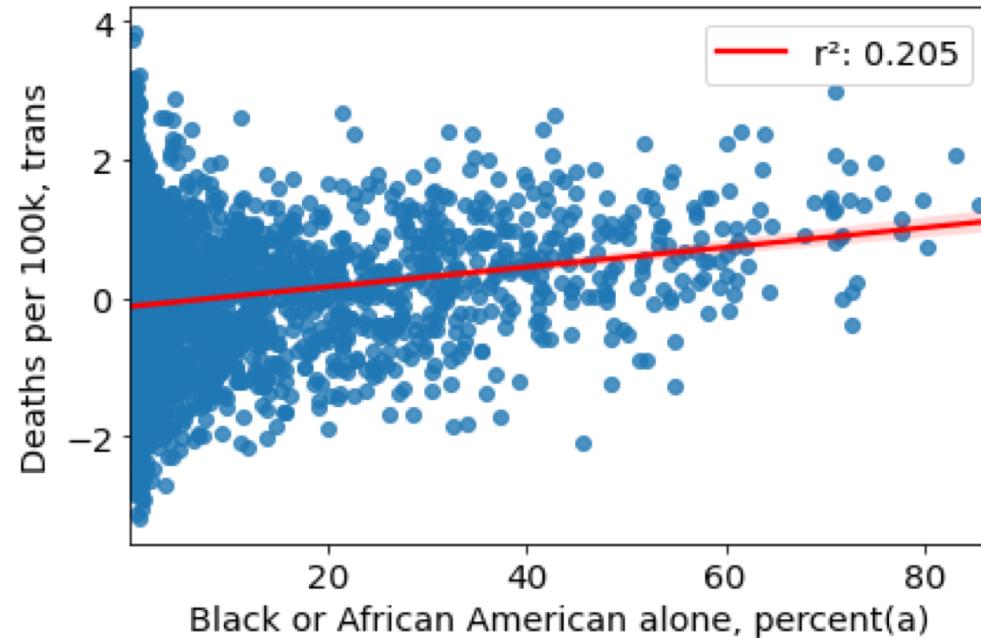
What influences the number of deaths?

- Age
- Race
- Income
- Education
- Health care
- Etc.
- A lot of these are significantly correlated to each other
- By trial and error I selected 12 features for my model

Selected Features

Race/Age
Income Related
Education/Information

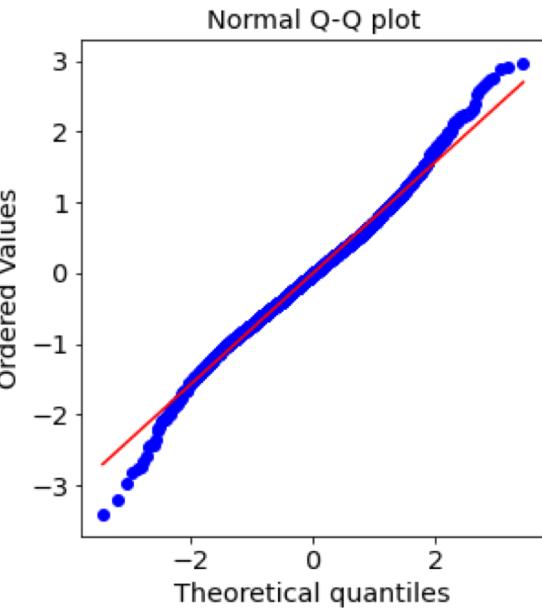
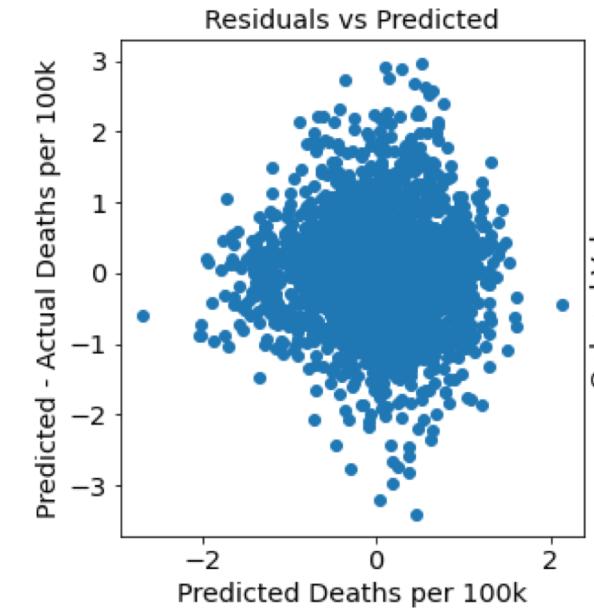
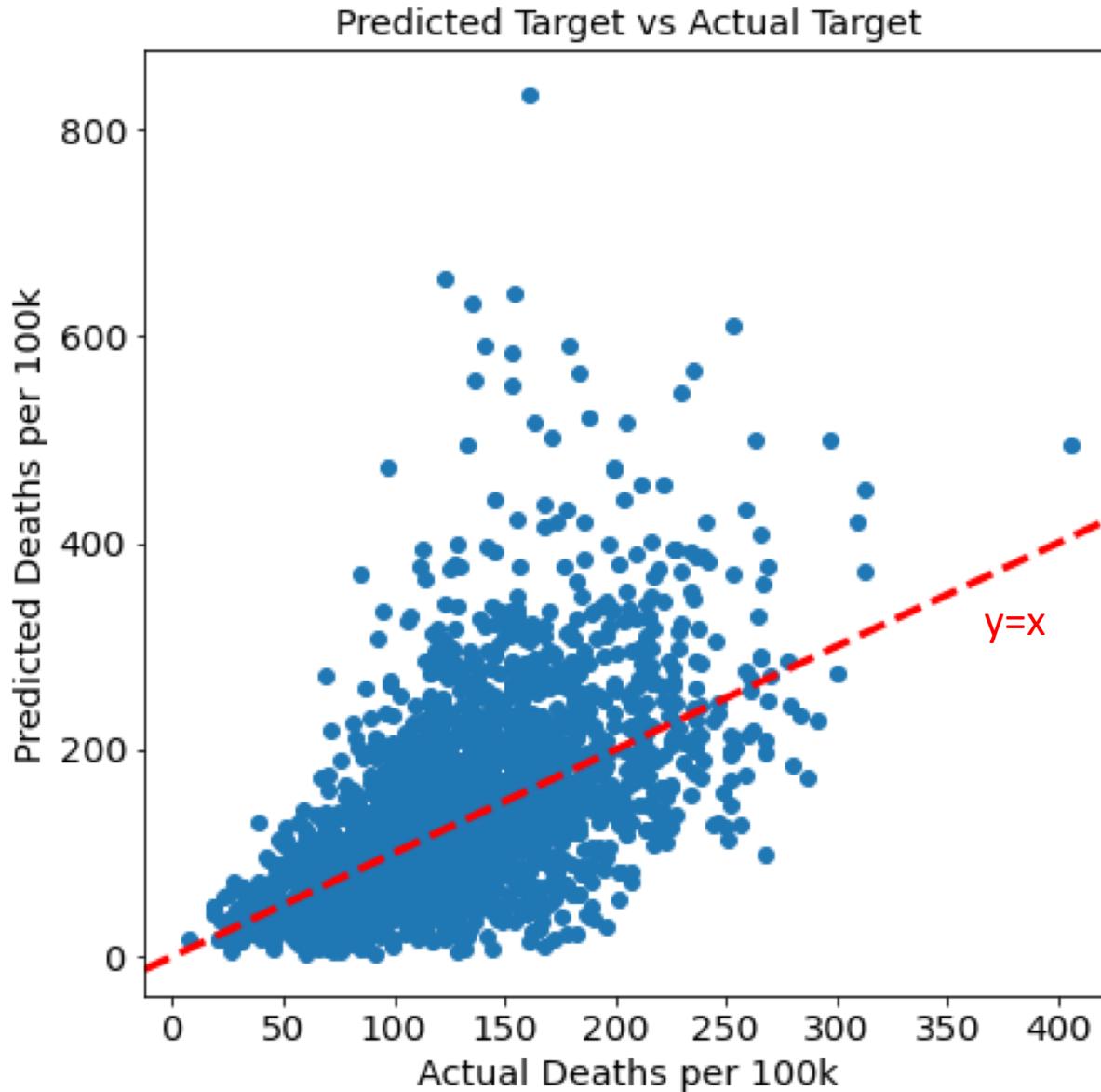
| Number | Features | Correlation to target |
|--------|---|-----------------------|
| 1 | Persons 65 years and over, percent | 0.047 |
| 2 | Black or African American alone, percent | 0.205 |
| 3 | Two or More Races, percent | -0.278 |
| 4 | White alone, not Hispanic or Latino, percent | -0.201 |
| 5 | Persons without health insurance, under age 65 years, percent | 0.199 |
| 6 | Median value of owner-occupied housing units, 2015-2019 | -0.369 |
| 7 | Median household income (in 2019 dollars), 2015-2019 | -0.289 |
| 8 | Living in same house 1 year ago, percent of persons age 1 year+, 2015-2019 | 0.186 |
| 9 | Households with a computer, percent, 2015-2019 | -0.353 |
| 10 | Households with a broadband Internet subscription, percent, 2015-2019 | -0.348 |
| 11 | High school graduate or higher, percent of persons age 25 years+, 2015-2019 | -0.276 |
| 12 | Bachelor's degree or higher, percent of persons age 25 years+, 2015-2019 | -0.314 |



Model Choice: Lasso with standard polynomial features

- Polynomial degree 2, including all poly features, not just interaction terms
- $\lambda = 4e-4$
- Ultimately I selected the model which performed the best, regardless of complexity
 - Verified with cross validation
- I decided this because this was an analysis on life/death data, and every little bit matters
 - Did choose Lasso over Ridge for better interpretability

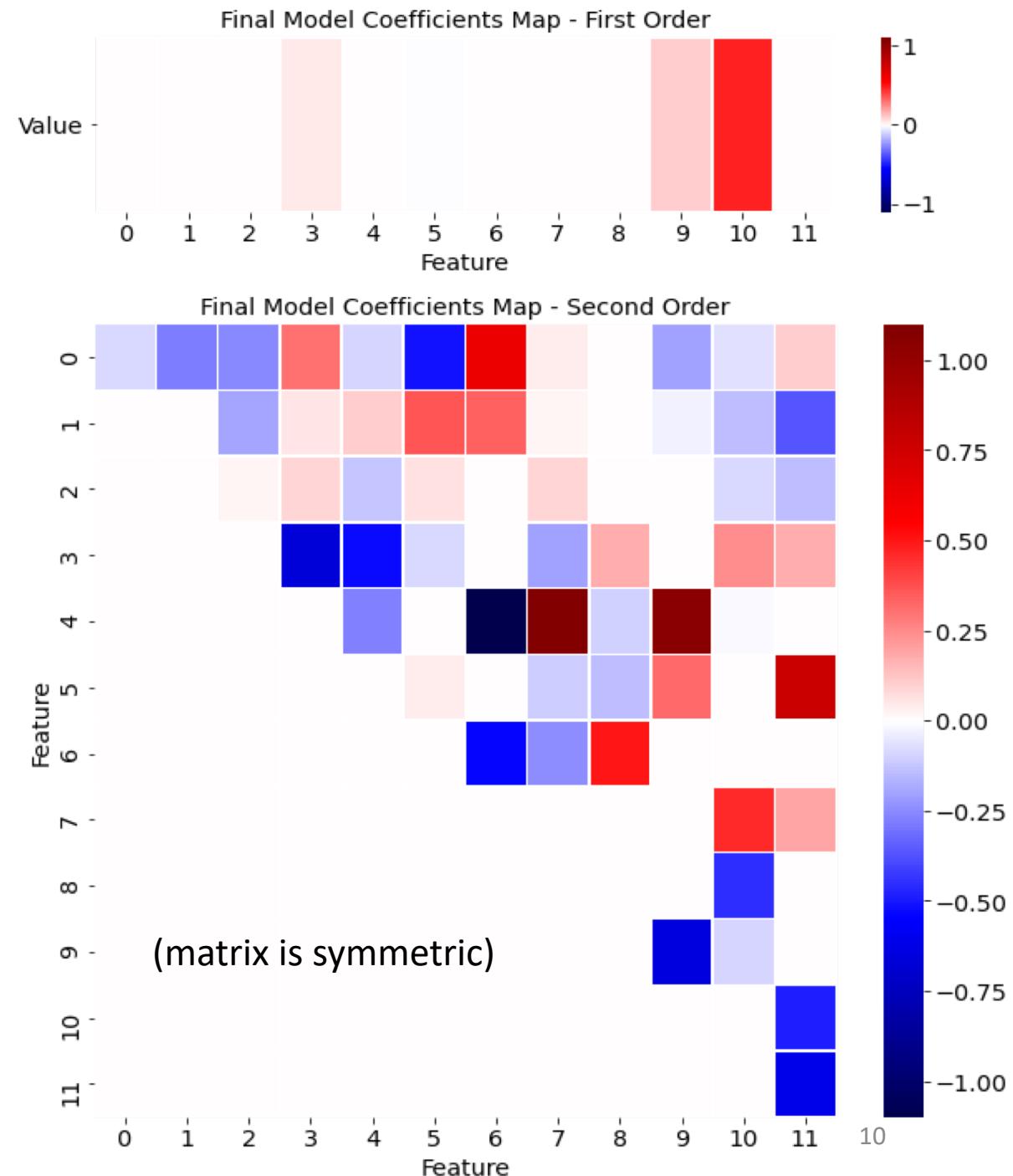
Model Performance



- The predicted vs actual plot shows that my model tends to overestimate the number of deaths
- Residuals are relatively happy
- Variance is high

Polynomial Coefficients

| # | Feature |
|----|---|
| 0 | Persons 65 years and over, percent |
| 1 | Black or African American alone, percent |
| 2 | Two or More Races, percent |
| 3 | White alone, not Hispanic or Latino, percent |
| 4 | Median value of owner-occupied housing units, 2015-2019 |
| 5 | Living in same house 1 year ago, percent of persons age 1 year+, 2015-2019 |
| 6 | Households with a computer, percent, 2015-2019 |
| 7 | Households with a broadband Internet subscription, percent, 2015-2019 |
| 8 | High school graduate or higher, percent of persons age 25 years+, 2015-2019 |
| 9 | Bachelor's degree or higher, percent of persons age 25 years+, 2015-2019 |
| 10 | Median household income (in 2019 dollars), 2015-2019 |
| 11 | Persons without health insurance, under age 65 years, percent |



Final Model Results

- $r^2 = 36\%$ on training data, 32% on the test data
 - The unexplained variance is high
 - I'm not surprised by this considering the complexity of the problem
- MAE = 182.1 Deaths per 100k
- RMSE = 205.5 Deaths per 100k

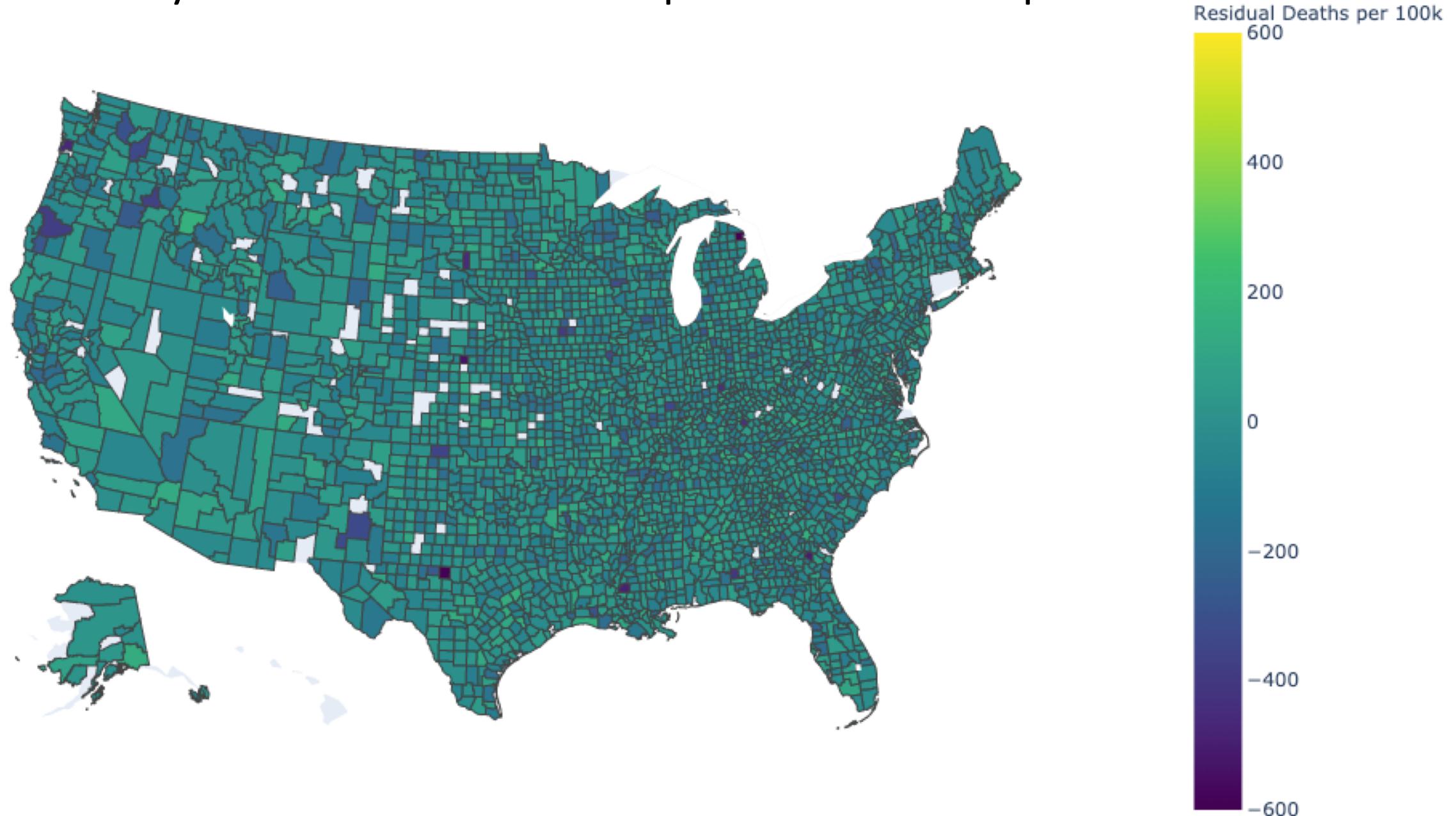
Future Work

- Investigate trend of model to overestimate the number of deaths
- Ultimately the model would need to be improved significantly for it to be of any real use
 - More features including health care related statistics, time-dependent features, etc.
 - More sophisticated calculations on those features
 - Finer granularity: cities or census tracts

End - Questions?

Appendix

US County Residual Deaths per 100k People



Simpler Model – LassoCV with standard features

('Persons 65 years and over, percent', 0.074)
('Black or African American alone, percent(a)', -0.047)
('Two or More Races, percent', -0.228)
('White alone, not Hispanic or Latino, percent', -0.327)
('Median value of owner-occupied housing units, 2015-2019', -0.322)
('Living in same house 1 year ago, percent of persons age 1 year+, 2015-2019', 0.002)
('Households with a computer, percent, 2015-2019', -0.193)
('Households with a broadband Internet subscription, percent, 2015-2019', 0.019)
('High school graduate or higher, percent of persons age 25 years+, 2015-2019', 0.084)
("Bachelor's degree or higher, percent of persons age 25 years+, 2015-2019", -0.057)
('Persons without health insurance, under age 65 years, percent', 0.045)
('Median household income (in 2019 dollars), 2015-2019', 0.162)

$$\lambda = 0.0014$$

$$r^2 = 27\%$$

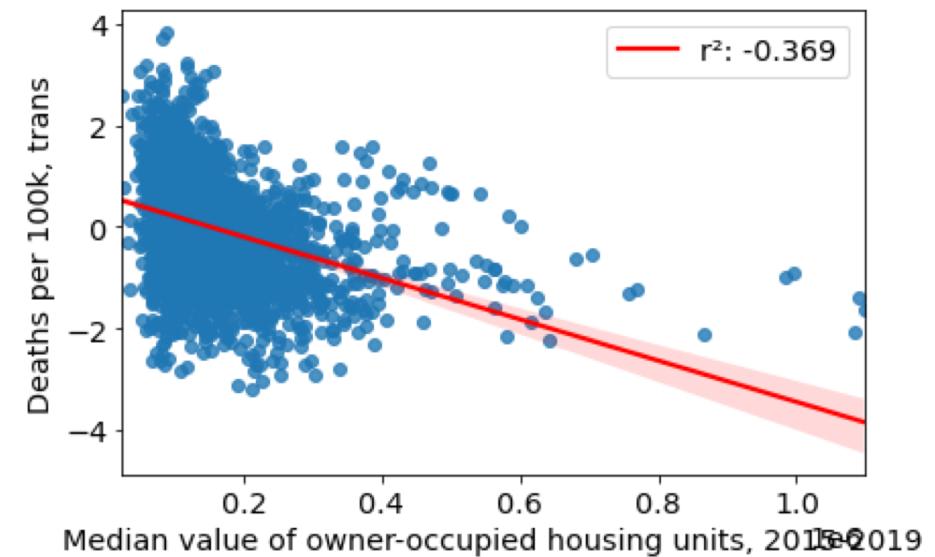
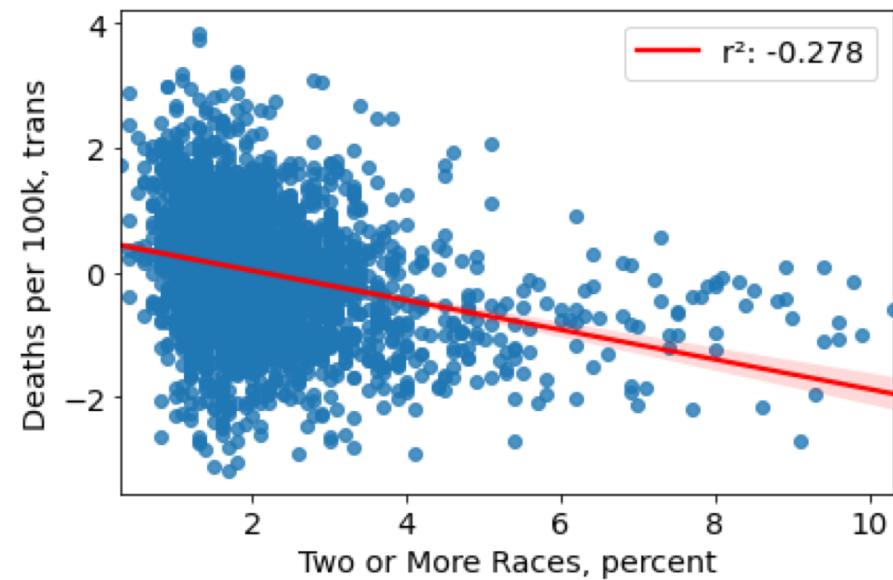
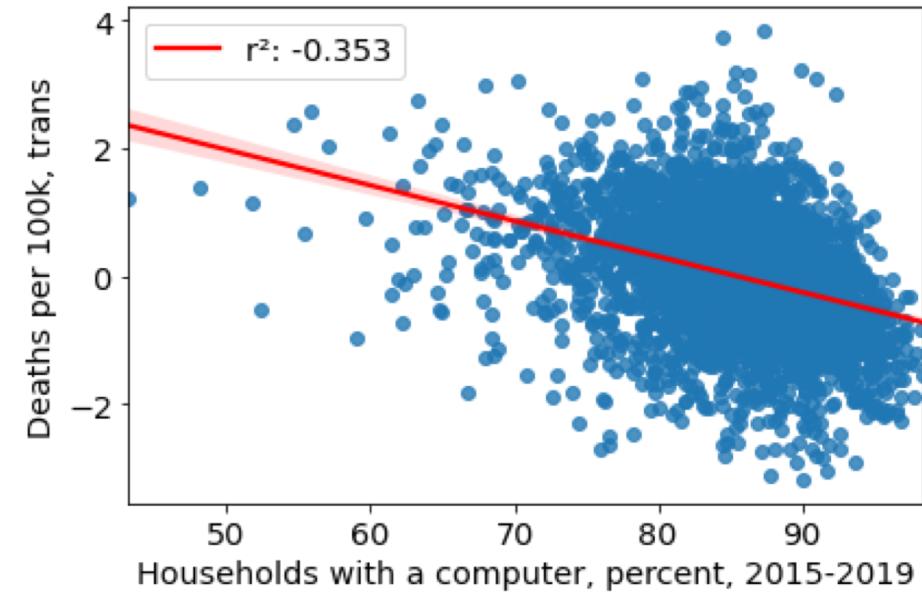
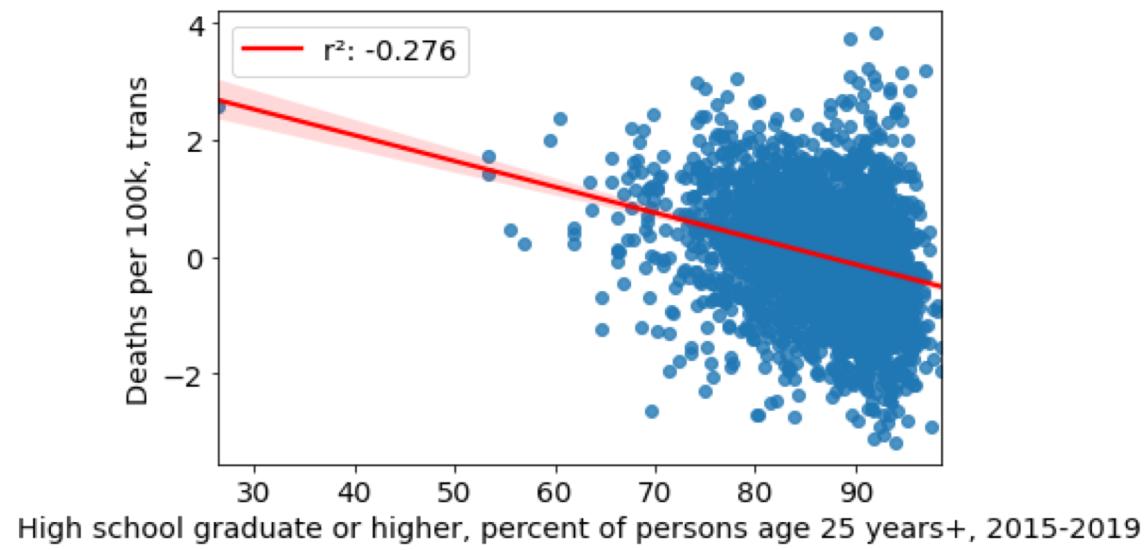
$$MAE = 188.2 \text{ Deaths per 100k}$$

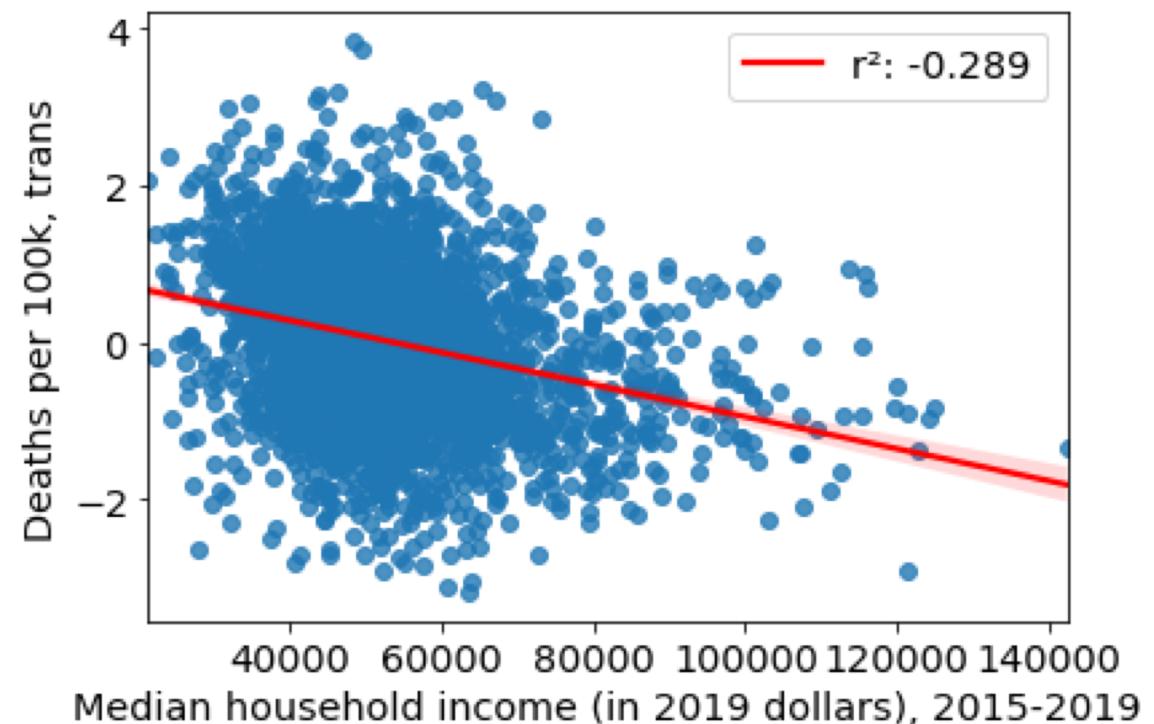
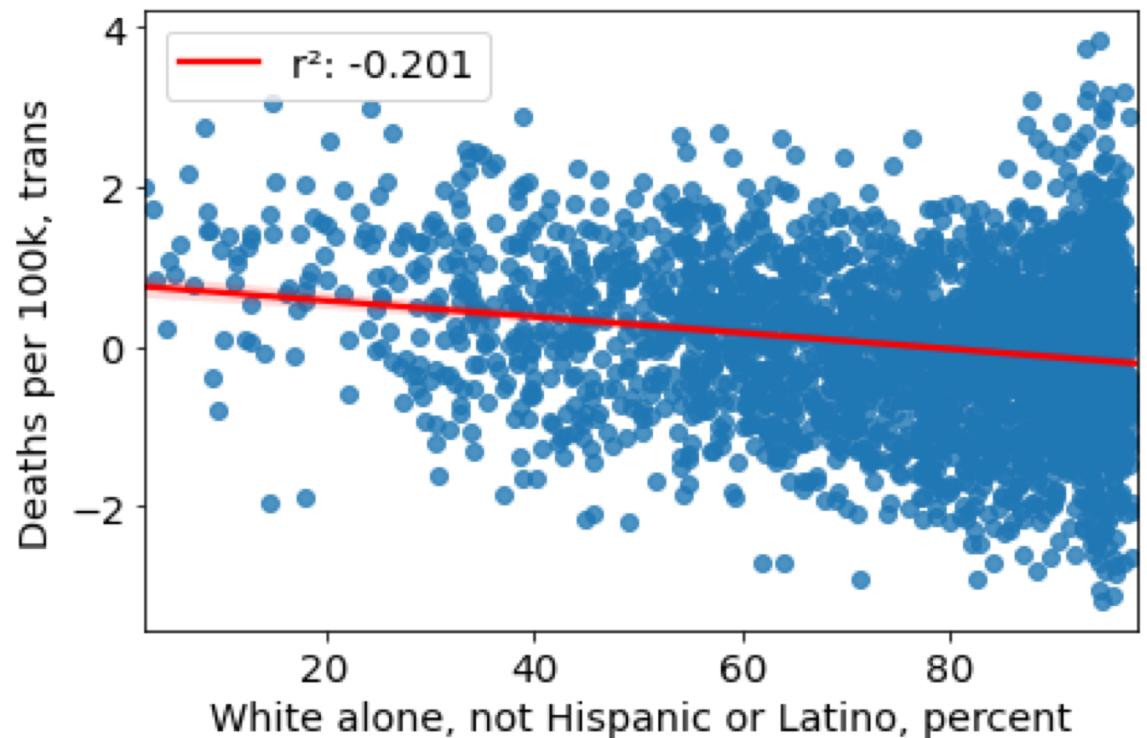
$$RMSE = 211.8 \text{ Deaths per 100k}$$

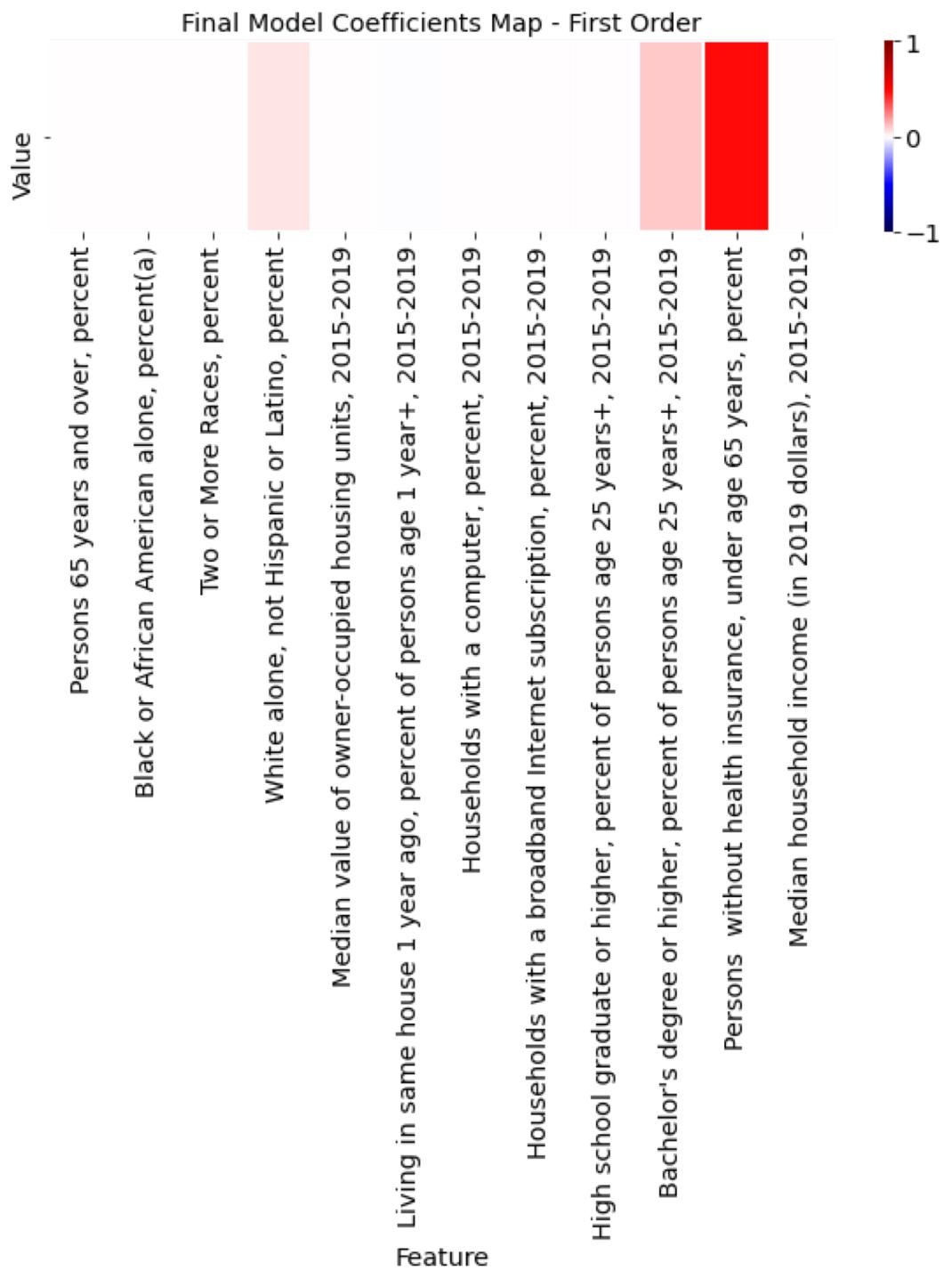
Model progression (on training data)

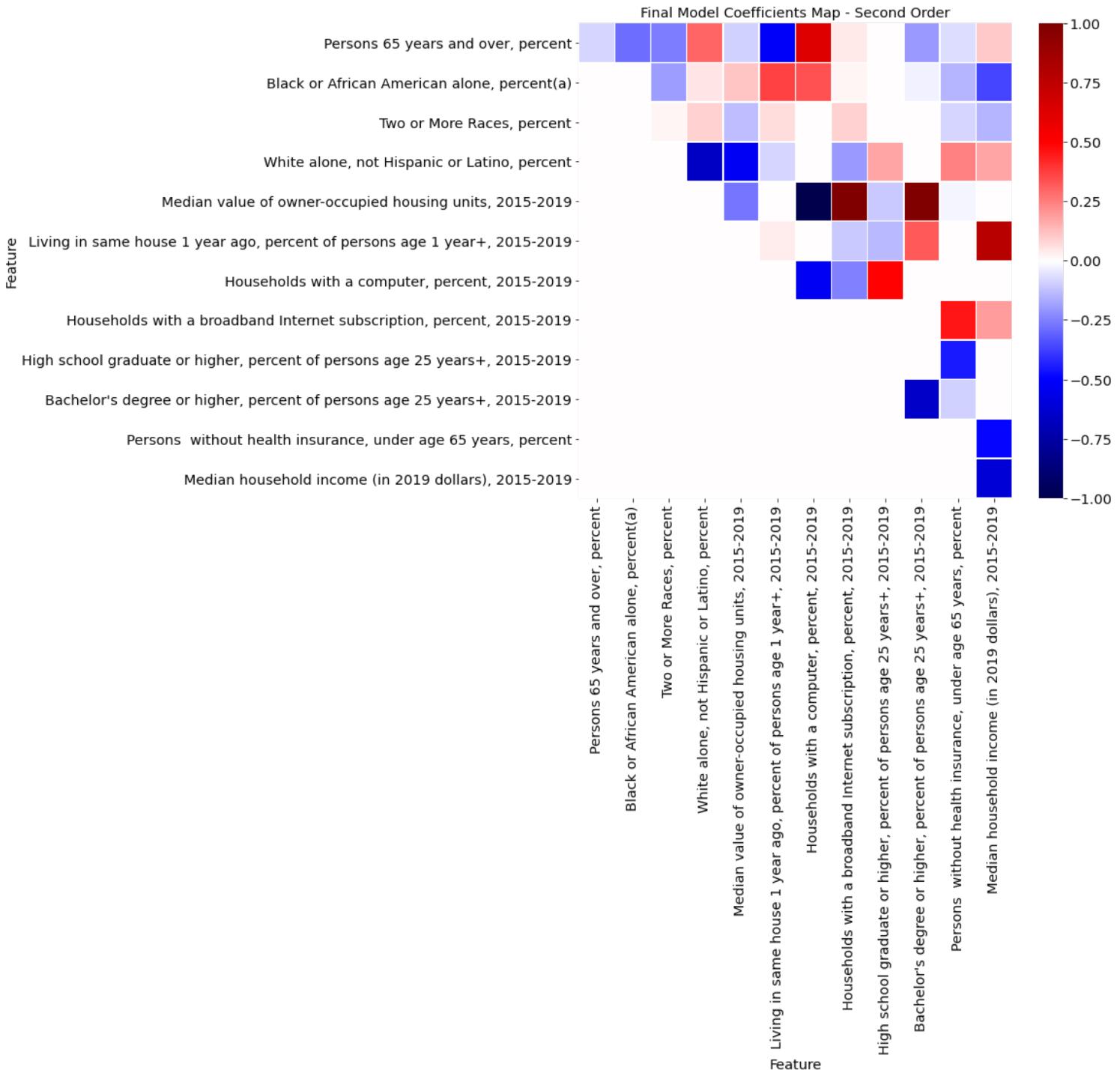
- ~0.1 with simple linear regression using most correlated features
- ~0.2 after reducing features manually
- ~0.27 after Box-Cox transform of target
- ~0.315 after poly transform of features
- ~0.36 using Lasso on poly features

- Ridge and Lasso were basically equivalent
- I chose Lasso because it zeroed out a good amount of the polynomial features









Box-Cox Transform:

$$x_i^{(\lambda)} = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(x_i) & \text{if } \lambda = 0, \end{cases}$$