**Question & Assessment**

Ozone is quite harmful to human health when found on the surface of the planet. For this reason, it is important to be able to predict the ozone level for a community. By predicting the maximum 1-hour average ozone level (ppm) forecasters will be able to put out media announcements to the community that those with certain health conditions should stay indoors. Ozone could be rated based on the measurement as to how much it may impact people with health concerns. <https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics> In addition, communities that are working to reduce ozone may want to know when to encourage a ride-sharing program to help stop ozone from being generated, for example, because vehicles contribute to the pollution. Predicting the level is going to aid community leaders in guiding such programming. For this reason, the target audience of ozone level predictions is individuals who have health problems that are exacerbated by ozone and community leaders who are working to reduce levels of ozone.

I selected the quantitative predictor because having more precise knowledge of the quantity of ozone in the atmosphere will be more useful than a high/low value. That is, I thought the precision would be important to the community and could have numerous applications such as categorizing quantities into different buckets or putting the specific predicted parts per million in the media.

**Analysis**

The two methods utilized in the analysis are **Method 1**: Linear Regression with Subsets and **Method 2**: Penalized Regression (LASSO and Ridge Regression).

**Method 1** is appropriate because there are linear relationships between the predictors and response variable, and it finds the best model for the various number of subset predictors over some other regression methods. This can improve interpretability and tradeoff between bias and variance. More specifically, this method was chosen because it allows us to formulate the best set of predictors. The formula can then be easily interpreted and utilized by a user who knows the values for the variables used to make a prediction.

**Method 2** is appropriate because it can provide high prediction accuracy and stable results. Because we did see a few outliers in the data, the penalized regression method will help to limit the influence these have on the predicted values. This penalty helps us get more accurate predictions when we are applying the model to new data. This method is also appropriate because we have many predictors and it tends to do well with many. Another advantage to why Method 2 is preferred for this data over other methods is that it allows (through Ridge Regression) us to determine predictors to exclude, which may be helpful if there are unimportant variables being considered in the model.

Within the ozone data, there are a variety of variables that can be used to predict the response variable. Examples include the 500 millibar pressure height (meters), Humidity (%), inversion base height, and wind speed (mph), to name a few. Many of the variables are numerical observations related to weather, such as temperature in Sandberg, CA. For this reason, there are some correlations between the values. For example, there is a correlation 80% of the time between the temperature in Sandberg and the 500 millibar pressure height. There is also strong correlation between the humidity and the pressure gradient from LAX to Daggett, CA (69%). This would make sense that some weather variables are associated with each other.

In the analysis, we considered all other variables to be predictor variables. We excluded the inversion base temperature variable because its variance is already explained within the scope of other variables. The weekday variable was excluded because it was the only categorical variable within the dataset. From a practical standpoint, weekday did not seem very helpful to a user of the final model and is not conducive to penalized regression. However, the other variables are quite conducive to use in penalized regression and they were kept.

It is important to note we did log transform some variables to account for the uneven spread of values, as a standard statistical method. This included transforming the predicted value of the variable HourAverageMax, which is indicated by the prefix of “log” whenever we transformed variables.

Overall, our modeling considered the two methods described above to predict the value of the ozone level in parts per million using up to seven different predictor variables. By performing a modeling process called “Cross-Validation” we determined the best model by incorporating as much data as possible to create the model. We then tested the model against the remaining sample data to see how well it performed.

Finally, after determining the best model to use, we ran a second cross-validation process to assess how well our entire methodology holds up. That is, we assessed the overall ability of our modeling to predict new data in the future.

**Results & Selection**

The analysis, using all available data, revealed the best model to be from Method 1 which was a linear regression model. The model used the variables: LogHourAverageMax, tempSandburg, visibility, loghumidity, and loginversionBaseHeight. (one response and four predictors) The final fitted model formula is:

LogHourAverageMax = 0.2041 + 0.0318\*tempSandburg - 0.0007\*visibility + 0.3072\*loghumidity – 0.1449\*loginversionBaseHeight

To apply this model to new data, simply input the known predictor values (applying logarithms as indicated). The result will be the **logarithmic** **value** of HourAverageMax. To get this in the form of PPM, simply take 10 to the power of this number (ex: 10^**LogHourAverageMax** = HourAverageMax (PPM)). The reason we utilized logarithmic values was to ensure we had uniformity as it helps the model.

The best model formula indicates that as the temperature in Sandburg and humidity increases, so does the average maximum parts per million (ppm). As visibility and inversion base height increase, the average ppm decreases. By knowing these four predictors, you can find the predicted value of the 1-hour average maximum ozone in ppm.

This best model, selected by our process, was selected by a measure relating to how well the predicted values compared with the actual values of the known sample data. That measure, for our model, was 0.1806 and was the lowest among a total of 207 various models that were assessed across our methods. (See chart on next page for a comparison of model performance.)

Additionally, we assessed the modeling process by performing a second “wrap-around” layer of cross-validation to predict all the values in what is called an “honest” way. This process shows us that our modeling process performs fairly well when predicting truly new data as it gives us a measurement that shows our modeling *process* explains about 68% of the variance in our ozone average max variable with the predictors in an honest way.

The figure below shows the three model groups, Regression Subsets, Ridge Regression, and LASSO; the model represented by the red dot scored the best (lowest) amongst all the models used in the cross-validation.

