Spring 2023 Data C102 Final Project

Research Topic: Chronic Diseases & Air Qualities

David Shen

Kaixin Lei

Oscar Li

Xiaomeng Xu

Data Overview

Data Sources 1: "The U.S. Chronic Disease Indicators: Asthma" ("the Asthma dataset")

The CDC's asthma dataset contains information on asthma-related topics such as mortality, hospitalization, and vaccination rates, broken down by gender and racial/ethnic groups for each state from 2010 to 2021. We focus on the "overall age-adjusted mortality cases per million" and "age-adjusted asthma prevalence" as they have the most complete data among the 50 states.

We used age-adjusted rates since they won't be influenced by age distribution differences according to the CDC, and we excluded data from different race groups due to missing values. Dropping NaN will skew our results, be biased towards a certain ethnicity, and result in incomplete outcome variables. Instead, we use overall age-adjusted mortality cases as our outcome variable (y) and find predictors (X) from state-level annual census data.

Data Sources 2: "American Community Survey 1-Year Data" ("the Census dataset")

The U.S. Census Bureau generates census datasets annually from 2010 to 2019 and 2021, with the 2020 data excluded due to poor data quality and incompleteness. Using APIs, we downloaded state-level census data for each year from Data Profile and selected relevant variables. All the census data we used was stored in CSV files named by year.

Data Cleaning & Processing (Data Sources 1+2):

1. Add a corresponding year column to the 11 census datasets and append them into one dataframe: "census".
2. For the asthma dataset: filter "Age-adjusted Rate" and pivot "age_adjusted asthma". We notice that there are many missing values in different groups.
3. Merge age_adjusted asthma and census datasets in the corresponding year and state columns. Filter "Asthma mortality rate" / "Asthma prevalence rate".
4. Drop duplicate columns and NaN values.

The granularity of the dataset is a pair of years and states. Each row contains selected census variables and the aged-adjusted mortality for a state in a given year. The output dataset called "data_mortality" is used for the modeling part.

Both datasets have a significant degree of privacy since the data are aggregated by state, removing all personally identifiable information (PII) and preventing the identification of individuals.

Limitations:

1. Data for certain groups of people are excluded. The Asthma dataset does not have enough information on minority groups for some questions, and the census data exclude hard-to-count populations such as homeless people, non-English speakers, undocumented immigrants, etc. These census data were collected through convenience sampling, such as "surveys, phone calls and door-to-door canvassing".
2. While people might know they are participating in the census, they may not know how their data will be used, and the same goes for participants in the asthma dataset.
3. Insufficient data and features. If available, we would like to add multi-dimensional features to our dataset, such as familiar near-fatal asthma history, adherence to asthma medications, oral corticosteroid usage, etc.

Data Sources 3: "Daily Census Tract-Level PM2.5 Concentrations, 2011-2014" ("the PM2.5 dataset")

Each row of the dataset records the daily PM2.5 concentration at a specific location, presumably based on sensor measurements. It's a comprehensive collection of all the PM2.5 data across the U.S., so we consider it a census. There is no privacy concern as the dataset does not directly involve human subjects.

Data Cleaning & Processing:

Due to the large size of the dataset (with 100+ million rows), we pre-processed it in an aggregated, low-granularity fashion: taking the means of the estimated 24-hour PM2.5 concentration by year and state. Our method of aggregating data helps reduce the variance.

Limitations:

1. We note that unlike human subjects, we can never know the true underlying distribution/population of PM2.5. Since urban, developed areas are more likely to have weather/sensor stations monitoring the air quality, there exists a potential source of bias in the locations of the measurements.
2. Aggregating data removes extreme values, such as a sudden spike in PM2.5 in a specific county someday, which could lead to interesting results with further study.

Research Questions

Question 1: Predicting asthma age-adjusted mortality cases per million from annual census data for each state.

GLMs and nonparametric methods are a **good fit** for predicting the number of asthma mortality cases. The general goal of GLMs is to use predictor X to predict predictand y by estimating some weights β. Parametric methods such as OLS make assumptions about the relationship between data points, whereas non-parametric methods such as Bayesian GLMs and random forests make fewer assumptions when making predictions with the goal of finding the best predictor.

Real-world decisions, such as medical resource allocation, healthcare improvement, and asthma action plans, could benefit from the prediction of asthma mortality to help prevent asthma-related fatalities and improve asthma healthcare outcomes.

However, these methods have their **limitations**. Frequentist GLMs may perform poorly if assumptions don't hold, while random forests are hard to interpret. Bayesian modeling may also not perform as well without prior distribution knowledge. More limitations are discussed later in the report.

Question 2: What is the causal effect of mean daily PM2.5 concentrations in each state on the prevalence of asthma in the U.S.? We hypothesize that higher levels of daily PM2.5 concentration cause higher prevalence of asthma.

We wish to identify PM2.5 as a potential health risk, so developing an argument that PM2.5 causes higher prevalence of asthma is much more compelling than simply finding correlations. Hence, causal inference method is a **good fit**.

Real-world decisions such as environmental policy, resource allocation, and health interventions could be made by answering this question.

- The results of this problem can be used to inform decisions on environmental policies aimed at reducing PM2.5 emissions.
- The results of this problem can be used by states with higher PM2.5 concentration levels to determine how to allocate/prioritize funding for projects that reduce PM2.5 emissions, improve air quality, and promote asthma prevention.
- The results of this question can inform health interventions for patients with asthma and help patients manage their asthma symptoms more effectively.

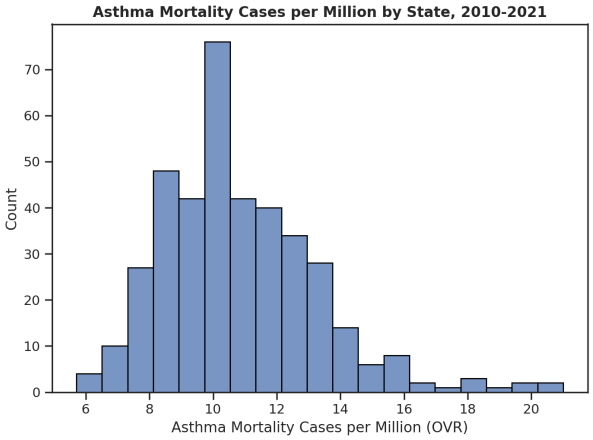
The **limitations** of the method we chose are:

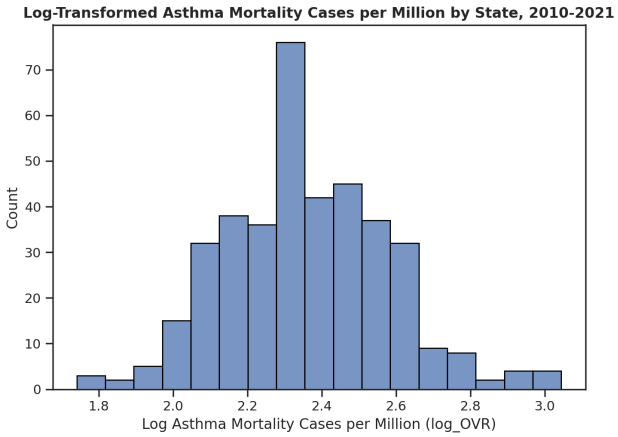
1. It may require domain knowledge to determine confounders.
2. It may not be possible to identify and measure all confounding variables.
3. It may not be able to control for selection bias.

Without identifying all the confounders, we might have a faulty interpretation of study outcomes. The causal inference might go wrong under selection bias, where certain groups are more likely to choose treatment. Selection bias can lead to an inaccurate estimate of the causal effect.

EDA

Quantitative Variables for GLM:

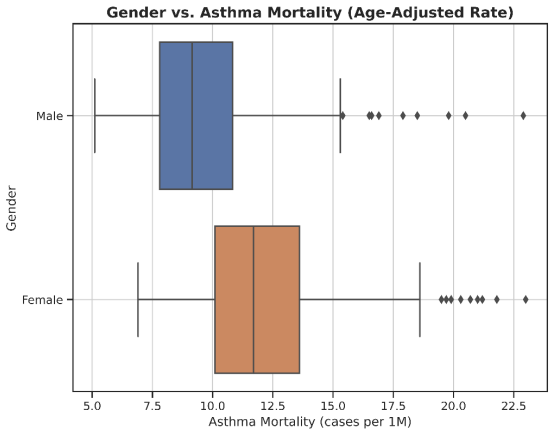


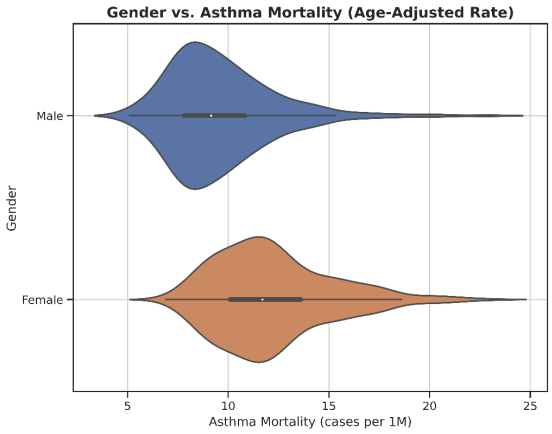


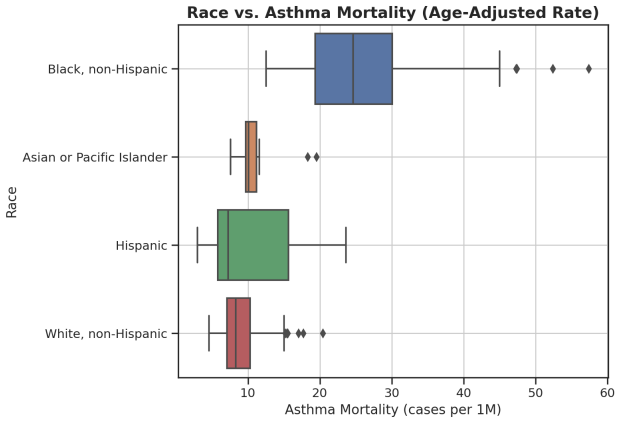
Trends & Follow-up: Overall asthma mortality cases per million (OVR) is the predictand of our model. According to the first histogram, it has a right-skewed distribution. Based on the Tukey-Mosteller Bulge Diagram, we take a log transformation on the dependent variable OVR to have an approximately normal distribution.

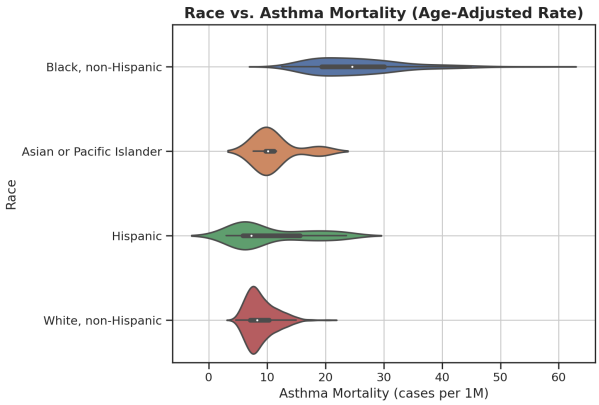
Relevancy: Since we use the linear regression model as our frequentist GLMs in the following part, normalizing OVR by applying a logarithmic scale might help to prove the linearity assumption of GLM.

Categorical Variables for GLM:







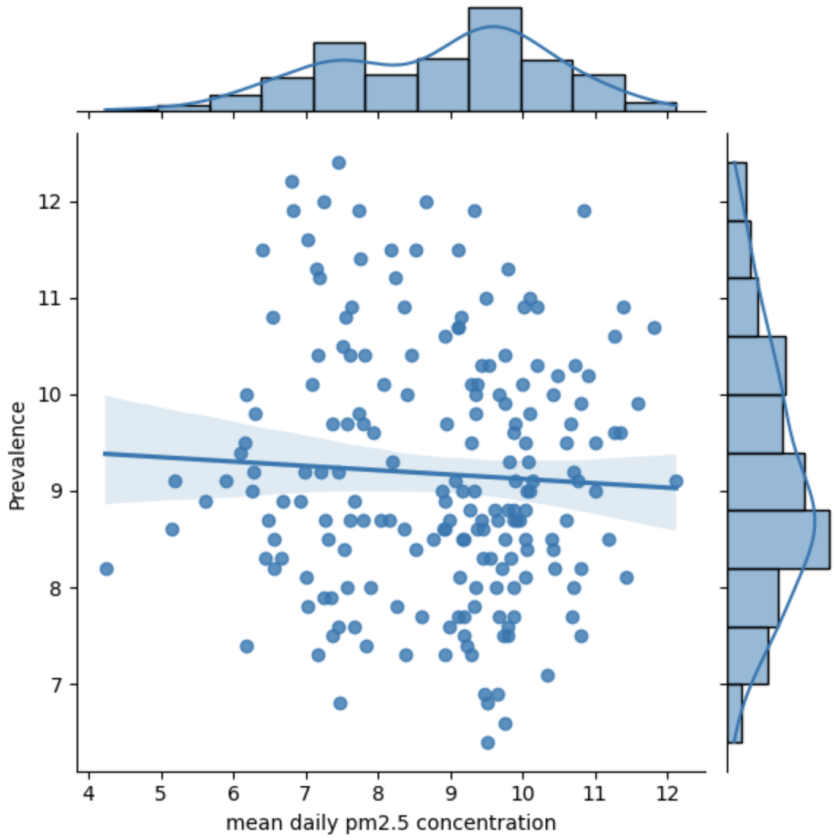


Trend: We noticed from the first two graphs that blacks and non-Hispanics have a higher average number of asthma deaths per million than other races, and the last two graphs indicate that females have more asthma deaths per million than males.

Follow-up & Relevancy: To understand the reasons for these trends, we plan to consider factors such as type of work, education, and health insurance coverage in the census data. Adding gender and race variables to our predictors could improve our model's accuracy, but we must be careful of any algorithmic bias towards minority groups.

Quantitative Variables for Causal Inference:

Jointplot: Age-Adjusted Prevalence vs Mean Estimated 24-Hour Average PM2.5 Concentration in $\mu g/m^3$



Data-Cleaning: We are mainly using two datasets: the U.S. Chronic Disease Indicators: Asthma ("the Asthma dataset") and Daily Census Tract-Level PM2.5 Concentrations, 2011-2014 ("the PM2.5 dataset"). The data cleaning process is summarized as follows:

1. The PM2.5 dataset has more than 100 million rows, with each data point representing the estimated 24-hour PM2.5 concentration at a specific location (i.e. a sensor station). Due to the large size of the PM2.5 dataset, we took the means of the estimated 24-hour PM2.5 concentration by year and state.
2. We filtered the Asthma dataset by restricting it to 2011-2014, the same time period as the PM2.5 dataset. We also only considered valid, *overall* asthma prevalence rate data.
3. We standardize the representation of states to their abbreviations by using the Federal Information Processing System (FIPS) Codes.
4. We merged the PM2.5 dataset and the Asthma dataset by state and year.
5. We plotted Prevalence vs Mean Daily PM2.5 Concentration, with each data point representing the values *per year* and *per state*. Note that we do not have 4 years * 50 states = 200 data points due to missing data in the Asthma dataset.

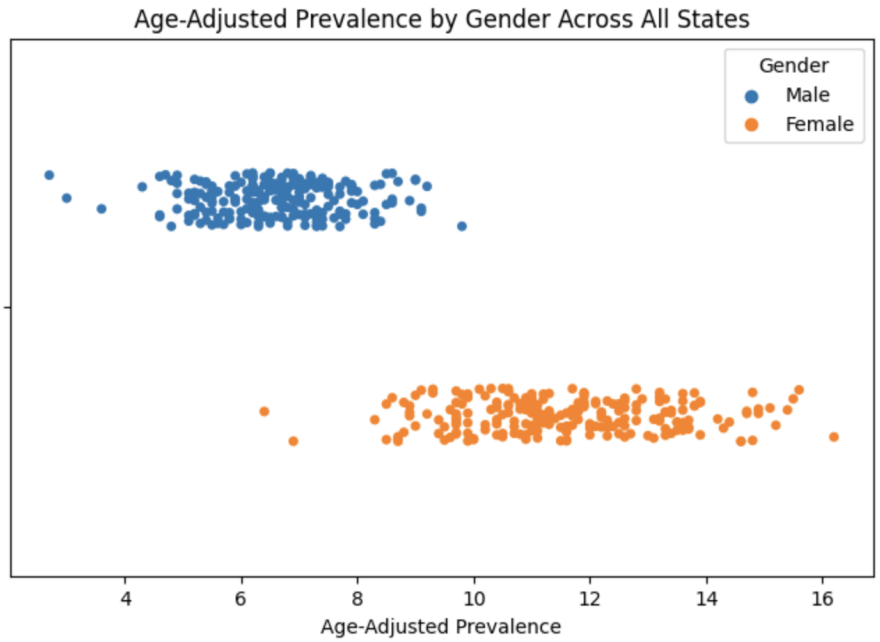
Trends: We observed that there is a negative correlation between the mean estimated 24-hour average PM2.5 concentration in $\mu g/m^3$ and the prevalence (Age-Adjusted). The correlation coefficient is -0.0381 with a p-value of 0.640. Since this p-value is larger than 0.05, we would conclude that there is not a statistically significant correlation between the two variables. We can see that many data points are far from the regression line. The top marginal histogram is a bimodal distribution since there are two peaks on the plot, which tells us the average PM2.5 concentration in most states is around 7.3 in $\mu g/m^3$ and 9.7 in $\mu g/m^3$. The right marginal histogram is left-skewed. The right marginal histogram tells us that the mode prevalence (Age-Adjusted) is around 8.5%.

Follow-up: We would like to follow up on the negative correlation between the mean estimated 24-hour average PM2.5 concentration in $\mu g/m^3$ and the prevalence (Age-Adjusted) since this relationship is against the popular idea that PM2.5 imposes a health risk. We would like to identify potential confounders and perform inverse propensity weighting to further analyze the question.

Relevant to Research Questions: The joint plot is motivating for our research question because it shows that there might be some confounders that can affect our treatment and outcome. Intuitively, there should be a positive correlation between the mean estimated 24-hour average PM2.5 concentration in $\mu g/m^3$ and the prevalence (Age-Adjusted). We need to figure out what kinds of factors produce a negative correlation.

Categorical Variables for Causal Inference:

Stripplot: Age-Adjusted Prevalence by Gender Across All States



Data-Cleaning: We are using the Asthma dataset. Similar to the prevalence vs PM2.5 plot, we restricted the Asthma dataset to 2011-2014. We considered valid asthma prevalence data based on *gender*. We then visualized the data in a strip plot, with each data point representing the age-adjusted prevalence *per year*, *per state,* and *per gender*.

Trends: Overall, females have a higher asthma prevalence than males. Males have many data points clustered around 5-8%, and very few above 9%. In comparison, females have a lot more data points above 9%, with one outlier going as far as more than 16%.

Follow-up: We would like to see how the result of the regression analysis of prevalence vs PM2.5 concentration would change, with the inclusion of gender as a covariate. We are also curious about how other asthma-related statistics, like the hospitalization rate and mortality rate, differ by gender.

Relevant to Research Questions: The objective is to help identify gender as a potential confounder in our causal inference. Intuitively, gender can influence the treatment (PM2.5 concentration). Different genders might be exposed to different levels of PM2.5 due to their occupations, oftentimes as a result of gender inequality. Gender can also influence the outcome (asthma prevalence) for biological/medical reasons like human anatomy. The above plot shows that different genders indeed have different outcomes, so we should further investigate it as a confounder.

Option C: Prediction with GLMs and nonparametric methods

a. Methods

We aim to accurately predict the number of asthma mortality cases per million using a combination of relevant census variables and an asthma dataset for each state and year. Through exploratory data analysis, we recognized the need for additional census data to provide better insight into the factors that contribute to higher asthma mortality cases among certain ethnicity and gender groups. By doing so, we hope to develop a better understanding of the various factors that contribute to higher asthma mortality rates and develop effective interventions to address them. Our objective is to build a model that can establish strong relationships between our features and the outcome variable, OVR, to improve our prediction accuracy.

To achieve accurate predictions, we initially selected 27 economical and social variables from the census data that we believed would help explain the outcome variable, such as poverty_family percent estimate, unemployed_PE, etc. Our modeling process involved testing different feature combinations and evaluating their corresponding model-fit and performance metrics.

1. Frequentist and Bayesian GLM

Justification:

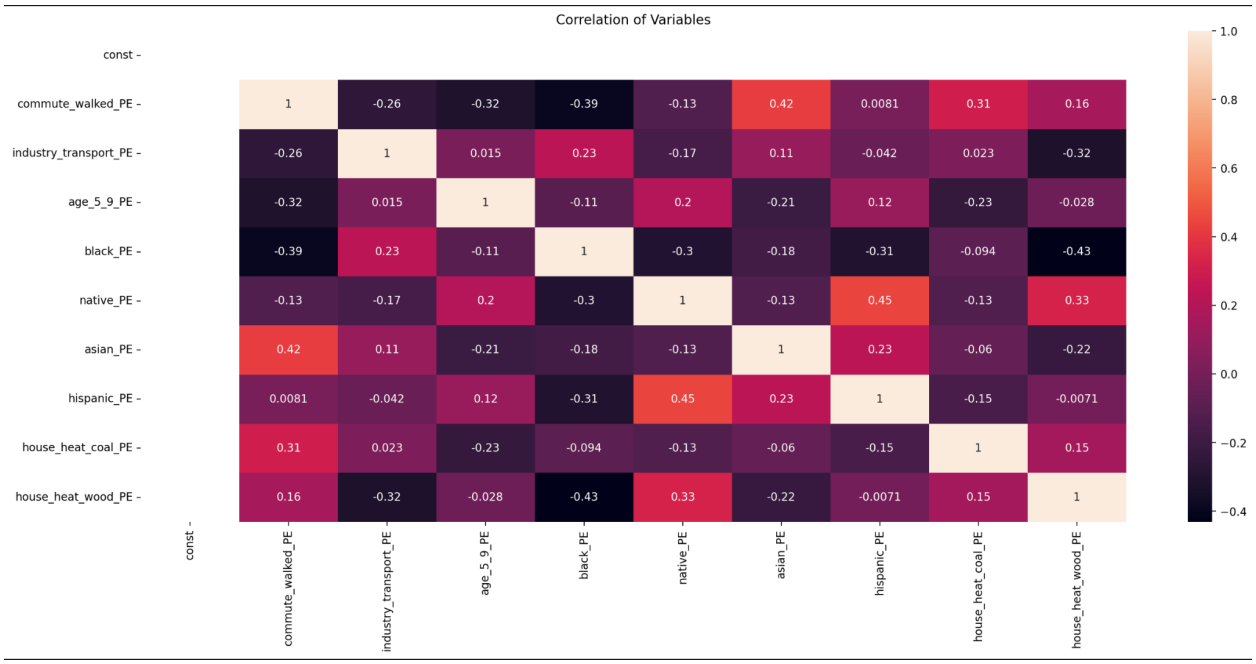
We will be using a Frequentist and Bayesian linear regression model for our analysis because our outcome variable is continuous and can take on any real value. In contrast to other GLMs, which are designed for binary or count data, linear regression allows for flexible choices of likelihood and link functions that can be tailored to our specific dataset. However, this method is computationally intensive and can be uncertain in determining variable coefficients, which will be discussed further in our limitations section.

Feature Assumptions:

1. There is no multicollinearity among our filtered variables.
2. There is no autocorrelation of error terms.

We checked for multicollinearity among our selected features using a correlation heatmap and variance inflation factor (VIF), which measures the extent to which the variance of the regression coefficients is inflated due to multicollinearity. We filtered out variables with a VIF value greater than 100, leaving us with around 7 variables.

To test for autocorrelation, we used the Durbin-Watson test, which produces a value between 0 and 4. Our test showed that there is little to no autocorrelation in our data with a value of 1.97, which falls within the range of values indicating no autocorrelation.



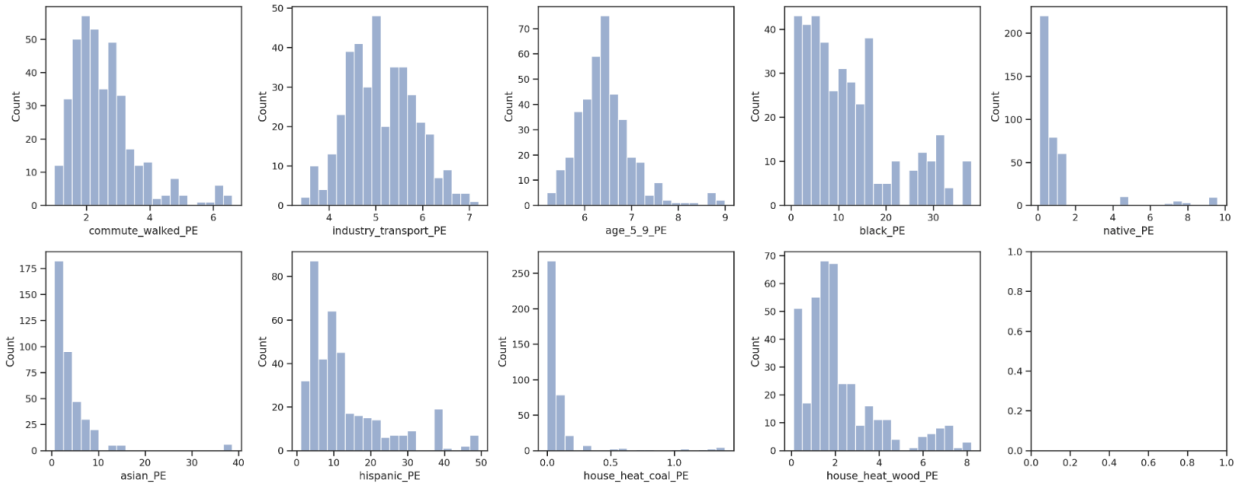
```
Performing Durbin-Watson Test
Values of 1.5 < d < 2.5 generally show that there is no autocorrelation in the data
0 to 2< is positive autocorrelation
>2 to 4 is negative autocorrelation
-------------------------------------
Durbin-Watson: 1.9769991166382594
Little to no autocorrelation

Assumption satisfied
```

Feature Engineering:

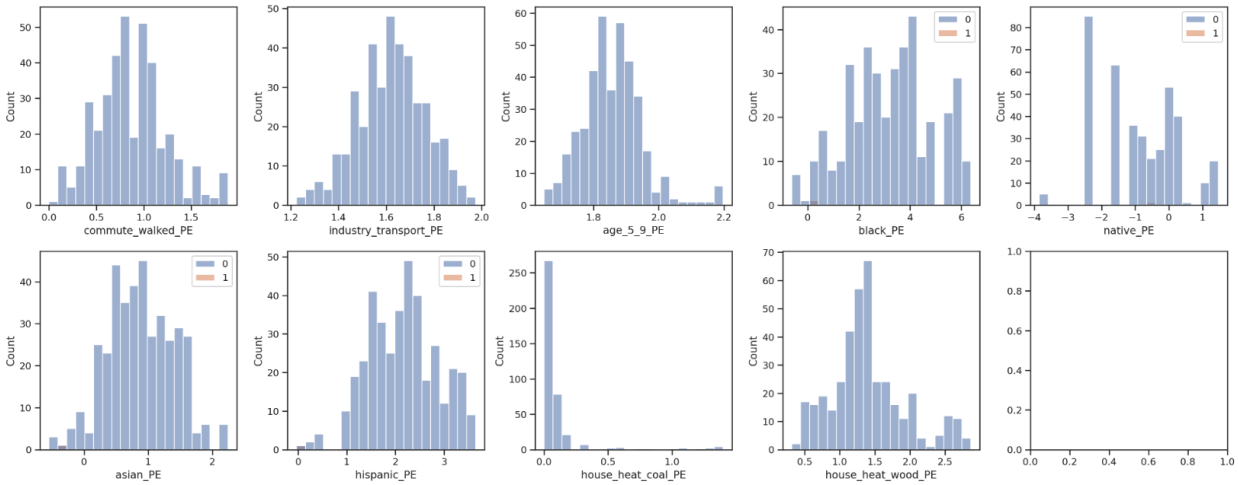
There were also attempts of feature transformation and adding PM2.5 data from our casual inference data. However, these transformation and addition of PM2.5, not only made our dataframe size reduce in half, and were insignificant according to our p-value significancy.

Before Transformation:



After transformation:

1. transformed the first three variables using np.log.
2. Race variables using boxcox to put each race on the same scale
3. np.sqrt house_heat_wood_PE because of their right skewness

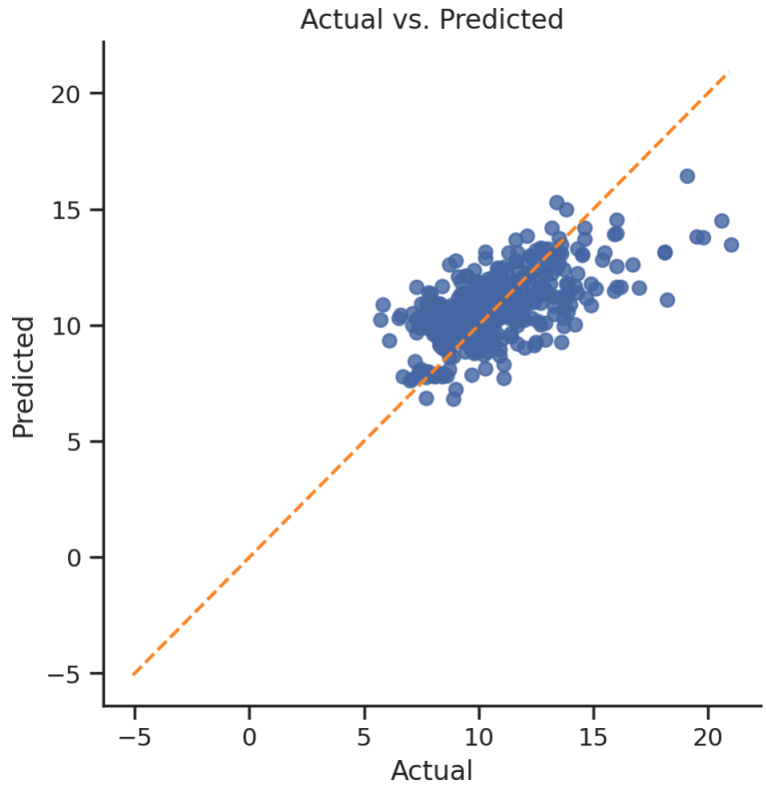


Model Assumptions:

In our frequentist GLM model, we used a Gaussian distribution as our likelihood and identity link function, which is similar to an OLS but with fewer assumptions required according to this reference. We made the following assumptions in our GLM model:

1. The overall mortality cases in OVR are independently distributed.
2. OVR can follow an exponential family distribution, but we are assuming normal from its residual histogram because Gaussian has better statistical properties: Refer to the EDA section.
3. There exists a linear relationship between the transformed predictor variable in terms of link function and the outcome variable.
4. Residuals need to be independent but NOT normally distributed.
5. Homogeneity of variance does NOT need to be satisfied.

To demonstrate linearity, we created a scatter plot of actual vs predicted:



Before applying any variable transformation, there seems to be a linear relationship. However, this relationship is not perfect where there is some bias towards higher values around 16 - 20.

For Bayesian regression, we used a default flat prior and normal likelihood because we did not have relevant public health expertise. We chose a normal likelihood, assuming that the vast majority of y-values we see will be within 3 standard deviations from the prediction X * beta. We used a flat prior to cover all possible ranges of values, which may not be optimal in certain cases but allows for more flexibility in our model.

All assumptions tested and visualized are inspired by the model summary plot and Jeff Macaluso's post.

Performance:
We used RMSE to assess model performance out of the sample for frequentist models. We also evaluate our Bayesian models' performance by posterior predictive checks and compare the mean highest density intervals (HDI) for each variable in the Bayesian models with the frequentist's best model.

2. Non-parametric Method - Random Forests

We chose random forests as our nonparametric method.

Justification:
Random forests obtain decisions by majority voting on classification and averaging in regression among a bunch of decision trees. Thus, it can use regression analysis to predict asthma mortality cases.

Compared to the k-NN algorithm, random forests can work in high-dimensional spaces; they prevent overfitting compared to decision trees; and they require less computational cost than neural networks.

Assumption:
As a non-parametric method, the random forest doesn't make any assumptions about the distribution of data, but it relies on a common assumption that sampling is representative.

Performance:
We use RMSE to evaluate the performance of random forests.

b. Results

The best frequentist GLM for our analysis is the one that uses a Gaussian likelihood and Identity link function with untransformed filtered variables. This model has a training error of 0.1864 and a test error of 0.1586 in terms of root mean square error (RMSE).

The best Bayesian GLM has a Gaussian likelihood and a flat prior with untransformed filtered variables. The mean highest density interval (HDI) is closest to the frequentist's best model, and the posterior predictive distribution is less noisy and concentrated around the true distribution.

The best non-parametric model is a random forest using all features and a log-transformed outcome variable, with a training error of 0.0630 and a test error of 0.1476 in RMSE.

We used confidence intervals, p-value significance, and mean HDI to estimate variable coefficients, and its corresponding estimates indicate an increase of 1 unit in X resulting in a corresponding increase in log_OVR.

c. Discussion

1. Model Comparison and Differences

Random forests outperformed frequentist and Bayesian GLMs with the lowest test RMSE, likely due to their ability to handle noisy data and large numbers of features. The choice of model depends on the purpose of the analysis, with GLMs being more interpretable for policy-making and Bayesian GLMs providing more uncertainty estimates but being more computationally intensive. Frequentist GLMs provide point estimates for coefficients and are less computationally intensive.

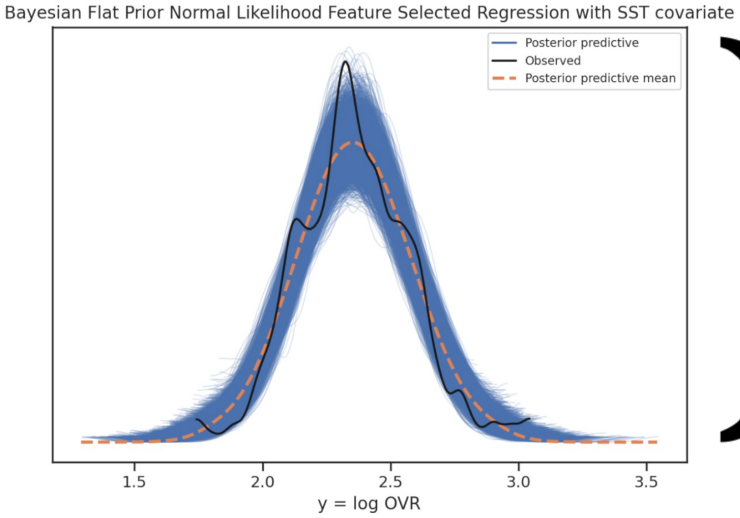
2. Model Fits

We used different methods to fit the model to the data. For frequentist models, we used summary functions and metrics like log-likelihood, pseudo-R-square, and RMSE. Bayesian models were evaluated using HDI distributions to determine coefficient peaks. Non-parametric methods have no assumptions but may require a log transformation of the outcome variable for better predictions.

3. Interpret results

Frequentist GLM training RMSE: 0.1864 → $e^{0.1864} \approx 1.2049$
Random forest training RMSE: 0.0630 → $e^{0.0630} \approx 1.0650$
Frequentist GLM test RMSE: 0.1586 → $e^{0.1586} \approx 1.1719$
Random forest test RMSE: 0.1476 → $e^{0.1476} \approx 1.1590$

} Difference between the predicted and true number of age-adjusted asthma mortality cases per million.



} Difference between the predicted and true number of age-adjusted asthma mortality cases per million.

4. Limitation

1. The frequentist GLM assumes no multicollinearity and no autocorrelation, which may not hold in the real-world dataset.
2. The Bayesian GLM may suffer from a lack of prior knowledge about the data.
3. Random forests may be challenging to interpret and store due to their complexity.
4. The analysis was conducted on a small dataset with limited feature categories.

Additional data related to asthma mortality might be useful, such as family asthma history, asthma healthcare, and resources, etc.

5. Model Improvements

Our initial model was to model all variables, but as we perform feature engineering and matching frequentist model assumptions, we use RMSE to filter variables and add variables against log OVR.

Although random forests are "out-of-the-box" and do not require feature selection or feature engineering, inspired by this post, we tried to improve model performance by using different combinations as follows:

1. All features VS. Selected features (non-log outcome variable)

All features + non-log:

```
Training set error for random forest: 0.7297973450452977
Test set error for random forest:     1.5782184643982342
```

Selected features + non-log:

```
Training set error for random forest: 0.7240977738403744
Test set error for random forest:     1.6856993577700388
```

2. Non-log VS. Log outcome (using all features, proved to be better in the above case)

Non-log outcome + all features:

```
Training set error for random forest: 0.7297973450452977
Test set error for random forest:     1.5782184643982342
```

Log-transformed outcome + all features:

```
Training set error for random forest: 0.06300288786133221
Test set error for random forest:     0.14759178865159248
```

3. Decision Trees VS. Random Forests (all features + log outcome)

Decision tree:

```
Training set error for decision tree: 0.0
Test set error for decision tree:     0.21153515849647855
```

Random forest:

```
Training set error for random forest: 0.06300288786133221
Test set error for random forest:     0.14759178865159248
```

→ We can see that the random forest with all features and a log-transformed outcome has the lowest test error (RMSE).

1. Frequentist GLM Transformed Variable

```
Training set error for linear model: 0.19956275558576336
Test set error for linear model:     0.1667642563642375
```

2. Frequentist GLM Untransformed Filtered Variable

```
Training set error for linear model: 0.18636193014088825
Test set error for linear model:     0.15862360407072915
```

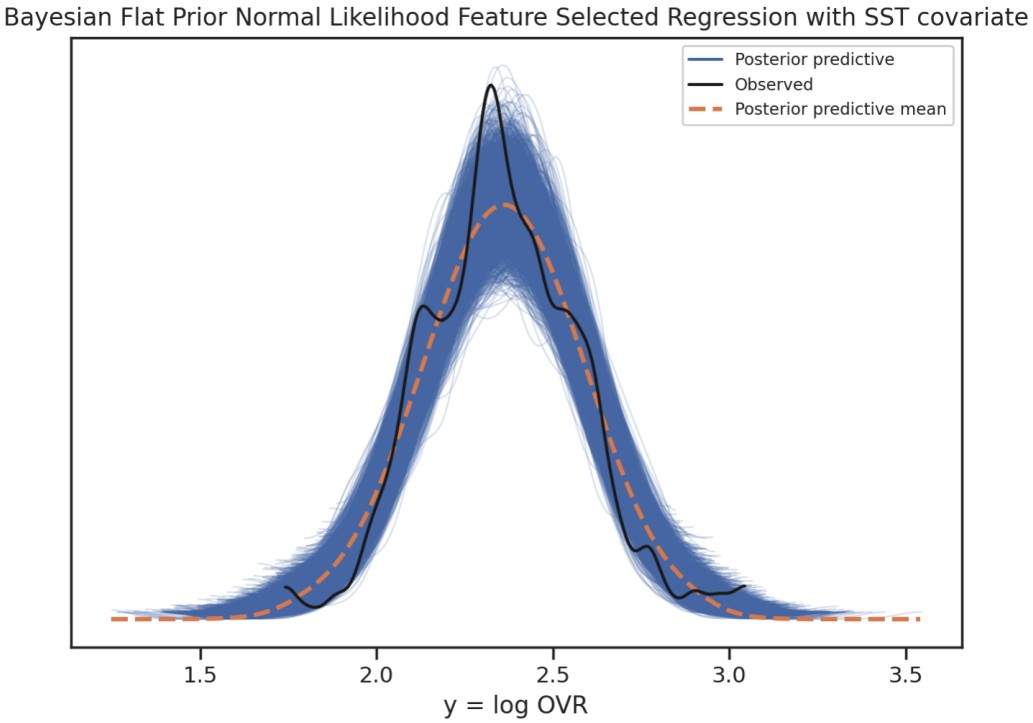
3. Frequentist GLM Transformed Filtered Variable + PM2.5

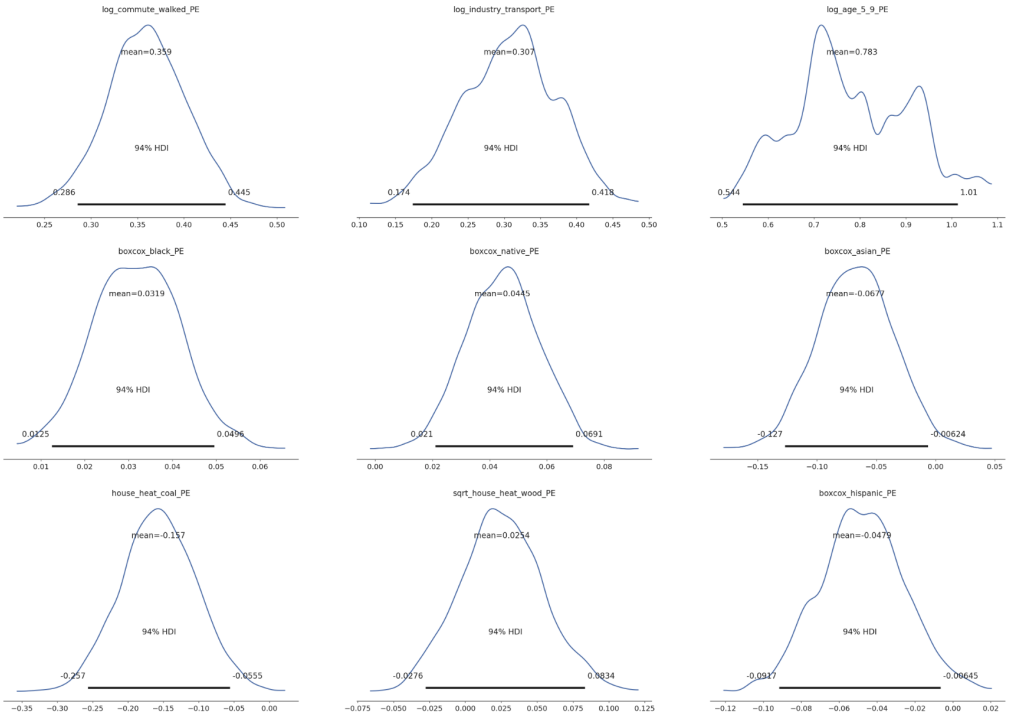
```
Training set error for linear model: 0.3031925451346434
Test set error for linear model:     0.22980252455553105
```

We can see here that transformation and adding the PM2.5 censor variable did not help our model predict accurate results. The best method would be Frequentist GLM Untransformed Filtered Variable.

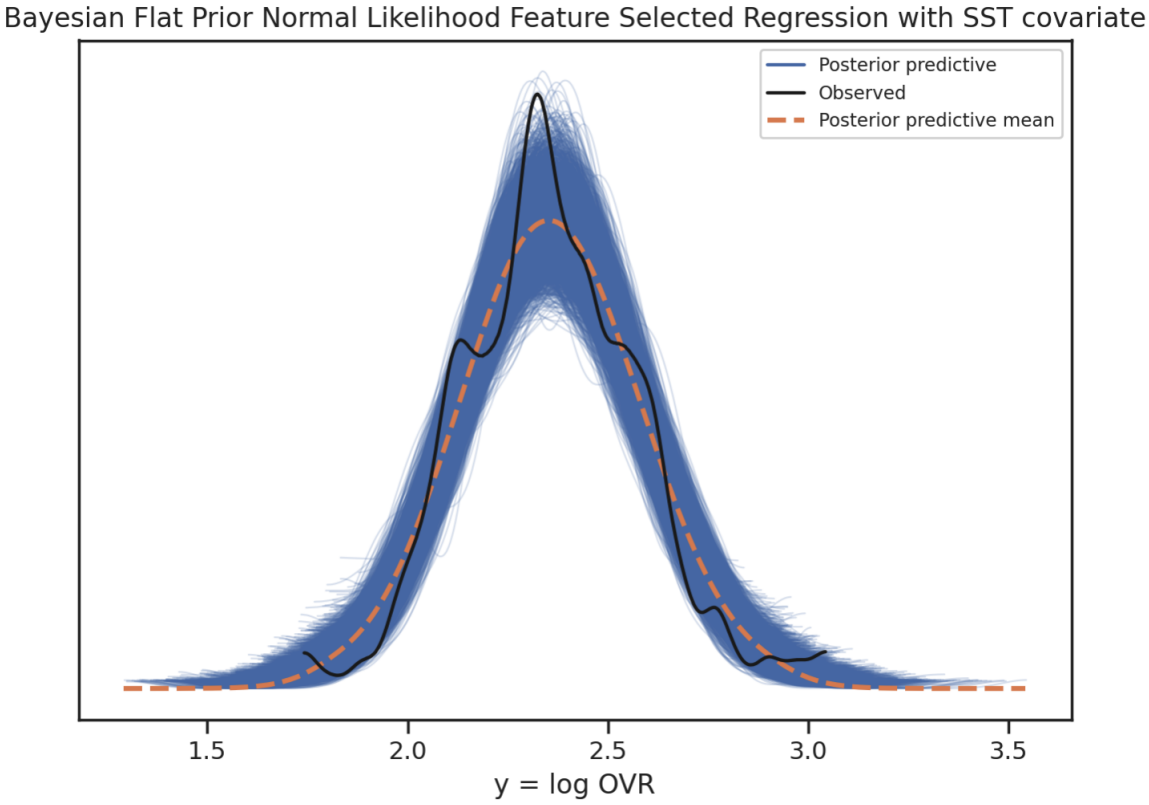
For Bayesian GLM, we used Posterior Predictive Checks and compared coefficients' HDIwith the frequentist GLM best model.

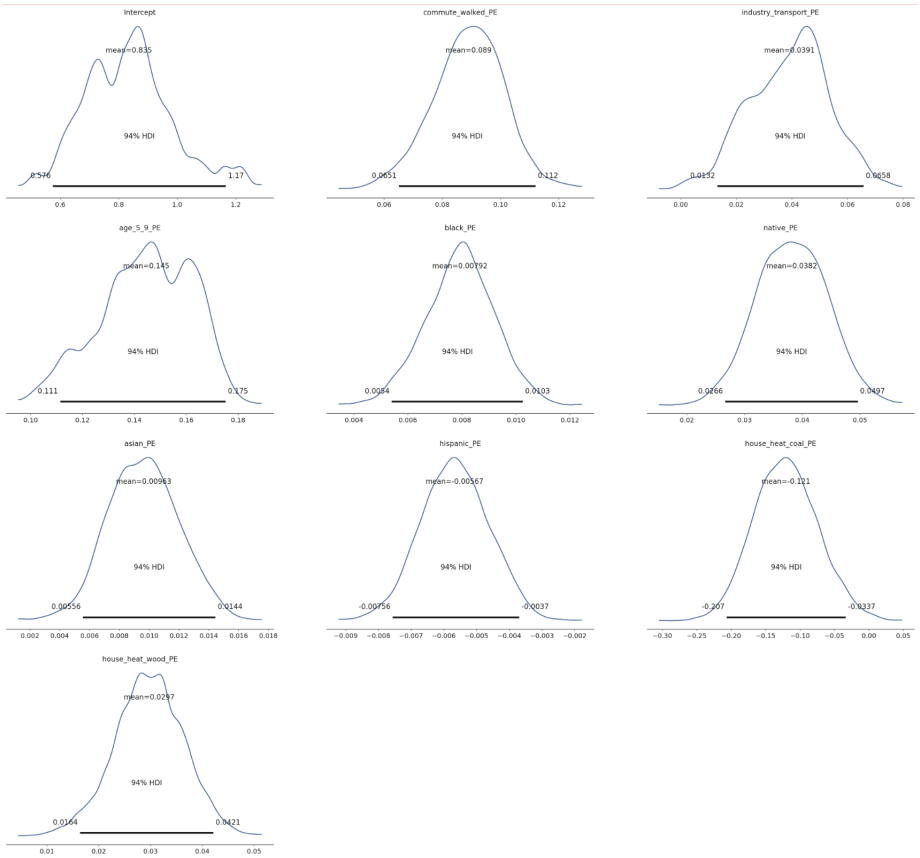
1. Bayesian GLM Transformed Variable



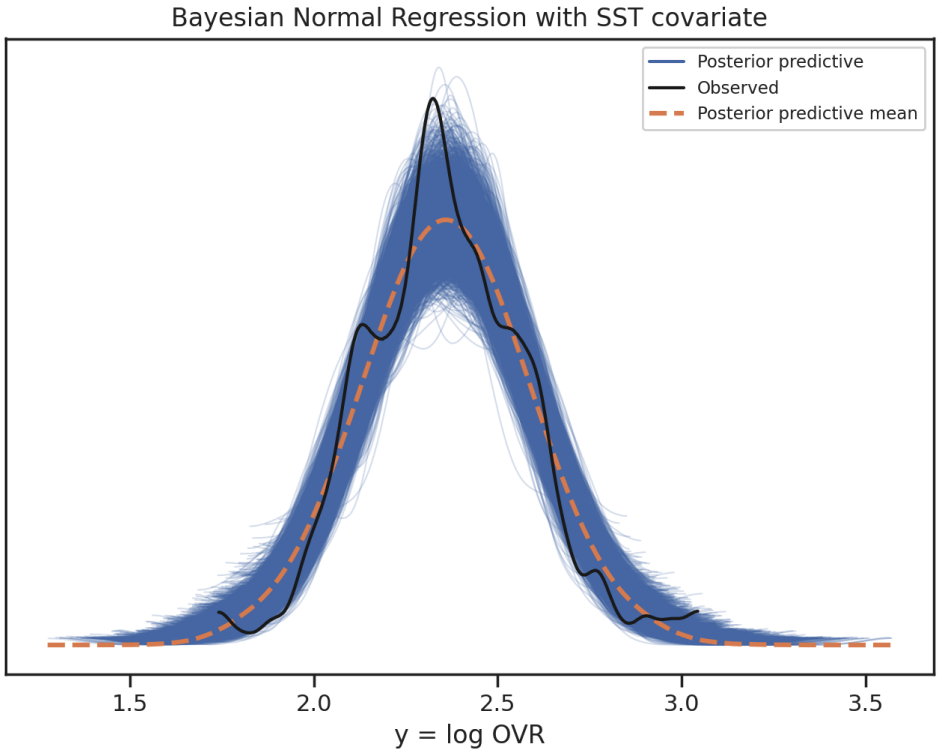


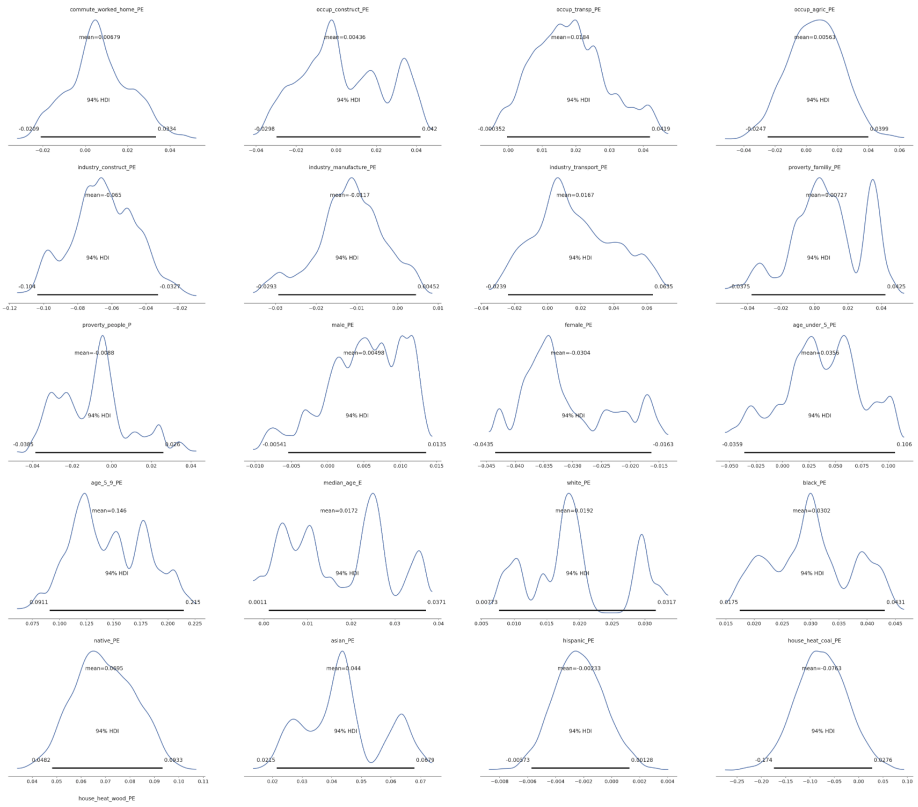
2. Bayesian GLM Untransformed Filtered Variable





3. Bayesian GLM all features





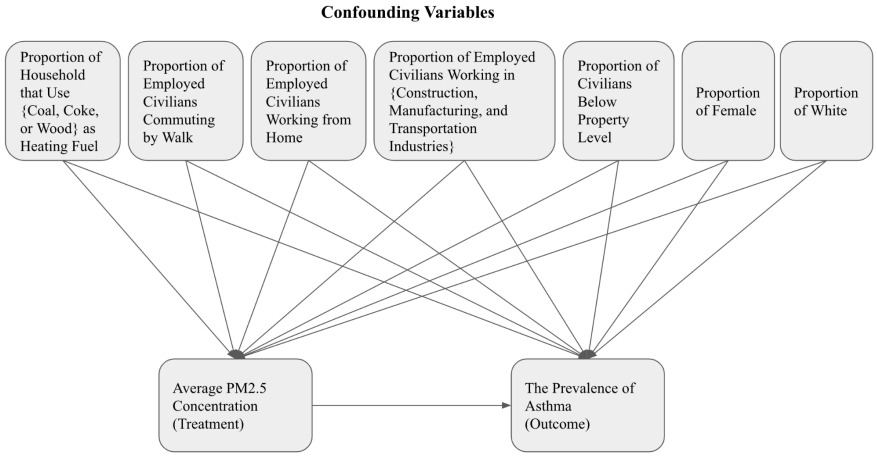
Based on the results of the predictive posterior check and numerical HDI, we conclude that the untransformed Bayesian regression model is the best Bayesian model for our dataset because it has the lowest noise in the posterior predictive checks, the coefficient estimate has one peak in HDI, and it aligns well with the frequentist version.

Option D: Causal Inference

General Research Question: What is the causal effect of mean daily PM2.5 concentrations in each state on the prevalence of asthma in the U.S.?

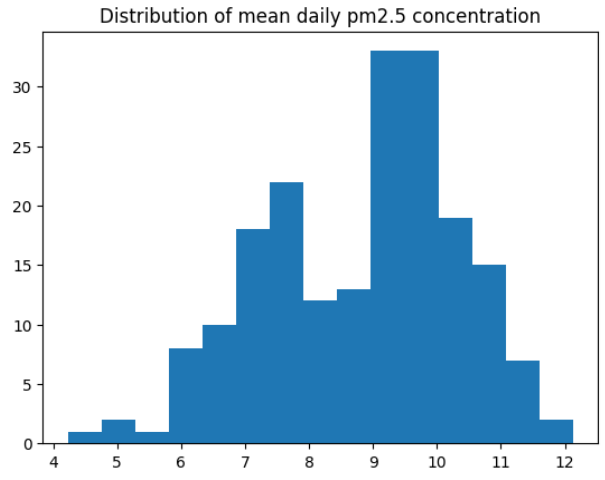
a. Methods

- Treatment variables: Whether the population of a specific state during a specific year is exposed to an annualized daily average PM2.5 concentration of more than 10 μg/m^3(>10 μg/m^3 is positive treatment, ≤ 10 μg/m^3 is no treatment)
- Outcome variables: The prevalence of asthma
- Unit: per state per year
- Confounding variables:
    - Proportion of households that use {coal, coke, or wood} as heating fuel
    - Proportion of employed civilians commuting by walk
    - Proportion of employed civilians working from home
    - Proportion of employed civilians working in {construction, manufacturing, and transportation industries}
    - Proportion of female
    - Proportion of white
    - Proportion of civilians below poverty level
- Methods used to adjust for confounders:
    - We first fit a logistic regression model to predict Pr (daily exposure to PM2.5 above 10 μg/m^3 | confounders) to obtain the propensity scores.
    - We trim our propensity scores to [0.05, 0.95].
    - We then use Inverse Propensity Weighting (IPW) to compute the adjusted average treatment effect (ATE).
- Colliders: None
- The causal DAG for the variables:

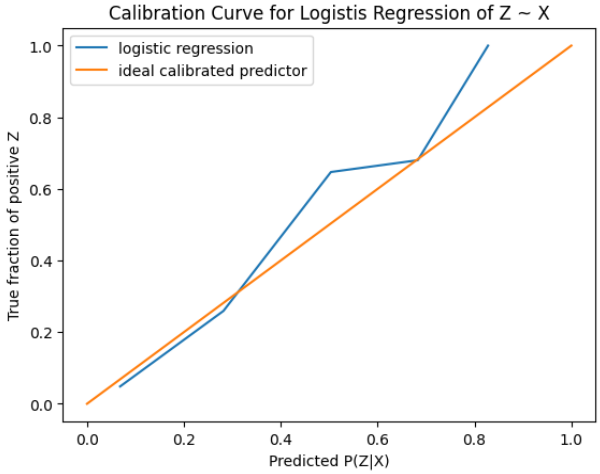


b. Assumptions and Results

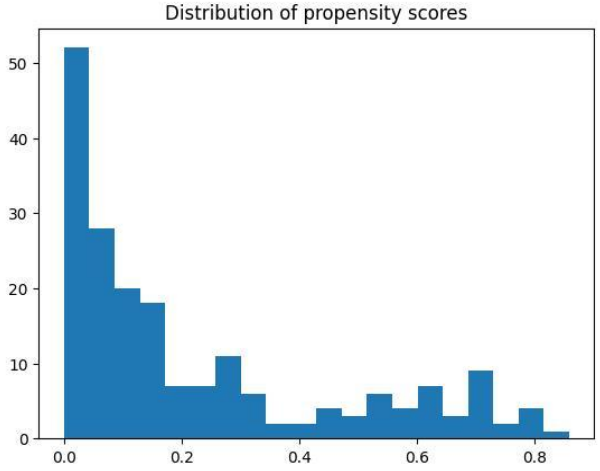
We assume that the outcome is unconfounded given the confounding variables listed above. We also assume our propensity score is an accurate model of P (Z = 1|X = x). It follows that IPW would give us a good estimate of the causal effect. Using IPW, we found that one unit increase in annualized average PM2.5 concentration causes the prevalence of asthma to increase by roughly 0.82(%).



Based on the distribution of the mean daily PM2.5 concentration, we set the threshold for our treatment as:
If the population of a specific state during a specific year is exposed to an annualized daily average PM2.5 concentration of more than 10 $\mu g/m^3$, Treatment (Z) = 1.
If it is less than 10 $\mu g/m^3$, Treatment (Z) = 0.
By setting this threshold, around 22.45% of our data points are greater than the threshold, and 77.55% of our data points are less than the threshold.

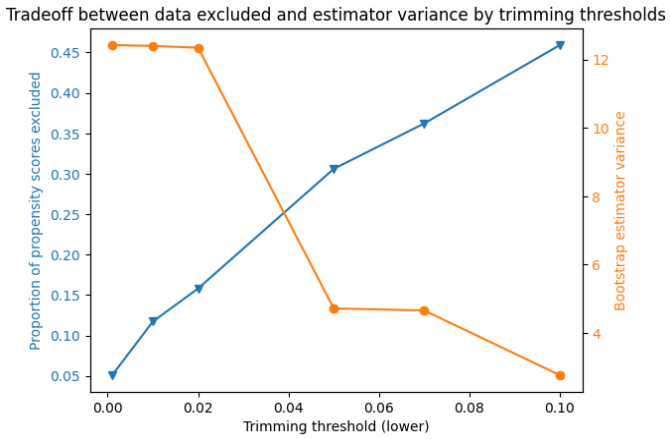


We used the Logistic Regression to compute the propensity score by fitting our confounding variables into the treatment.
We then used the calibration curve on the left to see that our model is generally well-calibrated.



The confounding variables we use to model the treatment can differentiate them from the general population.
Due to label imbalance in treatment (as most states in most years have a low PM2.5 concentration), we can see that a lot of our propensity scores are close to zero.

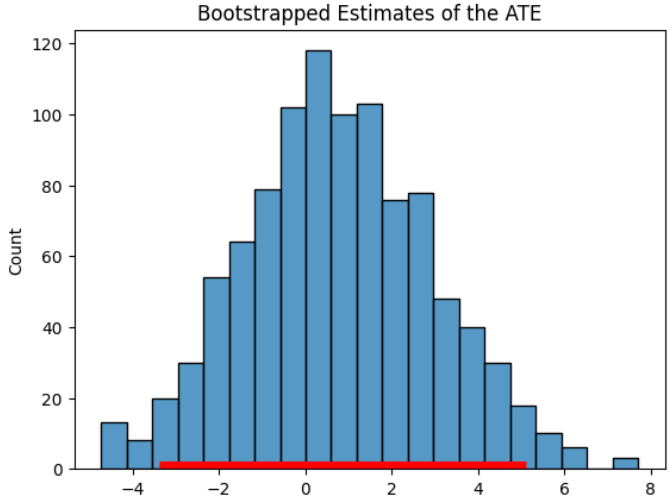
Trimming Procedure for IPW



Since we have a significant number of propensity scores that are close to zero, we aim to find a trimming threshold where 1. we do not trim too much of our data, and 2. we obtain an estimator with a reasonably small variance. Hence, from the plot to the left, we can see that trimming to [0.05, 0.95] gives us a balance between the two criterions.

We trimmed our propensity score by removing any data points with propensity scores that were too low or too high and keeping the propensity scores between 0.05 and 0.95. Finally, we computed the inverse propensity weighting estimate for the ATE. The estimated average treatment effect is **0.82(%)**.

Uncertainty Quantification for IPW:



By performing bootstrapping on our ATE estimator, we found that our estimator has a variance of 4.72, with a 95% confidence interval of [-3.393, 5.117].
The ATE estimator has more extreme positive values than negative values.

c. Discussion

The hypothesis we are investigating is that people who are exposed to an annualized daily average PM2.5 concentration of more than 10 μg/m^3 would more likely have a higher prevalence of asthma. However, as detailed in the EDA section, without the confounding variables, there seems to be a negative correlation between them, which is evidence against our hypotheses.

After taking into consideration the confounding variables that would affect our treatment, we found a positive correlation, which is more in line with the intuition that PM2.5 could harm our respiratory system and lead to diseases like asthma.

Here are some limitations of our method:

1. We may not be able to identify and measure all confounding variables.
2. Label imbalance led the logistic regression model for the propensity score to output very small probabilities, which is problematic for computing IPW.
3. Since each unit is one state in one year, the granularity of our data is high. It will be preferable to have data on individuals, but it is not feasible due to privacy concerns.

Additional data that would help further answer the causal question: We identified the confounding variables based on our domain knowledge. It would be useful to have data to establish the causal relationship between confounders and treatment.

Through IPW, we obtained an ATE of 0.82%, which is a positive causal relationship between our treatment and the outcome. By inspecting the bootstrap ATE distribution, we can see that the ATE estimator has more extreme positive values than negative values. However, considering that the variance of our estimator is high, we suggest that further study (ideally with data of higher granularity) is needed to confirm the causal relationship we found.

Conclusion

1. Key Findings

GLM:
We developed three models to study the relationship between asthma mortality and its features. The random forest model performed best with an **RMSE of about 1, indicating an average error of ±1 cases per million**. Frequentist GLM models are more interpretable, while non-parametric methods predict a precise mortality rate, and Bayesian models perform better with expert knowledge.

Causal Inference:
We merged the dataset for PM2.5 concentration and the dataset for asthma prevalence to analyze the causal relationship between the two variables on a per-state per-year basis. We identified many confounding variables, like commute time, and incorporated the census data to account for them. Using inverse propensity weighting, we found that an annualized daily PM2.5 concentration of more than $10\mu g/m^3$ causes **0.82% more** asthma prevalence.

2. Generalizability, scope, limitations

GLM:
Our study is limited to the specific time period and geographic area of the United States from 2010-2021. Our findings are based on the available data sources, and may not be generalizable to other countries or regions with different healthcare and social factors.

Casual Inference:
As hinted in the results section, our analysis is limited by the data source. Due to privacy concerns, all census data are aggregated, giving us data with relatively low granularity. Moreover, the PM2.5 data is limited to 2011-2014, so some parts of our findings could be outdated. Since our data is limited to the U.S., a causal relationship between PM2.5 exposure and asthma prevalence in the U.S. may not be applicable to other countries with different demographics or environmental conditions.

3. Call-to-action

GLM:
Our model allows us to predict state-level asthma mortality using census data. We can use a color-coding system to classify states based on predicted asthma mortality(per million cases): <= 5 marked state green; 5 - 10: yellow; 10 -20: red. These ranges were determined based on our dataset and statistics from asthma dataset.

States marked yellow are recommended to increase public awareness campaigns, provide additional funding for research, and develop legislation and regulations to promote asthma prevention and management.

For states marked red, immediate actions are needed, such as increased production of asthma medications and increased access to medical resources.

Causal Inference:
Considering the overall asthma prevalence in the U.S. to be 7.7% [Asthma Facts], PM2.5 is clearly a health risk. Based on our results, we think improving air quality by implementing policies and regulations to reduce PM2.5 emissions is important. For example, increasing the use of public transportation, incentivizing the use of electric vehicles, or investing in developing more clean energy sources.

4. Data Merging

The data merging process for casual inference is explained in its EDA sections. For GLM, we combined census and asthma data to provide a more comprehensive analysis of the factors contributing to asthma mortality rates. One benefit of this approach is that we were able to explore relationships beyond ethnicity factors and identify significant variables in the census data. However, a potential consequence is that if the assumptions for merging the data were incorrect, the combination may not be accurate. Additionally, some column variables in the data were not well-defined, as we initially mistaken % per 1M instead of cases per 1M, and these could have affected the accuracy of our analysis.

5. Future studies and lessons learned

For casual inference, exploring the interactions between PM2.5 exposure and other risk factors contributing to the prevalence of asthma is one future study that can be built upon in our work, such as dietary habits, genetics, etc. This could support targeted prevention and intervention strategies by identifying vulnerable populations.

For GLM, although we were able to predict state-level asthma mortality rates, we were unable to analyze them at a more granular level due to data limitations. With more individual-level data, additional factors such as pollution rates and weather conditions could be considered to narrow down predictions. However, unforeseen events can still impact our predictions, and time series forecasting methods may be useful to account for unexpected events over time. Overall, our study provides insight into the potential factors that contribute to asthma mortality rates in the United States, but further research is needed to validate and expand upon our findings.

Throughout this project, we gained valuable experience in building models with accurate assumptions and preprocessing features to improve our predictions. We learned how to clean a big dataset, such as census data. We have strengthened our knowledge of causal inference learned in class and applied what we learned to real-world problems during the project. Additionally, we learned about each other's expectations of the project and planned our work accordingly to complete it on time.

Works Cited

Census Data API: /Data/2021/Acs/Acs1/Profile/Groups. api.census.gov/data/2021/acs/acs1/profile/groups.html.

Daily Census Tract-Level PM2.5 Concentrations, 2011-2014 | Data | Centers for Disease Control and Prevention. 25 July 2018, data.cdc.gov/Environmental-Health-Toxicology/Daily-Census-Tract-Level-PM2-5-Concentrations-2011/fcqm-xrf4.

Davide, ND. "Log-Transforming Target Var for Training a Random Forest Regressor." *Cross Validated*, stats.stackexchange.com/questions/447863/log-transforming-target-var-for-training-a-random-forest-regressor.

Macaluso, Jeff. "Testing Linear Regression Assumptions in Python." *Jeff Macaluso*, 27 May 2018, jeffmacaluso.github.io/post/LinearRegressionAssumptions.

Mortality Mapping Help - Section 3.2 | WISQARS | Injury Center | CDC. www.cdc.gov/injury/wisqars/mapping_help/age_adjusted.html.

Kjhealy. "Fips-codes/state_and_county_fips_master.csv at Master · Kjhealy/Fips-codes." GitHub, 1 Jan. 2018, github.com/kjhealy/fips-codes/blob/master/state_and_county_fips_master.csv.

U.S. Chronic Disease Indicators: Asthma | Chronic Disease And Health Promotion Data And Indicators. 30 Jan. 2023, chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-Asthma/us8e-ubyj.

US Census Bureau. "American Community Survey 1-Year Data (2005-2021)." Census.gov, 8 Sept. 2022, www.census.gov/data/developers/data-sets/acs-1year.html.