PROJECT REPORT

(Project 1 - Regression Analysis)

(EAS 508 – Statistical Learning and Data Mining I)

Team Members:

1. Jahnavi Gangu (jgangu@buffalo.edu)
2. Sai Teja Ankuru (saitejaa@buffalo.edu)
3. Utkarsh Mathur (umathur@buffalo.edu)

Instructor:

Dr. Leonid Khinkis

---

1. INTRODUCTION

- The aim of the project is to practice regression analysis skills on 2 datasets, one of which was provided by the instructor while the other dataset was picked on teams discretion. Our team opted for Gas Emission Data Set.
- The health dataset (provided by the instructor) had target variable named “X1”, 4 predictors named “X2”, “X3”, “X4”, & “X5”, and 53 observations. On the other hand, Gas Emission Data Set has “NOX” as its target variable, 10 predictor “AT”, “AP”, “AH”, “AFDP”, “GTEP”, “CDP”, “TIT”, “TAT”, “TEY”, & “CO”, and 36733 observations almost uniformly spanning over 5 years.
- We have performed regression analysis with 2 key agendas:
  a. Comparing performances of various models.
  b. Impact of number of observations and predictors on each model.
- We have used 7 models namely, Linear Regression, Polynomial Regression, Principal Component Regression, Partial Least Squares Regression, Cubic Spline, and Support Vector Machines along with various techniques like Subset Selection, Lasso, Ridge, and Cross-Validation to reduce test MSE of models.

2. DATA HANDLING

- We began our analysis by importing the health data set and gas emission data set.

- For Gas Emission data set, we have included the year of data collection as an additional predictor raising the count to predictors to 11.
- As the gas emission dataset is divided into 5 files for each year the data was collected, we concatenated these dataset partitions into a single dataset.
- Using 80-20 split, we partitioned our datasets into their respective training and test datasets. The Health data set has 35 training and 18 test observations whereas the Gas Emission data set has 29357 training and 7376 test observations, respectively.

3. RESULTS & ANALYSIS

MODELS	Health Data Test MSE	Gas Emission Data Test MSE
Multiple Linear Regression (LR)	1.109641	54.84162
LR after Subset Selection	1.414834	54.84162
LR with Lasso	1.706216	54.900184
LR with Ridge	1.706216	58.760404
Polynomial Regression (PR)	**0.941459**	3250.5496
PR with Lasso	1.626555	54.900184
PR with Ridge	1.701605	58.760405
Principal Component Regression	1.252638	57.917343
Partial Least Squares Regression	1.395639	55.064885
Cubic Spline	54.71878	25.361899
Support Vector Machine (Linear Kernel)	1.597123	58.226347
Support Vector Machine (Polynomial Kernel)	9.452273	35.891706
Support Vector Machine (Radial Kernel)	1.177949	**15.295987**

(Table 1 – Test MSEs for all the models trained on Health Data Set and Gas Emission Data Set)

3.1. MULTIPLE LINEAR REGRESSION (LR)

- The first model we built is linear regression to establish a baseline model to test other models against.
- For both the datasets we trained the linear regression model on all the available predictors.

3.2. LR AFTER SUBSET SELECTION

- Subset Selection techniques of Best Subset Selection, Forward Subset Selection, and Backward Subset Selection were used in correspondence with C_p, BIC, and Adjusted R^2 to select optimal subset of predictor.
- For Health dataset, we selected on "X5" as the predictor based on the C_p, BIC, and Adjusted R^2 metrics, whereas, for Gas Emission dataset, we selected 8 predictors ("AT", "AP", "AH", "TIT", "TAT", "TEY", "CO", "year").

3.3. POLYNOMIAL REGRESSION

- For the Health dataset, polynomial regression outperforms all the other models while it is quite the very opposite for the Gas Emission dataset.

3.4. LASSO & RIDGE

- We have used cross-validation to determine the best value of λ.
- For the Health dataset, LR with Lasso and LR with Ridge delivered similar results, whereas, PR with Lasso has lower test MSE than PR with Ridge.
- For Gas Emission data set, Lasso outperformed Ridge for both LR and PR. They also return the same test MSE suggesting that the variability induced polynomial terms over linear terms have less impact on the performance of models on larger datasets.

3.5. PRINCIPLE COMPONENT REGRESSION (PCR) & PARTIAL LEAST SQUARES REGRESSION (PLSR)

- For Health Data, PCR gives better results than most of the models. This may indicates that the number of predictors are very large with respect to the number of observation.
- For the Gas Emission Data, PCR performs sub-par.
- PLSR has average performance for both Health data and Gas Emission data adding little value to the analysis.

3.6. CUBIC SPLINE

- For the Health data, cubic splines gives absurdly high test MSE which can be justified by the high p-values of all the predictors and very few observations.
- For the Gas Emission data, cubic spline performs exceptionally well on test data.

3.7. SUPPORT VECTOR MACHINES

- With **Linear Kernel**, SVM returns high test MSE for both the dataset.
- With **Polynomial Kernel**, SVM returns a high test MSE for Health data where a very low test MSE for the Gas Emission dataset.

- With **Radial Kernel**, SVM returns low test MSE for both datasets specially the lowest test MSE for Gas Emission data.

4. CONCLUSIONS

- The best model for Health data is **Polynomial Regression** (Test MSE = 0.941459) but Linear Regression (Test MSE = 1.109641) and Radial Kernel Support Vector Machine (Test MSE = 1.177949) fits the data very well too.
- For the Gas Emission data best fitting model is **Radial Kernel Support Vector Machine** (Test MSE = 15.295987) but Cubic Spline fits very well (Test MSE = 25.361899) too.
- The Health data has 35 training observations out of the 53 observations whereas the Gas Emission data has 29357 training observations out of 36733 observations, therefore, this setting of datasets given us a good idea about model behavior in different observation density environments.
- This analysis sheds light on the curse of dimensionality. The training set of Health data has 35 observations for 4 predictors, which gives models like Cubic Splines and Polynomial Kernel SVM the curve of dimensionality as these models perform exceptionally well in the Gas Emission data where there 29357 training observation just for 11 predictors.

5. REFERENCES

1) UCI Machine Learning Repository: Gas Turbine CO and NOx Emission Data Set Data Set
2) Home - RDocumentation
3) Resources - Second Edition — An Introduction to Statistical Learning (statlearning.com)
4) Elements of Statistical Learning: data mining, inference, and prediction. 2nd Edition. (su.domains)
5) RPubs - ISLR - Chapter 6 Solutions
6) RPubs - ISLR - Chapter 7 Solutions
7) RPubs - ISLR - Chapter 9 Solutions