US Air Pollution Report

To understand and maintain US air pollution from linear model level

Wayne Wang

INTRODUCTION

Since I am working alone on my project, I have reduced the nonessential analysis from inference and only included the relevant and significant performance models. Hope you have fun reading this report! I think I really made something work :)

Air pollution has always been a key to our development and as of today, countless people have dedicated themselves to preventing further environmental damage. Induced by different factors, **air pollution levels in different states are widely different**, thanks to their individual differences in policies, political interests, and more. **But specifically, are they different from each other in statistical ways? Are CO AQI worse in CA than in Alaska? Which model would be better at predicting pollution levels if we want to capture such a difference?** To answer these questions, I will use the **bootstrap** procedure with **linear**, **shrinkage**, and **non-linear** models to try to interpret the dataset and conduct **model selection** with **bootstrapped confidence intervals**.

From the previous EDA, I have shown visible trends for each type of pollutant with respect to the different States. **This suggests that there are potential methods that we can use to try to capture the nature of air pollution in different states so that people and government can obtain a better understanding of the air quality of some states on a broader scale with a scientific lens.**

Specifically, I choose to work on the US Air Pollution data in the project datasets to explore how to correctly apply models of different choices to predict air pollution levels of different states. This dataset includes documented Air Pollution data like CO from 2000 - 2016. I will introduce this project step by step below.

In this project, I hope to:

1. Engineer dataset to use "State Code" and "Pollutant" as dependent variables to try to predict **" Mean AQI" (referred to as AQI in later parts)**
2. Use multiple models tested with **Bootstrap** to compare the performance and see if we can help make sense of air pollution

These two questions can be answered with models with shrinkage methods and non-linear approaches, and could potentially benefit people by providing a data analytic view of pollution, combined with real-world incidents we could take impactful actions in the future to prevent or predict air pollution.

Engineering:

As per my introduction, I want to focus on 1. State Level difference, 2. Pollutant difference, and see that by having state and pollutant type as independent variables how I can predict the dependent variable, which is the AQI level of that specific pollutant in that state, with different models. To do so, I first engineered my dataset to look like this:

	State Code	Pollutant	AQI
0	1	CO	3.850288
1	1	NO2	21.232246
2	1	O3	36.845170
3	1	SO2	7.005115
4	2	CO	6.528340

Now I have the dataset ready for further analysis, I can explore the performances of different models.

Models:

Now that we have the dataset ready, we want to primarily focus on applying three types of models provided in the project spec. The models I have chosen are the simple linear model, the ridge regression model, and lasso regression model, and the random forests. I want to apply these three models and test their performances by applying the **bootstrap** procedure, which means I will be running a bootstrap sample 200 times to generate a confidence interval for r squared and a confidence interval for RMSE, in this way, we can better account for the errors and explain model performance. The general setup for the models are:

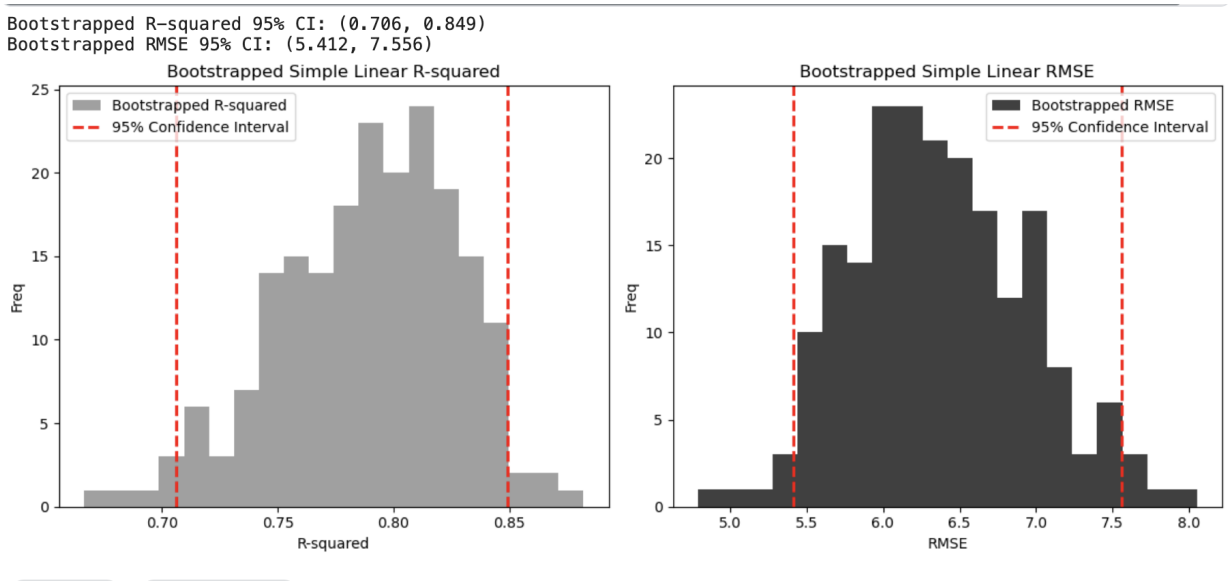
1. **Simple Linear Model**: AQI ~ State Code + Pollutant
2. **Shrinkage: Ridge Regression and Lasso Regression**: AQI ~ State Code + Pollutant
3. **Random Forests**

Simple Linear Model:

I will perform **Bootstrap** to find the **confidence intervals** of all the models to be used in

the following part, and interpret the results, the detailed steps will be:

1. Prepare the data for regression by one-hot encoding, pick X to be the state and pollutant to be specific in pollution type and location
2. Fit a model of interest
3. Calculate the R-squared and RMSE values per each iteration
4. Perform bootstrapping to estimate the confidence intervals for R-squared and RMSE
5. Plot the confidence intervals for both R_squared and RMSE



As we can see, the distribution of R_squared is centered around 0.8, with a wide spread from 0.75 to 0.85, whereas the distribution for RMSE is centered at 6.5, with an even spread from 5.5 to 7, this means the R_squared and RMSE are covering a wide range and potentially not as precise as we want since it accounts to a lot of variabilities.

We can see that the linear regression model does a fairly good job of explaining the variability in the AQI values, as suggested by the good R-squared value. The RMSE indicates that the model's predictions are generally accurate, but there is still some unexplained variability that is very high in terms of RMSE. The bootstrapped confidence intervals distribution here suggests that the model's performance is not perfect, but it does provide a reasonable starting point for understanding the relationship between the states + pollutant and AQI values. This means there are potential improvements we can make to the model, one direct response to this question would be to try and use the

shrinkage methods, specifically Ridge and Lasso.

Shrinkage Methods Models:

Alpha Selection:

To test **shrinkage methods** with **Bootstrap**, we need to first find the best alphas for both shrinkage methods, using **cross-validation**.

Luckily there are built-in methods that provide CV-tested optimal values for the alphas of both shrinkage methods.

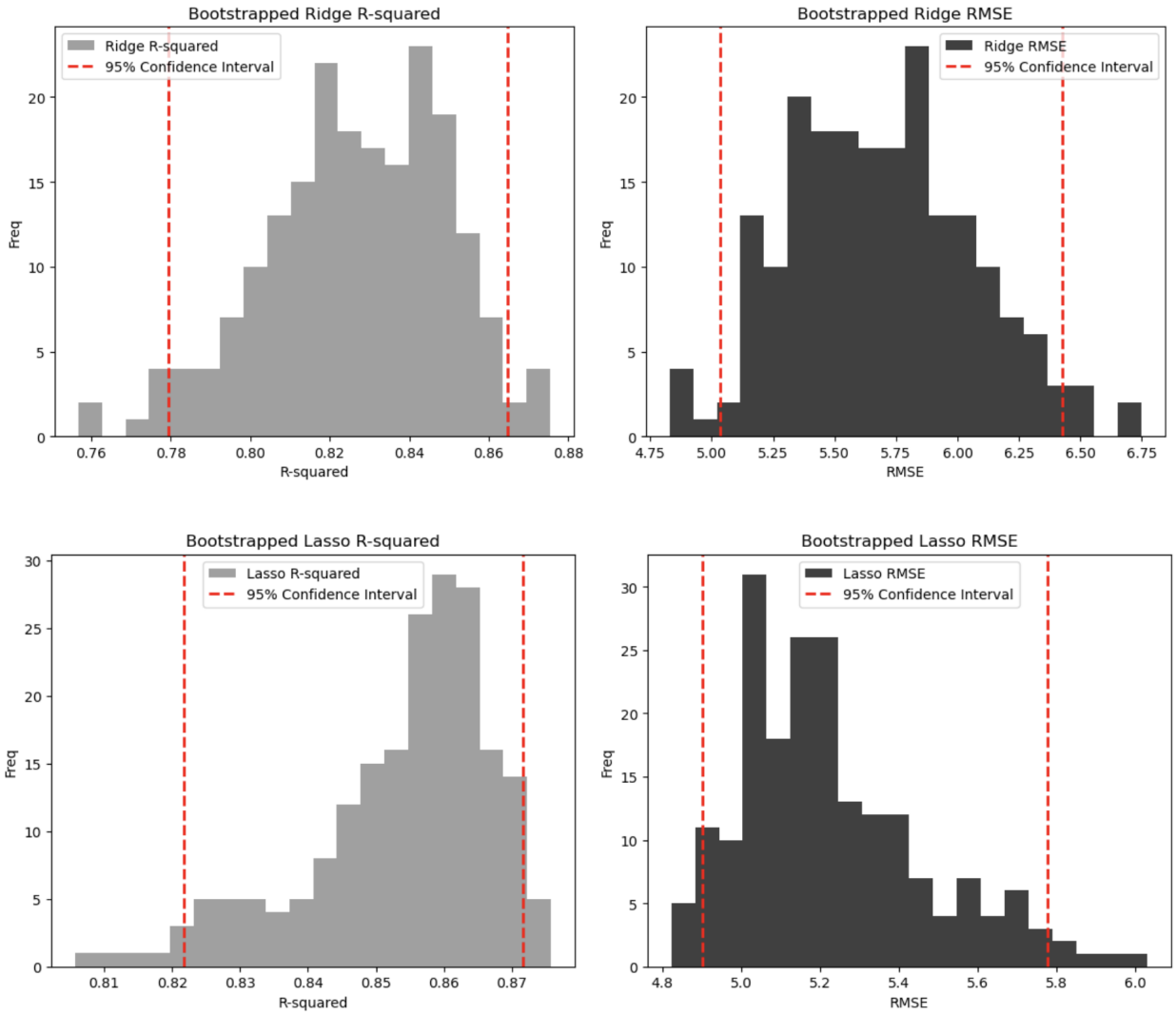
```
Best alpha for Ridge Regression: 1.3219
Best alpha for Lasso Regression: 0.1417
```

Now we know the optimal alpha for Ridge Regression is 1.32 and the optimal for Lasso is 0.14, we can put them into our model. Next, I will use Bootstrap to find the **confidence interval of R_squared and RMSE** of both models and compare them to a simple linear model.

Results:

```
Ridge Regression:
Bootstrapped R-squared 95% CI: [0.77939518 0.86461828]
Bootstrapped RMSE 95% CI: [5.03495714 6.42723314]

Lasso Regression:
Bootstrapped R-squared 95% CI: [0.82170154 0.87168869]
Bootstrapped RMSE 95% CI: [4.901726   5.77817339]
```



What we can get from the above 4 graphs are these:

R_squared of Ridge Regression: centers around 0.81, moderately improved compared to the simple linear model, with a better distribution since it focuses more around the center, but still wide ranged covering from 0.77 to 0.86.

RMSE of Ridge Regression: centers around 5.7, moderately improved compared to the simple linear model, with a similar distribution that is almost evenly distributed, indicating instability.

R_squared of Lasso Regression: centers at 0.86, much better than both previous models, and distribution is more centered at the higher values than lower ones, the range also covers a higher range, from 0.82 to 0.88, this is a significant improvement.

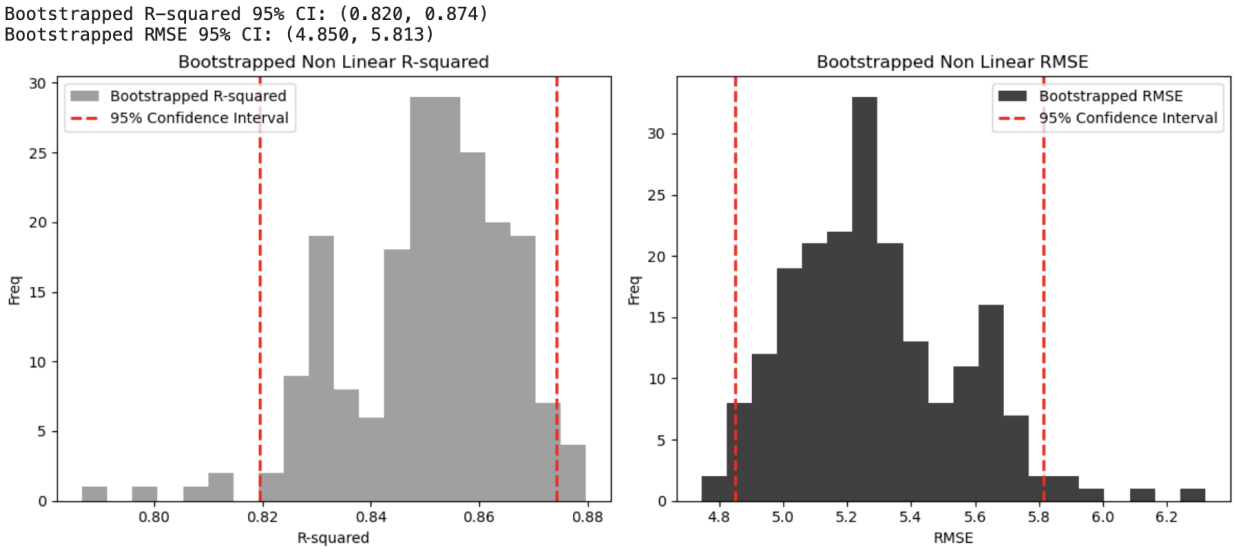
RMSE of Lasso Regression: centers around 5.1, with a central distribution at 5 to 5.3, which means not only the RMSE are lower, but the overall range is lower, and the distribution is more promising.

As we can see, Ridge Regression does worse than Lasso Regression since it ranges greater in terms of RMSE and less in terms of R_squared. As a comparison, they still both outperform the simple linear regression model from before, which means regularization methods work for this dataset and improve our ability to predict and understand the relationship between state and pollutants. The overall distribution for Lasso Regression Model is more reliable since the R_squared is more centered at around 0.865, with the RMSE being relatively stable centering at about 5, which is a huge improvement compared to the distribution of Ridge and Simple Linear Regression.

But can we improve further than this?

Random Forests:

Now I plan to run the same bootstrap method to test the non-linear model performance against all previous models, which in this case is a random forest with the bootstrap procedure.



As we can see, the non-linear R_squared is even better than most of our previous models! This is probably due to how certain states have certain types of factories or labs that emit specific types of pollutants. The RMSE is also lower in this model, and the confidence interval for both R_squared and RMSE are more reliable than a linear model. However,

we can see that the distribution of R_squared and RMSE are not as desirable as the Lasso Regression, since the Random Forrest model's r^2 and RMSE are more scattered but with similar expected values as those from Lasso, so preferably we would choose Lasso for pollution AQI predictions.

Conclusion:

Currently, we have found that in terms of this specific dataset, **the Lasso Regression model and Random Forrest perform the best** with almost the same performance in terms of high R squared and low RMSE. This means now with these two models we can confidently predict the AQI for different pollutants and states in the United States. Both Lasso Regression and Random Forest models provide us with valuable insights into the relationships between the input features and AQI levels. By leveraging these models, we can identify the most significant factors contributing to air pollution levels and inform policymakers to focus on these areas for more effective interventions. **The Lasso Regression model, in particular, is helpful in feature selection as it tends to set the coefficients of less important features to zero.** This can lead to a more interpretable model and help us better understand the underlying mechanisms of air pollution. **On the other hand, the Random Forest model can capture more complex interactions between features and is more resilient to overfitting due to its ensemble nature.**

Overall, both Lasso Regression and Random Forest models offer promising results for predicting AQI levels in the United States. By using these models, we can make data-driven decisions to improve air quality and work towards a healthier environment for everyone.