A comparative approach for model selection and prediction using presence only data in R

Vinayak Chaturvedi, Nitin Sharma and Stavan Anjaria

University of New Hampshire, Durham

Course: CS850 Machine Learning

ABSTRACT

This paper is a comparative summary of experiments conducted using multiple machine learning models upon presence only data for predicting distribution of a species in New England region. Each region is divided into rectangular regions ***L***. Each region can be represented by unique Latitude and Longitude values. The following models have been taken into consideration:

- Subset Selection
- Polynomial Regression
- Lasso Regression

(I) Brief Introduction

A lot of the data available to humans are presence-only i.e. they are reported on an observation event E. We have selected three regression methods to achieve the prediction task. Ability to represent a model with minimum features and test error is important to achieve good prediction values. Such assumptions are key factors in determining the frequency data missing for cells.

(II) Data Preprocessing

Data Filtering: For our implementation we are using the synthetic data set for analysis. Post importing required packages for the Mini-Project we used Microsoft Excel tool for data preprocessing. In order to obtain a consolidated data table containing the target "Frequency" along with features we used VLOOKUP in Excel and created a file "Filtered_Landscape.csv" containing target and features for which target values were available. We will use this dataset to perform predication tasks. We have also used set.seed(1) as default to reproduce same results on each trial.

Establishing Relationship with Frequency: We produced scatterplot matrix for each feature against target using ggpairs() in R. From the Pearson Correlation obtained we can see that all the features appear to have non-linear relationship with target (Figure 1). However, it assumes that variables under consideration are normally distributed. Additionally, Pearson Correlation is sensitive to outliers as well. Hence, we yet proceed to use methods which are linear in nature as well.

Handling Missing Data: To deal with missing data, deletion and imputation methods were studied. The synthetic data set did not have any missing data. Excel IFNA and R's na.omit() with dim() were used to confirm the dataset did not have any missing values.

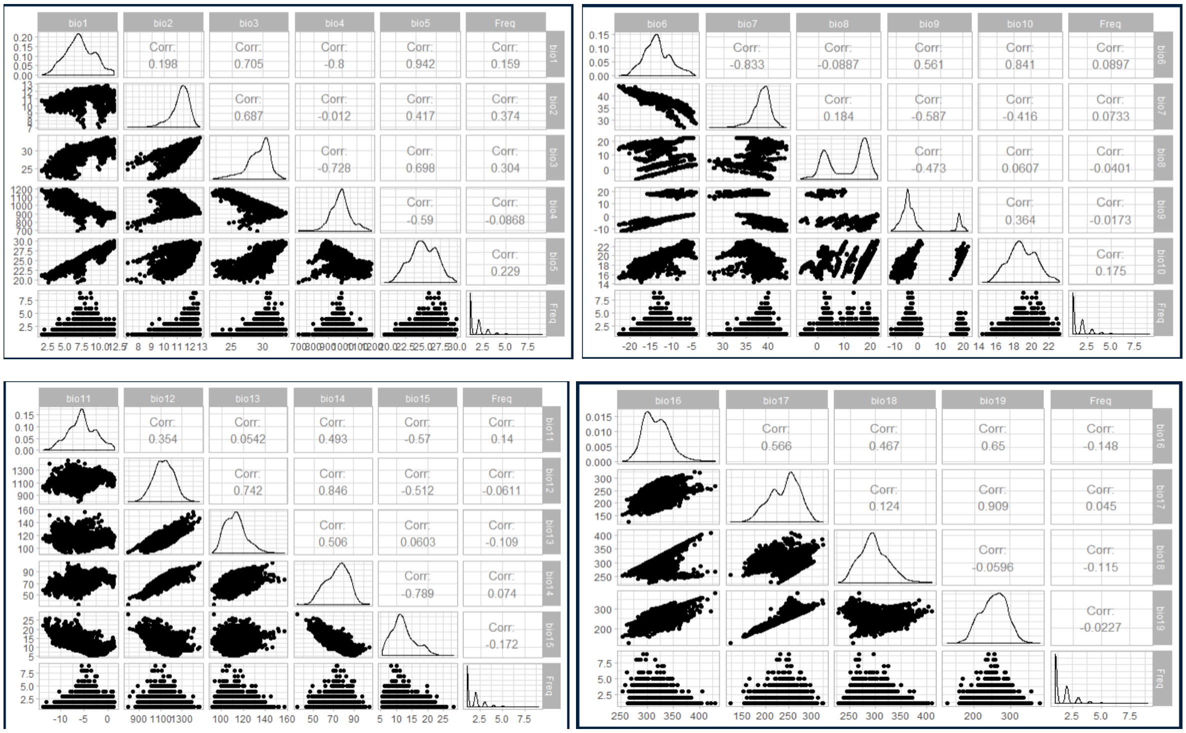


Figure 1: ggpairs() of different features with frequency was plotted to understand the relationship within the data.

(III) Methods Implemented

Subset Selection:

Experiments with multiple subset selection techniques were performed and in general certain steps were followed: "Filtered_Landscape.csv" was read and data was split into training and test since it is optimal to fit the model on the training data rather than complete dataset. R squared, Cp and bic statistics were calculated and plot() function was used for visualization. Cross-validation was performed and coefficients of model with minimum RMSE were extracted using **coef().** A prediction file was generated using the intercept and respective features. For our experiments we have consistently selected methods with maximum of 9 predictors.

Following variations in Subset Selection were performed:

- **Best Subset Selection with all predictors**
  In this method we allowed all features from Bio1 to Bio19.
- **Best Subset Selection with relevant predictors**
  Some features can be represented as a combination of others and hence manual feature selection was applied. We used

bio1+bio3+bio4+bio7++bio10+bio11+bio12+bio13+bio14+bio15+bio16+bio17 as argument while fitting and predicting.

- **Forward and Backward Subset Selection**

Lasso Regression

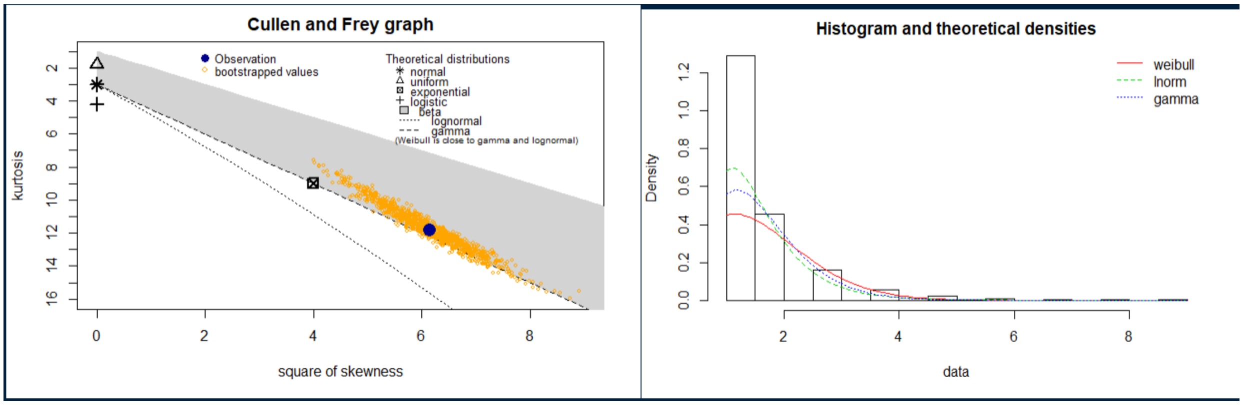
We fitted the lasso regression using glmnet package.

Polynomial Regression

Polynomial regression was fitted with caret package. Polynomials of different orders are used to obtain RMSE.

(IV) Prevalence and Likelihood Estimation

Prevalence (also known as occurrence probability) is not identifiable from presence-only data. It is a spatially averaged occurrence probability (e.g. Ward et al. 2009). The likelihood estimation is applied through Bayes rule. Andrew Royle's paper on "Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions" provides an algorithm for likelihood estimation in R. We attempted to apply the algorithm on our data set. However, we noticed that the resulting values in the paper z(x) has been assumed to come from a normal distribution. We proceeded to find the most appropriate distribution fit for our results. For our ease of implementation, we used **fitdistrplus** package in R to find the appropriate fit. We obtained the Cullen and Frey graph for our data and noticed that corresponding skewness-kurtosis plot generated has positive skewness and a kurtosis not far from 6. Hence, Weibull, gamma and lognormal distributions are considered. The plots are represented in Figure 2.



From the Figure 2 we can notice that the lognormal distribution represents the observed data closely. We did not proceed to analyze Likelihood for this distribution.

(V) Results

Table 1 below summarizes the Adjusted R squared, Cp, bic and RMSE for all the methods where applicable.

Method Implementation	**R2**	**Cp**	**Bic**	**RMSE (coef if applicable)**	**Selection Criteria?**
Subset with all predictors	10	10	7	0.8722131 (8)	Not Selected
Subset with manual selection	9	9	5	0.8697909 (9)	Selected
Forward Subset Selection	10	9	7	0.8730941	Not Selected
Backward Subset Selection	10	9	7	0.8725426	Not Selected
Lasso	-	-	-	0.8739276	Not Selected
Polynomial (Model1)	-	-	-	0.9504023	Not Selected
Polynomial (Model2)	-	-	-	0.9514861	Not Selected
Polynomial (Model3)	-	-	-	0.8706975	Not Selected

Table 1: Analysis of machine learning methods on presence only data set

From the table above we can see that Subset Selection with manual selection of predictors is having minimum RMSE at 9 variable selection (highlighted in yellow). Hence, we proceed with generation of output predictors by extracting the coefficients of this method and fitting a model that represents our data.

The output file “output_subset_best_manual.csv” contains the list of all predicted frequency values generated. Note that, we have applied text to columns to the data before submitting the file.

(VI) Conclusion

Post experimental analysis of various regression models we can conclude that different models can generate different results. Our selection was based on minimizing the test error using cross validation across all methods. This does not guarantee obtaining perfect results but the attempt to apply constraints like minimization of test errors greatly reduce the chances of output differing from actual values.

(VII) References

- https://en.wikipedia.org/wiki/Pearson_correlation_coefficient
- https://towardsdatascience.com/pearson-coefficient-of-correlation-explained-369991d93404
- https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4
- https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/
- http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-in-r/#k-fold-cross-validation
- https://cran.r-project.org/web/packages/fitdistrplus/vignettes/paper2JSS.pdf
- https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2041-210X.2011.00182.x
- Royle, J. A., Chandler, R. B., Yackulic, C., & Nichols, J. D. (2012). Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. Methods in Ecology and Evolution, 3, 545–554.
- Hastie, T., & Fithian, W. (2013). Inference from presence-only data; the ongoing controversy. Ecography, 36(8), 864–867.