STAT 444 PROJECT

BY STUART LIAM MIRANDA

Introduction. The impact of air pollution on public health has been studied extensively due to its significant impact on both public health and the environment. Pollutants such as nitrous oxides and ozone often serve as a major component of smog and pose serious health risks. In particular, ozone near the surface aggressively attacks human lung tissue and can cause long-term respiratory issues. Ozone may also react with other pollutants to create stronger reactions in the lungs. As a result, understanding the factors that may impact ozone levels is a matter of interest.

Data. The ozone dataset from The Elements of Statistical Learning textbook includes various environmental factors that are potentially linked to ozone levels. The dataset includes daily ozone concentration (ppb), wind speed (mph), daily maximum temperature (degrees Fahrenheit), and solar radiation (langleys) from 1973 from May to September. The covariates were measured in various locations from New York City, and collected by the New York State Department of Conservation and the National Weather Service.

Preliminary Analysis. A preliminary analysis of the covariates of our dataset versus the response variate was performed. We first fitted any linear models between each individual covariate versus ozone, resulting in 3 different linear models. From the plots of our models, there doesn't seem to be any strong linear relationships between the ozone and the individual covariates. When fitting a multiple linear regression model with all the covariates versus ozone, residual diagnostics showed that our model does not follow a normal distribution.

We also wanted to check whether multicollinearity exists by doing a VIF correlation matrix on the covariates in Figure 1 below. The values shared between every pair are low, implying that there is no significant problem with multicollinearity. Therefore, there is no need to remove any covariates from the model. We would expect that there would be a more complex relationship among our covariates and our outcome that would provide may be represented by complex, smooth curves in a flexible manner.

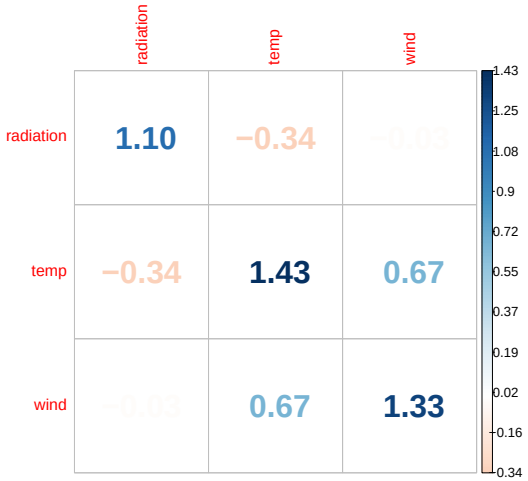

Figure 1: Multicollinearity Results not being significant

Research Question. The research question we aim to answer is **to what extent do solar radiation, wind, and temperature serve as reliable indicators of ozone levels in New York during the summer months of 1973?** This research question allows us to study the relationship between the covariates and ozone concentration near the surface, which may lead to a model that can be used to predict future ozone levels. Such a model would allow public health officials to put out announcements asking high-risk individuals to avoid going outdoors to avoid health risks. Given today's changing climate and persistent pollution in larger metropolitan areas, attempting to understand whether and to what extent certain environmental factors impact ozone levels is pertinent to public health.

Methods. Since there is no clear linear relationship between the covariates and the response variable, we will develop several models that can capture non-linear and complex relationships. We will be fitting our ozone data with 7 models: linear, ridge, LASSO, spline, additive, tree, and random forests. To determine the best model, we will examine the mean squared errors (MSE), generalized cross-validation scores (GCV), and adjusted R^2 values.

Linear Regression: We first start with a multiple linear regression model which is the simplest predictive modeling technique:

$$\hat{y} = \hat{f}(x) = X\hat{\beta} = \beta_1 x_{\text{radiation}} + \beta_2 x_{\text{temperature}} + \beta_3 x_{\text{wind}}$$

where we use the method of least squares to estimate the best β parameters:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} ||y - X\beta||^2$$

Ridge Regression: This is a more advanced method of predictive modeling where we apply linear regression to a set of transformed inputs.

$$\hat{y} = X\hat{\beta} = X(X^T X + \lambda I)^{-1} X^T y$$

Now we introduce the penalty term $\lambda > 0$ to prevent the size of the coefficients from being too large. This is done to limit the influence of linearly dependent variables on the predictions. So our coefficients used in our model are acquired as follows:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} ||y - X\beta||^2 + \lambda ||\beta||^2$$

It is a penalized version of our least squares objective function, where if we set $\lambda = 0$, then we would up with our linear regression model.

LASSO: Similar to the ridge regression, the LASSO is a method that simultaneously performs shrinkage with its additional shrinkage term as indicated below:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} ||y - X\beta||_2^2 + \lambda ||\beta||_1, \text{ where } ||\beta||_1 = \sum_{j=1}^{\beta} |\beta_j|$$

However, this time it also performs variable selection, resulting in more parsimonious models. This is because, for ridge regression, we cannot remove an input (make β coefficient only close but not equal to 0) from the model. This is indicated by the change of our 2-norm to the 1-norm for our penalty term, allowing our model to set coefficients to zero exactly.

Regression Spline: We use spline interpolants as a numeric analytic tool that smoothly interpolates a set of points using piecewise polynomials with knots at each point. We define $b_{j,p}$ as these piecewise polynomials that are continuously differentiable at a set of knots which we call our B-spline basis functions. Therefore, we can represent our model as:

$$E[Y|X = x] = f(x) = \sum_{j=1}^{k} b_{j,p}\beta_j$$

which describes a linear combination of the B-spline basis functions. This introduces non-linear relationships between variables without assuming a specific form for the relationship.

We can estimate our β parameters from:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int \{f''(x)\}^2 \, dx$$

where $X_{ij} = b_{j,p}$ is the "spline design matrix". So we fit this nonlinear regression model by choosing some knots, modeling the regression function as a spline function with those knots, and then estimating our spline weights by minimizing the penalized least squares above.

The second term is a measure of roughness that makes sure that we have a "calm" interpolant, where between the knots, they vary less than the polynomial. This form looks similar to the ridge regression model so we can use the same technique when estimating coefficients.

Additive Model: As a variation of the regression spline model, the regression function depends additively on multiple covariates, each modeled using separate spline functions:

$$E[y|X_1 = x_{\text{radiation}}, X_2 = x_{\text{temperature}}, X_3 = x_{\text{wind}}] = \alpha + f_1(x_{\text{radiation}}) + f_2(x_{\text{temperature}}) + f_3(x_{\text{wind}})$$

where we represent $f_j(x) = \sum_{l=1}^{d_j} b_{j,l}(x)\beta_{j,l}$ which are individual cubic B-splines subject to the linear constraint $\sum_{i=1}^{n} f_j(x_{ij}) = 0$ for $j = 1, 2, 3$

These spline functions ensure the estimated relationships are smooth, avoiding overfitting while capturing important patterns. Estimating the smoothing parameters can be modeled similarly to the regression spline method (fitting model by ridge regression):

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \underset{\alpha,\beta}{\operatorname{argmin}} ||y - \alpha 1_n - X\beta||_2^2 + \sum_{j=1}^{p} \lambda_j \int \{f_j''(x)\}^2 \, dx$$

Because the contribution of each feature to the response is interpreted separately, we can have an intuitive understanding of how each predictor affects the response, as shown below:

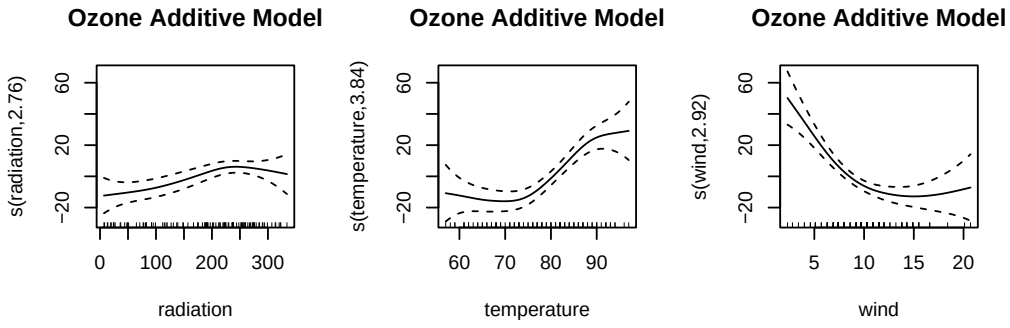

Figure 2: Ozone Additive Plots

In Figure 2, each plot indicates the relationship between each covariate and ozone. There are different nonlinear complex relationships between each pair: radiation has a flattened square-root relationship with ozone, temperature has a somewhat sigmoidal relationship with ozone, and wind appears to have a negative log relationship with ozone.

Looking at the rug plots, the evenly distributed ticks on the x-axis for temperature and wind indicate evenly distributed data, whereas there is a cluster of data points for radiation at roughly 250 langleys. This is the section with the tightest confidence interval. All the confidence intervals are narrow, indicating less uncertainty in the estimated effect. Note that there aren't any isolated tick marks, which implies that there aren't any outliers that may influence the shape of our curve.

It is also necessary to identify regions where each predictor may have a significant effect on ozone, indicated by steep changes in certain ranges of the x-axis. This is most evident for temperature at around 75-85 degrees Fahrenheit and wind at 8-12 mph, indicating that these are the ranges where they have the most effect among the predictors of ozone levels.

The magnitude of the effect of each predictor is also measured by how big the deviation from 0 is measured for each curve. From the plots, wind has the largest effect for smaller wind values, whereas temperature has the largest effect at higher temperatures. Radiation appears to have the smallest effect on ozone.

Trees: Regression trees are a form of decision trees used for prediction. They split the data into subsets based on the values of our covariates, aiming to partition the data into different subsections. A split on node t based on variable x_j at value s can be expressed as:

$$x_j < s \text{ or } x_j \geq s$$

Each split aims to minimize the residual sum of squares on the resulting two child nodes. The way we measure the residual sum of squares for a particular node t is:

$$\text{SSR}(t) = \sum_{i \in t} (y_i - \bar{y}_t)^2$$

This process is recursively repeated for each child node until a stopping criterion is met. In the tree model used with the ozone data, the stopping criterion is that the minimum number of observations in each node is 20 and the maximum number of terminal nodes is 4.

The predicted value for an observation falling into a leaf node t is the mean of the target variable for observations in that node:

$$\hat{y}_t = \frac{1}{|t|} \sum_{i \in t} y_i$$

Random Forests: This is an ensemble learning method that combines multiple trees to improve predictive accuracy. Each tree is trained on a bootstrap sample of data to ensure that each tree is different. At each split of the tree, a random subset of predictors is considered to find the best split. This further de-correlates the trees, enhancing the ensemble's performance.

For each tree, given inputs X, we denote the prediction of the bth tree to be $\hat{y}_b(x)$. The random forest effectively takes the average prediction of each of all the trees:

$$\hat{y}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_b(x)$$

where B is the total number of trees in the random forest. In our random forest model, we decided to vary the number of trees B to 5, 50, and 500 as a tuning parameter. Increasing the number of trees may potentially increase prediction accuracy but at the expense of increased computational costs and memory requirements, as it needs to store more trees.

One useful feature of the random forest is that we can determine feature importance, which indicates which covariates are most influential in making our predictions on ozone levels. This helps us understand the overall contribution of wind, temperature, and radiation separately.

The way this is measured is by the mean decrease impurity (MDI). Purity means that each subset after a split of a tree contains observations that are similar to each other. The aim for every split of a tree is to reduce impurity to make more precise predictions, since each node would have very similar values and variance is minimized. It is considered important if a feature is used to make a split that significantly reduces impurity. We average all the impurity reductions across all the trees in the random forest.

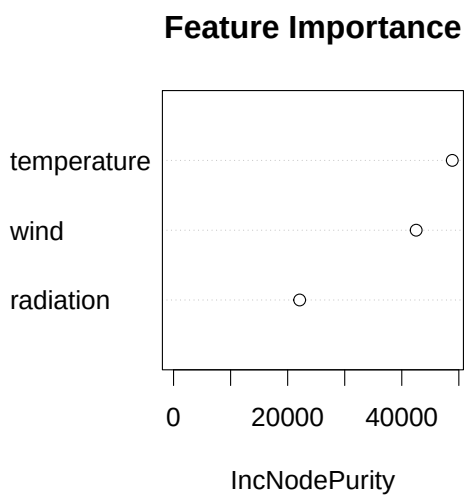

Figure 3: Feature Importance Plot

In Figure 3 above, it appears that temperature is the most important covariate, followed closely by wind, whereas radiation does not have much impact on ozone levels. This finding is synonymous with our conclusion from the additive model.

Results. The results of fitting all 7 models are displayed in Figure 4 on the next page.

The metrics for measuring how accurate our models fit the ozone data are mainly by mean squared error and generalized cross-validation (GCV). Adjusted R^2 determines the proportion of variance from the outcome that our covariates can explain.

GCV: This is calculated as part of the modeling fitting process for generalized additive models. A lower GCV score implies that the model balances fit and smoothness more effectively, so it is preferred.

$$\text{GCV}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - \text{trace}(H)/n}\right)^2$$

where H is the hat matrix corresponding to $H = X(X^TX)^{-1}X^T$.

From Figure 4, the two most successful models were the random forests and the additive spline models, wherein both models had the two lowest prediction errors and GCV error and

highest adjusted R^2 value, which are all preferable for determining the best model fit. Note that for random forests, increasing the number of trees from $n = 5$ to $n = 50$ reduced both errors greatly.

Model Comparisons

Model	Prediction_Error	GCV	Adjusted_R_Squared
Linear	432.10917	465.0203	0.5951713
Ridge Regression	433.80097	458.2357	0.4995825
Lasso Regression	432.52028	456.8829	0.4834381
Spline	272.67331	349.8134	0.7210810
Additive	276.95865	**292.5589**	0.7236839
Tree	333.78479	523.9588	0.6985570
Random Forest (n=5)	123.58988	**337.4335**	0.8883853
Random Forest (n=50)	91.52072	288.2722	0.9173471
Random Forest (n=500)	82.77643	290.3238	0.9252441

Figure 4: Table of Models with Corresponding Metrics

Random forest does well with large datasets containing many features. However, the random forest may overfit the training data because the ozone dataset only has 3 features with 111 observations. There may also be insufficient data to fit multiple trees, as each tree may not have distinct meaningful patterns. This is evident for $n = 500$ having slightly smaller errors than $n = 50$, even if we increased the number of trees drastically. However, our feature importance plot in Figure 3 did aid in the interpretability of the random forest in showing that temperature and wind were much more significant indicators of ozone levels than radiation.

On the other hand, using our additive model aids in answering our research question through the plots in Figure 2. We can determine the extent how which our covariates: wind, temperature, and radiation, may have significant effects on ozone levels depending on their measurements. In particular, wind and temperature seem to have a significant impact on ozone levels at specific ranges of miles per hour and degrees Fahrenheit respectively, whereas radiation doesn't seem to have a big effect on ozone levels.

Checking residual diagnostics: independent and identical distribution of errors, no multicollinearity, and our basis functions being cubic splines, all satisfy the assumptions for the additive model. Therefore, our additive model is most suited not only to fit our data but also to answer our research question on the extent of the effect of each covariate on ozone levels.

Conclusion. From the additive model plots (Figure 2) and feature importance plots (Figure 3), wind and temperature serve as significant indicators of ozone levels, whereas radiation doesn't have a significant impact on ozone. A limitation is the size of our ozone dataset is too small with too few features. A potential future work would be to include interaction terms within our models and gather more data to have our random forest model be used to its fullest extent, coupled with partial dependence plots and tree surrogates to improve the interpretation of random forests into how each a covariate influences predictions on ozone levels.