The model that we have built is a generalised linear model with the intended purpose of being able to predict a dependent variable based on a number of independent variables. More specifically, it is designed to allow the user to input which independent variables are of interest. These variables contain air quality related attributes, and the user is also able to select which attribute is the dependent one to be predicted. We have applied this model onto our given data set, which contains data on air quality, collected from March 2004 to April 2005.

The data was retrieved from a .txt file and read into FICO Xpress, which was then appropriately cleaned by removing all points labelled *-200*, which indicates a missing value. The attributes for date and time were not included for this model. Thereafter, the first 500 valid data points were imported and prepared to feed into the model by splitting the dependent and independent variables.

Dependent variable: Temperature (T)

10 Independent variables: CO(GT), PT08.S1(CO), NMHC(GT), C6H6(GT), PT08.S2(NMHC), NOx(GT), PT08.S3(NOx), NO2(GT), PT08.S4(NO2), PT08.S5(O3)

We firstly specified our decision variables and objective functions for each approach. We have three decision variables as illustrated below:

|  |  |
| --- | --- |
| **Decision Variables** | |
|  | the regression coefficient for gradient of independent variables |
|  | the regression coefficient for the Y intercept |
|  | the error or distance between the actual value (y) and the predicted value |

We introduced a new parameter, tolerance μ, which indicates whether the estimation of the model is sufficiently good and thus, it creates a range in which if the predicted value is between this upper and lower limit then the error is zero. Subsequently, two approaches are utilized. The first model minimizes the sum of errors, and the second model minimizes the maximum error.

The mathematical formulations of the two models are illustrated below:

**Model 1**

Objective: Min

Constraints:

(upper bound)

(lower bound)

**Model 2**

Objective: Min

Constraints:

(upper bound)

(lower bound)

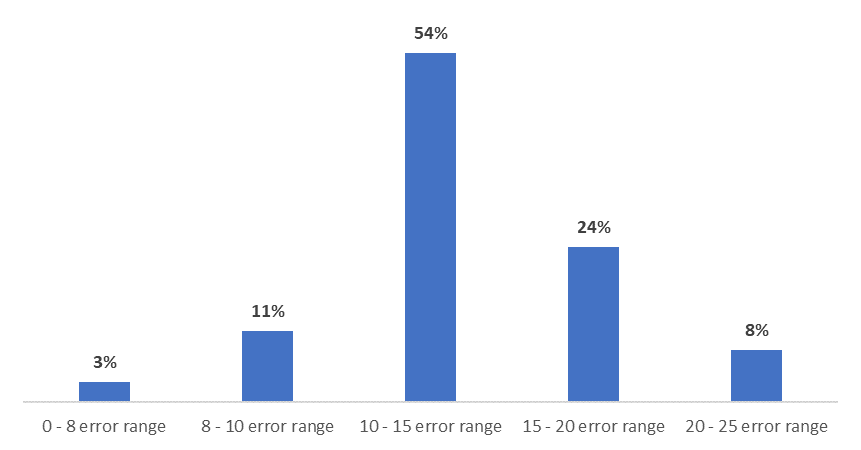
The constraints are identical for both approaches since the rationale for the constraints is the same and the only difference is the objective function. For building the constraints, we consider that if the actual data point (y) is under the predicted line then the error should be the distance between the predicted value and the upper tolerance point. On the other hand, if the actual value (y) is above the predicted line then the error should be the distance between the predicted value and the lower tolerance value.

In terms of results, it is worth noting that the maximum deviation between the predicted and the actual values of the first model is significantly lower (around 10.71) compared to the second model which is 90.98. In addition, a number of predictions of the first model (around 20%) fall within the tolerance limit thus the error is zero, while less instances from the second model fall within the tolerance limit around 9%. Therefore, the first model of minimizing the sum of errors seems to be more appropriate based on the training set. However, further testing of the models based on unseen data can assist in comparing the performance of the two approaches.

As a next step, we read additional 100 valid data points from the input dataset to further evaluate the performance of our first regression model minimizing the sum of errors. We choose the first regression model because the maximum error is smaller than the second model. We utilise the solutions of the model to predict the temperature variable and then we estimate the absolute error value including the same tolerance of 0.05.

Examining the error values derived from the test set, we can observe that there are no zero errors indicating that the predictions are outside the range of tolerance. In addition, the following graph shows that the average deviation between the predicted and actual values is around 13 units while the maximum deviation is almost 22. The majority of errors (78%) fluctuates between 10 and 20 units.

***Distribution of absolute errors based on the test set***



All things considered, the performance of the first model on the training set seems to be satisfactory, however, further examination on unseen data demonstrates that the predictions of the model are not so accurate. The size of the training set can justify why the model cannot generalize well, hence increasing the training set can lead to better prediction accuracy.