EXPLORATION OF OZONE CONCENTRATION

SIDNEY CHRISTENSEN & CARINNA ROEBUCK

PROJECT 2 STA4614

NOVEMBER 2023

ABSTRACT

Ozone pollution is a byproduct of coal and fossil fuel burning, occurring when internal combustion engines and fossil fuel power plants release volatile organic compounds and nitrogen oxide that interact with oxygen and UV rays (US EPA, 2016). In this study, the goal is to use eight different predictors from the environment of which they were sampled to predict ozone concentration. Exploring possible indicators of this pollution is pivotal as humans can experience respiratory complications from prolonged exposure. The aim of this study is to examine eight of the most suspected ozone concentration indicators/predictors and evaluate the significance of each. The final model will provide the subset of predictors that best explain the ozone concentration in any given environment. The concentrations in the data are sampled from Upland, California with each sample representing one day of the 330 days sampled throughout the year of 1976.

OBJECTIVES

Seeing that there are eight predictors in the model, there are many questions to be answered with the most important being which combination of predictors would most significantly explain the target (the daily maximum of the hourly-average ozone concentration.) During this process, the possibility of collinearity between the eight predictors is explored. After the study is concluded, the most significant factors in predicting atmospheric ozone concentration will be identified by the final model. The most prominent objective here is finding which atmospheric conditions most accurately predict ozone concentration so that the remedy for ozone pollution is more efficiently executed.

SOURCE OF DATA

The data used in this study is sourced from Leo Breiman, a consultant on the project from which the data is being used. The sampling was conducted in Upland, California circa 1976. It should be noted that the response variable (ozone) is the log of the daily maximum of the hourly-average ozone concentrations measured in PPM (parts per million). Additionally, there are 330 samples and not 365 because some samples (days) had missing values, and such samples were not included in the data provided. The table below provides the names of the eight predictors and their corresponding descriptions.

Predictor	Description
“vh”	Vandenberg 500 mb height: the height of the 500 millibar (mb) pressure level above the mean sea level at Vandenberg Air Force Base in California. It’s used for analyzing the upper-air patterns and tracking the movement of weather systems.
“wind”	Wind speed in meters per second
“humidity”	Humidity percentage
“temp”	Sandburg AFB temperature: The recorded temperature in fahrenheit at the Sandburg Air Force Base in Los Angeles
“ibh”	Inversion layer base height: An inversion layer forms when the air near the ground is colder than the air above it, trapping pollution in this inversion layer.
“dpg”	Daggot Pressure Gradient: describes in which direction and at what rate the pressure increases most rapidly around a certain location
“ibt”	Inversion base temperature: the temperature at the altitude that the inversion layer begins (the base of the inversion layer)
“vis”	Visibility measured in miles (the distance at which someone can observe objects in the atmosphere)

ENVIRONMENT OF STUDY

The statistical analysis for this study will be conducted in the R language using Google Colab. The version of R used is 4.3.1

DATA EXPLORATION

FREQUENCY EXPLORATION

In our frequency exploration, we ran primary analysis on the predictor variables, with Upland Maximum Ozone (ozone) in mind as our target variable of this study. The descriptive statistics for ozone reveal a mean concentration of 11.78 ppb, underscoring the central tendency of our dataset. Importantly, the median of 10.00 ppb provides additional context, indicating that the distribution may have a slightly positive skew. This observation prompts a closer inspection of potential outliers and the underlying distribution of ozone concentrations. Furthermore, the range of ozone values, spanning from 1.00 to 38.00 ppb, signifies substantial variability in atmospheric ozone levels over the studied period.

Turning our attention to predictor variables, the Vandenberg 500 mb Height, with a mean of 5750 meters and a range from 5320 to 5950 meters, highlights the altitude variations that may impact atmospheric conditions. Similarly, the Wind Speed variable, with a mean of 4.891 mph and a range of 0.000 to 21.000 mph, showcases the diversity in wind dynamics that could influence ozone dispersion. These statistics not only provide essential insights into the central tendencies and variabilities of key variables but also lay the groundwork for in-depth analyses, guiding our understanding of the complex interplay of factors influencing Upland Maximum Ozone.

GENERAL GRAPHS

In Figure 1, we observe four of our variables vh, wind, humidity, and temp as histograms and scatter plots. The dataset offers a summary of Vandenberg 500 mb Height (vh), Wind Speed (wind), Humidity (%), and Sandburg AFB Temperature (temp). The histogram for Vandenberg 500 mb Height (vh) shows a right skew in data, with a mean around 5,760 meters. Wind Speed (wind) reveals a right-skewed distribution, suggesting a prevalence of lower wind speeds over higher ones. Humidity follows an approximately normal distribution, indicating a balanced spread of humidity levels. Sandburg AFB Temperature in Fahrenheit exhibits a histogram resembling a normal distribution, implying a central tendency around the mean temperature. Additionally it can be noted that all of the scatterplots, with the exception of wind, are showing positive, strong correlations to the target variable ozone. This can pose some concern for multicollinearity issues in our dataset further on.

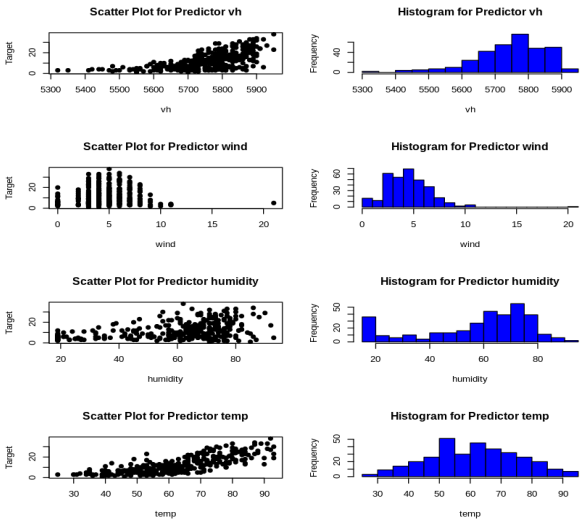

Fig 1. Histograms and scatter plots for variables Vandenberg 500 mb Height (vh), Wind Speed (wind), Humidity (%), and Sandburg AFB Temperature (temp).

Figure 2 shows our remaining variables inversion base height (ibh), daggot pressure gradient (dpg), inversion base temperature (ibt), and visibility (vis). Ibh's histogram reveals a right-skewed distribution, indicating a concentration of

lower inversion base heights. It should be noted that there is a large peak around the 5000 km area. Dpg, representing daggot pressure gradient, might exhibit a symmetric distribution around its mean. Ibt, denoting inversion base temperature, shows a histogram resembling a normal distribution, suggesting a central tendency in the data. Vis, representing visibility in miles, displays a right-skewed histogram, with lower visibility values being more prevalent. The scatter plot for ibt and ibh show strong correlations with the target variable, however dpg shows a weaker positive correlation, and vis shows a hard to read scatter plot.

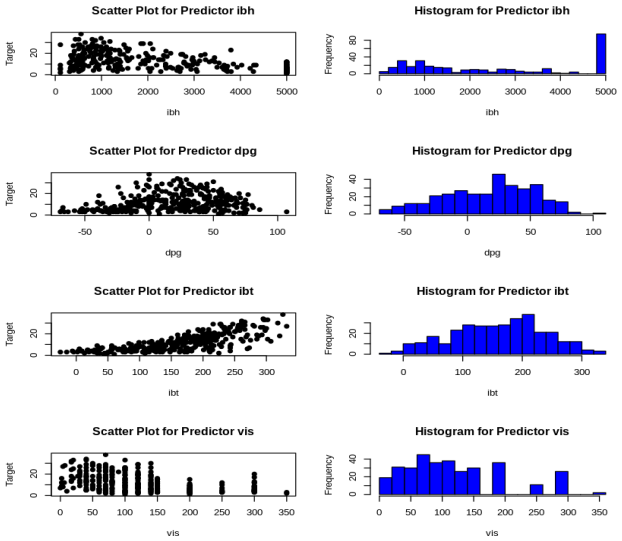

Fig 2. Histograms and scatter plots for variables Inversion Base Height (ibh), Daggot Pressure Gradient (dpg), Inversion Base Temperature (ibt), and Visibility (vis),

LINEARITY ASSUMPTIONS

Part of assessing the reliability of our dataset is to test for the linear regression assumptions. During our analysis we tested for the five linear regression assumptions. For any given fixed value of the variable X, the existence assumption explains that the presence of a random variable Y, characterized by a probability distribution with a finite mean ($\mu Y|X$) and variance ($\sigma Y|X$^2). This foundational assumption is inherently satisfied by all regression models, irrespective of their linearity. Furthermore, the independence condition is upheld in our study, as all observations were made only once during the 330-day study period, with a single recorded value for each predictor, ensuring the independence of Y

values. The homoscedasticity assumption is notably violated, as evident in our residual plot as the variance increases.

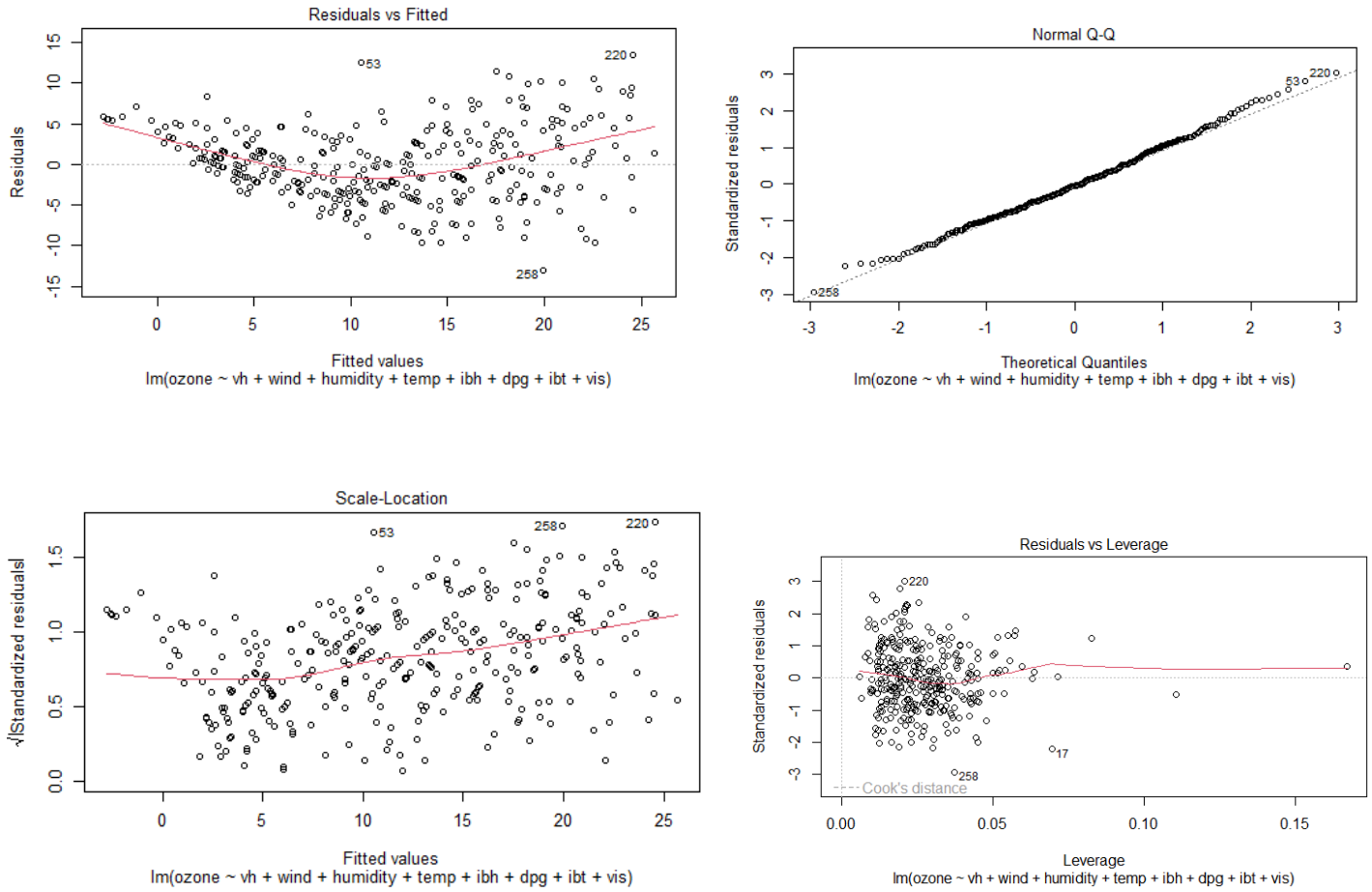

Fig 3. Output from Multiple Linear Regression of Residuals Fitted, QQ-Plot, Scale-Location, and Residuals vs Leverage

To mitigate this issue, we plan to implement variable removal models and methods. Acknowledging the current limitations in correcting homoscedasticity, it is important to note that the target variable has already undergone a log transformation. The assessment of normality, facilitated through QQ, shows that we more or less satisfy this function, with some concern being in the tail end of the data.

COLLINEARITY EVALATUION

The correlation matrix of the eight predictors exhibited significant r-square values. The VIF values calculated showed “ibt” and “temp” to have values around or greater than 10, with “ibt” having a VIF of 19.03 and temp with a VIF of 9.2. The context of our data; However, prompted us to refrain from removing “temp” since temperature is significantly related to ozone concentration and thus would make the model unreliable if removed. Once “ibt” was removed, heat maps of the correlation matrix helped show the decrease in collinearity exhibited across the two predictors that “ibt” was correlated with. Once “ibt” was removed, the heat maps show much less significant R-square values, confirming that all significant collinearity has been appropriately dealt with. It should be noted that “ibt” will only be removed

from the model building process when conducting stepwise selection and backward selection. However, the Ridge and LASSO models will have access to “ibt” as the aim is to see how these two methods rectify collinearity. Both methods are highly equipped to address multicollinearity in different ways, which will be explained further in the “Modeling” section of this report.

DATA PREPARATION

Data preparation for this study did not take significant effort as the data came already cleaned of missing values. This is because the researchers who collected the data left out any of the 365 days (observations) that did not have all eight predictors recorded. Only 30 days were missing, so the researchers went ahead with complete case analysis. To improve our initial models’ r-squared value, we identified influential points through the Cook’s distance method. We chose to remove these influential points as they also coincided with outliers observed. In terms of transformations, the response variable “ozone” had already been log transformed by the researchers. Unfortunately as mentioned in the “Data Exploration” section, this log transformation did not help rectify the linearity or homoscedasticity violations. My partner and I concluded that for the scope of this study it did not make sense to undo the researchers’ transformation and try another. Let us note that because of these violations, the model is less reliable without the transformation correcting the two violated assumptions. Lastly, dummy variables are not appropriate for this study as the predictors and response are continuous.

MODELING

Let us briefly note that upon univariate model exploration, the individual residual plots and scatter plots both exhibit violations of our regression assumptions. Additionally, most of the predictors follow a skewed (and for some bimodal) distribution which is confirmed through the individual QQ plots generated for each simple linear regression model. During general model building and exploration, interaction terms were not considered as multicollinearity was present beforehand, and predictor removal was used to rectify this. Our initial model is written as Upland Maximum Ozone = 27.985 - 0.007{vh}+ 0.067{wind} + 0.070{humidity} + 0.274{temp} - 0.0005{ibh} + 0.001{dpg}+ 0.025{ibt} - 0.007{vis}+ε. Our multiple r-squared from this initial model is 0.6912, explaining roughly 69% of the variance in our target variable ozone. This model also features a mean squared error of 16.01 and an extremely low p value of 2.2e-16.

In addition to improving our chosen selection criterion, we plan to remove variables from our initial model to improve our linear regression assumptions, specifically linearity and our heteroscedasticity issue. Now we move onto the result of our comprehensive model exploration. Four methods were used to explore the most appropriate model, which are backward selection, stepwise regression, LASSO regression and Ridge regression. LASSO and Ridge, although beyond the scope of this class, were included to explore their individual approaches to our collinearly-natured data. That being said, "ibt" was added back into the set of predictors only for these two specific methods. LASSO remedies multicollinearity through variable selection which also optimizes interpretability. Alternatively, Ridge regression shrinks all coefficients down towards zero accordingly with their collinearity and does not remove any variables. Thus, ridge is not an optimal model in this study since the model will not be interpretable. If it had a significantly higher R-square than the other models, we wanted to explore that possibility. On the other hand, "ibt" was excluded from the execution of backward selection and stepwise regression as these two methods are not built to address multicollinearity. Results show all four models having an R-square value of about .72, so ridge regression was not considered further as its R-square value does not stand out, thus giving the model poor interpretability. All four MSE values were all approximately 16. Backward selection had an MSE of 16.603, stepwise selection had an MSE of 16.232, LASSO had an MSE of 16.222 and Ridge had an MSE of 16.32. Thus, with all four models having approximately the same R-squared and MSE values, it came down to the degree of interpretability according to the number of predictors each model contained. Backwards selection gave a model containing three predictors while stepwise had four and LASSO had five. Thus, using a collective assessment of these three attributes, Backward selection was chosen to be the best model. The final model is as follows: Ozone Concentration = -11.089 + 0.79{humidity} + 0.335{temp} - 0.001{ibh} + error. This model features several criteria that are relatively the same as our initial model, with our R-square only

increasing by about 0.03 and our MSE staying the same.

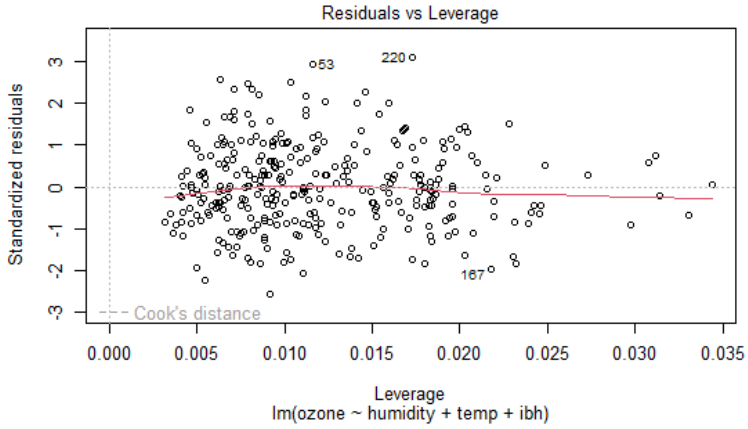

Fig 4. Residuals vs Leverage for Final Model

Removing variables however did help to address multicollinearity issues as well as improving our linearity and heteroscedasticity as shown in Fig 4.

DISCUSSION & CONCLUSION

Our final model (backward selection) tells us “humidity”, “temp”, and “ibh” are significant in predicting the ozone concentration on a given day in Los Angeles, California. We found in this study that temperature (temp), inversion layer base height (ibh) and humidity (humidity) are all necessary for explaining the variance in ozone concentration throughout Los Angeles California. Inversion layers form when warm air acts as a lid to the colder air beneath it. This causes an inversion layer, containing cold air and trapped pollution. Thus, a higher inversion layer base is correlated with a higher ozone concentration trapped in it. Temperature and humidity are controlled by nature, but ozone pollution trapped in the inversion layer can be remedied. Vehicle emissions and industrial processes, along with burning fossil fuels, are the ways humans contribute to ozone pollutants and this can be controlled with educating one another and creating environmental initiatives (US EPA, 2016). Ozone pollution is in our hands, but action is desperately needed. Training and testing sets were not used in this study as our knowledge of this process in the R language is not sufficient. Thus, model performance cannot be provided at this time. However, testing models is extremely important in knowing whether or not your model is able to generalize to other data. For further research, this method is strongly encouraged.

REFERENCES

Elements of Statistical Learning: data mining, inference, and prediction. 2nd Edition. (n.d.). Hastie.su.domains. https://hastie.su.domains/ElemStatLearn/

US EPA. (2016, March 21). *What is Ozone?* US EPA. https://www.epa.gov/ozone-pollution-and-your-patients-health/what-ozone

APPENDIX