Assignment 9

From the “pollution.csv” file, predict the mortality rate (y) from the remaining attributes. Use linear regression and then SVM to show how well you can do this prediction (using RMSE). Then tune the SVM and find the best model you can generate. Calculate the error to show if you were able to improve on regular SVM.

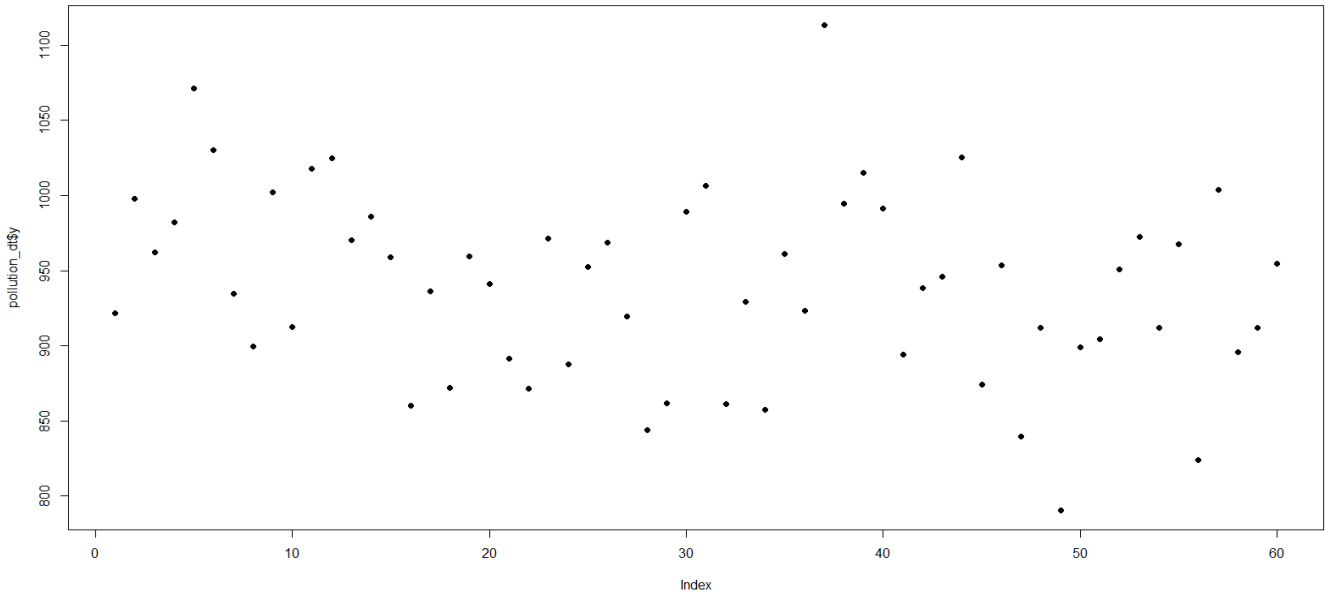
→

Linear Regression:

```
> pollution_dt  <- read.csv(file.choose(), header = TRUE, sep = ",",fileEncod
ing="UTF-8-BOM")

> View(pollution_dt)
> #Running the linear regression to get the model
> model1 = lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13++x14+x15, data=pol
lution_dt)
```

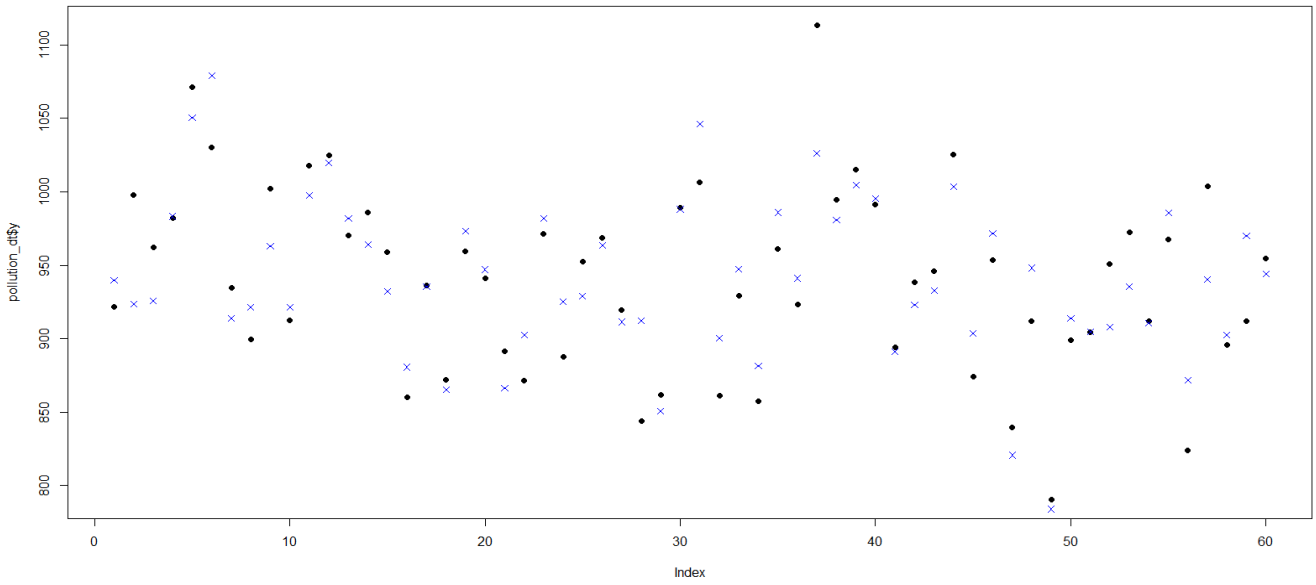
Representing the original Y values from the Pollution dataset.

```
> plot(pollution_dt$y,pch=16)
```




```
> # Prodict the Y values using the regression model build in the above step
> predY = predict(model1,pollution_dt)
```

```
> # Predicted Y values from the model (predY)
> predY
        1         2         3         4         5         6         7
8         9        10        11        12
 939.6121  923.3809  925.8145  983.2377 1050.2914 1078.9176  913.7189  921.31
47  962.9867  921.2663  997.4819 1019.7219
       13        14        15        16        17        18        19
20        21        22        23        24
 981.8444  963.9076  932.1272  880.6181  935.6055  865.1618  973.0040  947.07
82  866.2692  902.6774  981.8625  925.0714
       25        26        27        28        29        30        31
32        33        34        35        36
 928.9700  963.3068  911.4307  912.1184  850.6837  988.0713 1046.0043  900.44
96  947.2664  881.3805  985.9110  941.1068
       37        38        39        40        41        42        43
44        45        46        47        48
1026.1940  980.6364 1004.4299  995.1394  890.9952  923.1768  932.8288 1003.38
25  903.5031  971.5438  820.7878  948.1324
       49        50        51        52        53        54        55
56        57        58        59        60
 784.1330  914.0393  904.5584  907.9941  935.5534  910.6157  985.7851  871.69
77  940.3784  902.5954  969.8340  943.9045
> # Original Y values from the Pollution dataset
> pollution_dt$y
 [1]  921.87  997.88  962.35  982.29 1071.29 1030.38  934.70  899.53 1001.90
912.35 1017.61 1024.89  970.47  985.95
[15]  958.84  860.10  936.23  871.77  959.22  941.18  891.71  871.34  971.12
887.47  952.53  968.67  919.73  844.05
[29]  861.83  989.27 1006.49  861.44  929.15  857.62  961.01  923.23 1113.16
994.65 1015.02  991.29  893.99  938.50
[43]  946.19 1025.50  874.28  953.56  839.71  911.70  790.73  899.26  904.16
950.67  972.46  912.20  967.80  823.76
[57] 1003.50  895.70  911.82  954.44


> # Plot predicted Y values and original Y values on the plot in sequential o
rder of datapoints in the dataset
> points(predY, col="Blue", pch=4)
```

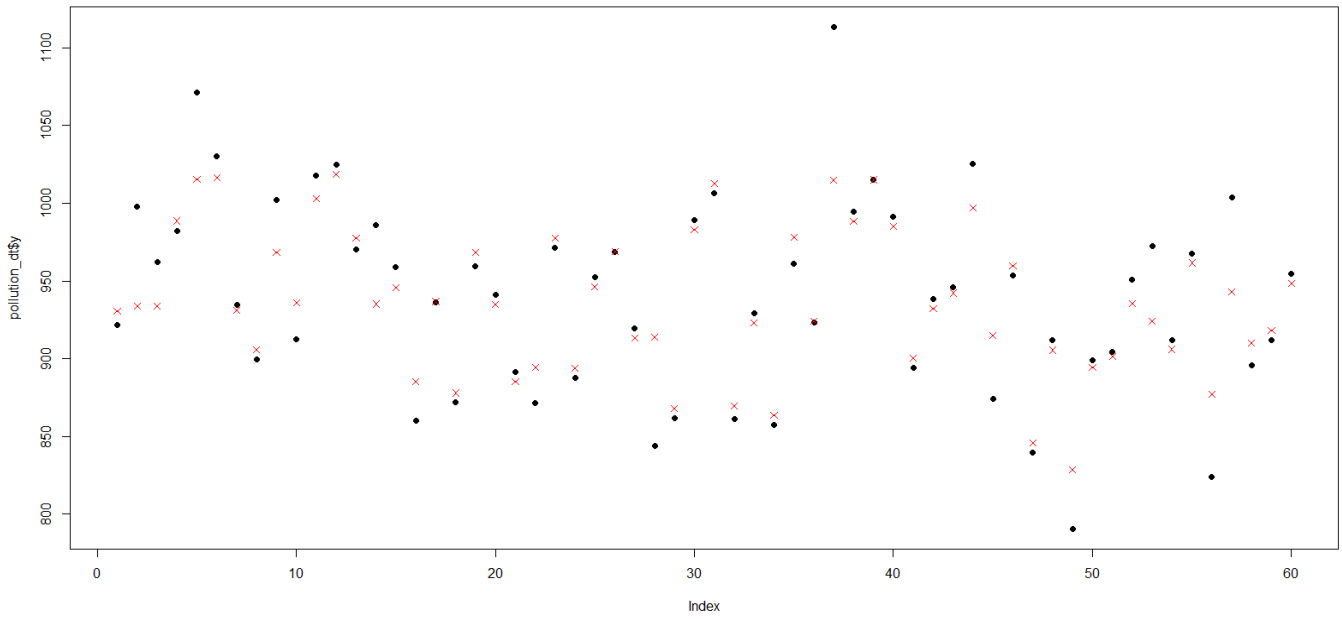


```
#RMSE error function
> rmse <- function(error)
+ {
+   sqrt(mean(error^2))
+ }


> error1 = model1$residuals
> lrPredMSE = rmse(error1)

# MSE of the predicted values using linear regression
> lrPredMSE
[1] 29.91123
```

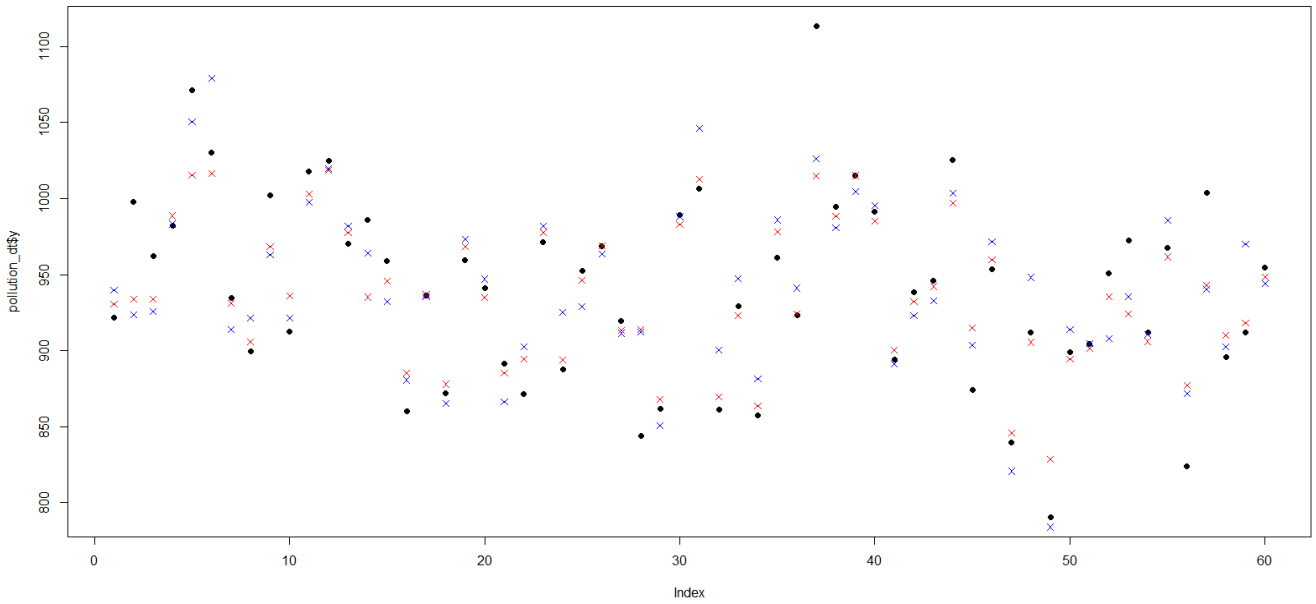
Thus, the MSE using linear regression model is 29.91123

Support Vector Regression:

```
> library(e1071)
>
> model2 = svm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13++x14+x15, polluti
on_dt)
>
> PredY2 = predict(model2,pollution_dt)
>
> #plot only Y values from the pollution dataset
> plot(pollution_dt$y,pch=16)
>
> #plot Predicted Y values using the SVM
> points(PredY2, col="Red", pch=4)
```



```
# Plot the SVM (Red), LR (Blue) and original Y (Black) values)
> points(predY, col="Blue", pch=4)
```




```
> error2 = pollution_dt$y - PredY2
>
> svmPredRMSE = rmse(error2)
> svmPredRMSE
[1] 26.94244
```

The MSE using the SVM is 26.94244.

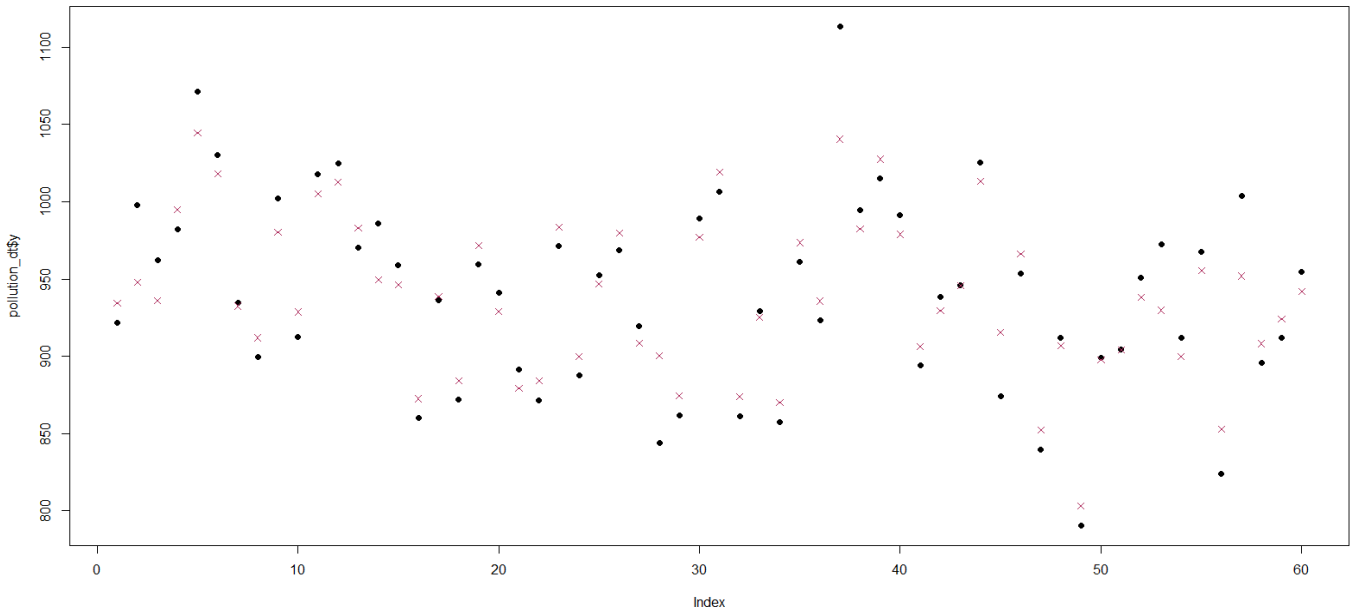
```
> lrPredMSE
[1] 29.91123
```

From the linear regression MSE value (29.91123) and SVM MSE value (26.94244), it is clear that we can get better values of MSE using the SVM. Thus, SVM fits this data better.

Tuning the SVM for best model

```
> model3 = tune(svm,y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13++x14+x15, da
ta=pollution_dt, ranges=list(epsilon=seq(0,1,0.1),cost=seq(1,10,1)))
>
> bestmodel = model3$best.model
> bestPred = predict(bestmodel,data=pollution_dt)
```

```
> #plot only original Y values from the pollution dataset
> plot(pollution_dt$y,pch=16)
>
> #plot redicted Y values using the Best tuned SVM model
> points(bestPred, col="maroon", pch=4)
```




```
> best_error  = pollution_dt$y - bestPred
>
> best_RMSE = rmse(best_error)
> best_RMSE
[1] 21.47147
```

The best tuned SVM model gives MSE as 21.47147.

```
> best_RMSE
[1] 21.47147
> svmPredRMSE
[1] 26.94244
> lrPredMSE
[1] 29.91123
```

From the above results, we can say that tuned SVM model given least MSE, thus we can improve the regular SVM by tuning the model.