# Algorithm Overview and Implementation

**By Poornima Ramachandran**

**Introduction**

The data assigned to us is the parking data of the Aarhus city in the Denmark. The data consist of four different data fields. The data fields include vehicle count, total spaces available and garage code. Since we have just three variables to get the analysis we have merged the new data set of pollution data and used the variable named Ozone content. Now we need to design a model which tells us the effects of delay in parking. Hence we need to run a linear regression to predict the effects of ozone layers in the environment.

**Algorithm**

The Linear **Regression** is algorithm we are using for the studying the impact of the ozone layers because of the delay in parking. Linear regression is a supervised machine learning algorithm. This algorithm has been a base for many more complex algorithms in machine learning. Linear Regression is an approach for modelling to establish a relationship between a scalar dependent variable and one or more explanatory variable. The dependent variable is usually represented as by Y and the explanatory variable are usually denoted by X. if the regression has only one explanatory variable then it is called Simple Linear regression. If the regression model has many explanatory variables then it is called Multiple Linear regression. In a linear regression explains the relationships between linear predictor functions where the unknown estimated parameters. Using final output of the prediction is always a straight line. Linear regression has been useful in many practical uses. Most applications fall into one of two broad categories:

* If the goal is prediction, forecast, or error reduction then the linear regression can be used to fit a predictive model on an observed data set on the dependent variable to the explanatory values. After developing a fit model more value can be given to ensure that the model works fine.
* The linear regression model can also be used to quantify the relationship between the dependent variable and the explanatory variable.

The linear regression model are often fitted using the concept of least square approach. They can also be fitted by minimizing the lack of fit.

There are a number of assumptions made to build a proper regression model. The assumptions are

* **Weak homogeneity:** This means that predictor variables or the explanatory variable can be treated as the fixed values rather than just random variables. This means that if we assume that the model is error free although this is not realistic we can make a conclusion we can predict the final output of the linear regression model free of errors.
* **Constant Variance :** This means that different response variable have the same variance in their same variance of errors regardless of the value of the predictor variables
* **Independence of errors:**  This assumption indicates that the errors of the response variables are not correlated with each other. This is a much stronger assumptions as any kind of correlation between the errors will make the prediction of the explanatory variable a bit difficult to build the final model.
* **Lack of Multicollinearity:** The lack of Multicollinearity in the explanatory variables will be make us assess the dependent variables will be in a way which we can ensure we get realistic results for an analysis.
* **Linearity:** This means that the mean value ofthe response variablein the linear combination of the parameter and the predicator variables. This is much less restrictive than the other assumptions.

Now the concept of the linear regression works on the estimation method to ensure that we have make predictions with maximum accuracy. A large number of procedure in machine learning the have been developed for parameter estimation inference of the results. The examples in this categories are the classification techniques and allocation methods. The estimation methods are of two types they are 1. Ordinary least square methods and 2. Generalized least square methods. As the accuracy of the linear regression model depends a lot on the least square parameter. The least square parameter has to be minimized to get the accurate model to forecast the dependent variable.

Now let us look at the estimation methods of the linear regression models in detail.

* **Ordinary Least Squares**: This is the simplest and the common estimator models to be used in the prediction of the dependable parameters in the linear regression model. This is very easy to compute and the easy to analyze. As the only action which we need to do here is to ensure that we minimize the least square errors to get accurate model for linear regression model.
* **Generalized least square**: This is an extension ordinary least square methods that allows efficient estimation of the beta values are different or are correlated in anyways. As long as the estimation values are not correlated the error terms will also be independent of each other.

The Linear regression model have found its applications in various fields such as biological, behavioral and social sciences to describe possible relationships between variables.

**Trend Lines**: A trend line represents the trends of the long term movements of the dependent variable after the explanatory have been accounted for. This gives us the exact predications of the dependent variables.

**Epidemiology:**  The early evidence which are related to tobacco smoking to mortality and morbidity came from the observational studies employing regression analysis.

**Finance**: The capital and asset price model uses linear regression as well as the concept of beta for analyzing and quantifying the risk of the model.

**Economics**: This is the one arena where the linear regression model is used as the predominant empirical tool in economics.

**Environmental** **Science**: Linear regression finds the applications in a wide range of environmental science applications. This was useful in the analysis of fish and benthic survey to measure the effects of pulp mill or metal mine effluent on the aquatic system.

Post applying the linear regression analysis on the parameters we use the time series analysis to build predict the effects of the delayed or inefficient parking systems on the environmental measures.

**Time Series Analysis**

A time series is a sequence taken at a successive equally spaced points in a time period. The time series analysis compares methods for analyzing times series data in order to extract the meaningful statistics and other characteristics of the data. Time series forecasting is used to predict future values based on the historical data we have. While the regression analysis is often employed in such a way that the current values are independent of the time series affect. A time series is a series of data particular data points indexed in time order. A time series is a sequence taken at successive equally spaced points in time.

The fields which has motivated the time series analysis are quantitative financial analysis, communication engineering for signal detection and estimation. It further found its use in pattern recognition, clustering, machine learning, anomaly detection and forecasting techniques. The other advantage of the time series plot is that it can be used for panel data. The panel data is a general class of multidimensional data where we can implement the time series data set in a one dimensional data panels. A data set can exhibits both panel data and time series plot. Further when the information is transferred across time often we get a good time series plot. One of the classic example of the statistical Markov model is which the system model is assumed to be unobserved stats and yet we can make a process a time series plot to get the information.

The major goals of the time series plot are as follows:

* Identify patterns in correlated data that is trends and seasonal variations.
* Understanding and modeling the data
* Prediction of short term trend analysis
* We can also perform intervention analysis which says how a single event can change the time series.
* And also how a deviation from the point can indicate a problem in the analysis or the forecast.

Understanding the mechanisms of a time series will allows us to build a mathematical model that can explain way the prediction and monitoring can be done. The assumptions we make to build a time series model is that it follows a systematic pattern. The most common patterns are generally trends and seasonality. Trends time series is either linear or quadratic in nature. To find the trends average or regression analysis is often used. Seasonality is a trend that repeats itself systematically over time. The second assumptions is that the data exhibits a random process so it is sometimes hard to understand the systematic patterns within the data. Observations made over time will be either discrete or continuous. One of the most important property of the time series is that the observations are dependent on the previous observation. This phenomenon is called auto correlation. Hence this will lead to the challenging process of identifying the auto correlation.

Now there two main approaches for analyzing the time series. They are 1. The time domain and 2. Frequency domain. The analysis in the time domain is done is often used for stochastic observations. The frequency analysis are used periodic and cyclical observations. As already stated above that the time series data is correlated which means measurement are related to the previous observation and one change in the parameter might affect the entire analysis.

**Pseudo code**

As we are using regression analysis for studying the effects of the delay in parking on the ozone layer.

1. Load the parking data into the system
2. Load the pollution data into the system
3. Merge the two data sets In the data frame
4. Name the data fields.
5. Take the Ozone count to the dependent variable
6. Perform the linear regression using lm function.
7. Check the P values of each parameter
8. Identify the parameters that affect the ozone count in the model
9. Plot the graph for the same.

The above pseudo code will be blueprint for our process of implementation of our analysis. From the pseudo code we can state the analysis in details. The dependent variable is the ozone content in the model. The explanatory variables are vehicle count and the total number of open spaces in parking spot.

**Work Flow of the analysis**

The work flow as shown in the figure indicates the work flow which is used in performing the analysis is described as follows:

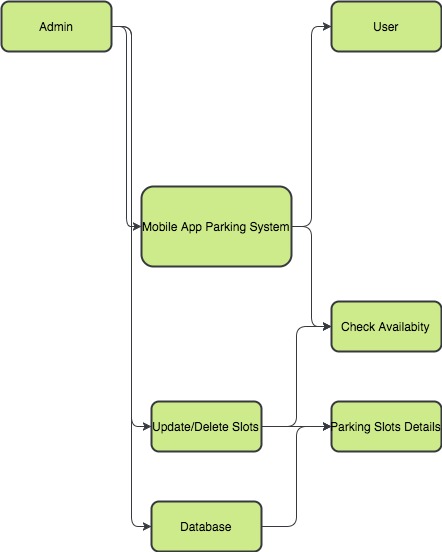
Level 1: The city’s transport the department will be building a data base with the garage codes the total number of slots available in the garage.

Level 2: Then we have a parking app which can be loaded on it the phones of all citizens. The app will act as an interface between the citizens and the centralized parking systems.

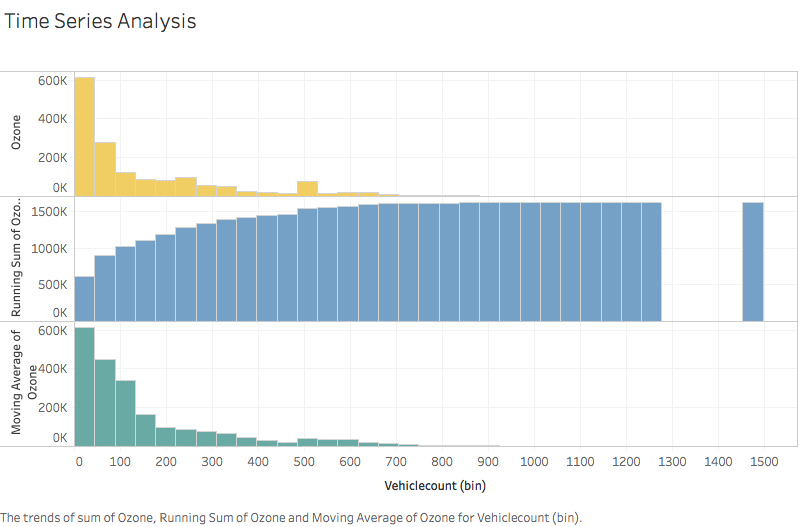
Level 3: The administrator and the user will have the access to the system.

Level 4: The citizen will be able to receive the real time update of the available parking spaces where the he wants to park and once he confirms he will get the Quick response code which will reserve a spot at parking space.

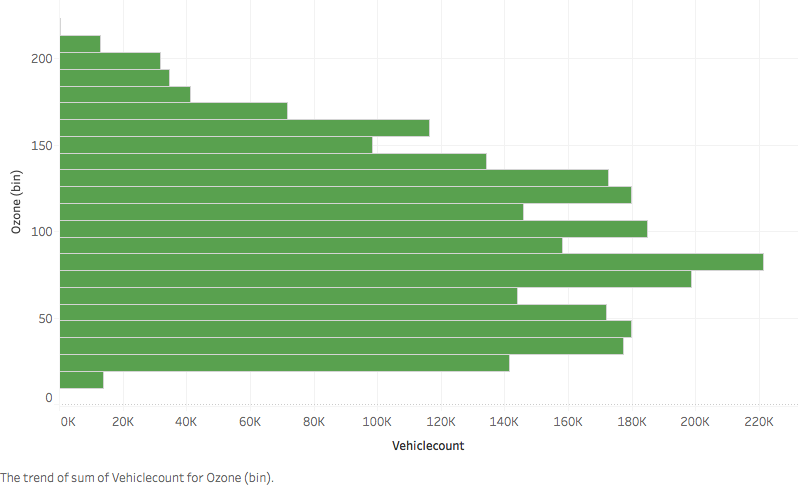
Level 5: Once confirmed the driver does not need to drive around to find a parking in the area.

****

**Time Series Analysis of the data**



The above mentioned time series analysis shows the trends of the ozone layer effects based on the running average of the vehicle count which is passing through the particular area.



This time series plot shows the effect the number of vehicles impacts the ozone index count in the atmosphere. It shows that as the number of the vehicles increases there is an increase in the air pollution which in turn is affecting the environment.

As the analysis shows that we need to ensure that we have a proper parking infrastructure in place as the number of vehicles increase in the city to ensure that we have a better environment for our future generation.