Piecewise Regression Analysis through Information Criteria using Mathematical Programming

Ioannis Gkioulekas, Lazaros G. Papageorgiou *

Centre for Process Systems Engineering, Department of Chemical Engineering, UCL (University College London), Torrington Place, London WC1E 7JE, UK

*Corresponding author. Tel: +44 20 7679 2563

E-mail addresses: ioannis.gkioulekas.16@ucl.ac.uk (I. Gkioulekas), l.papageorgiou@ucl.ac.uk (L.G. Papageorgiou)

Abstract: Regression is a predictive analysis tool that examines the relationship between independent and dependent variables. The goal of this analysis is to fit a mathematical function that describes how the value of the response changes when the values of the predictors vary. The simplest form of regression is linear regression which in the case multiple regression, tries to explain the data by simply fitting a hyperplane minimising the absolute error of the fitting. Piecewise regression analysis partitions the data into multiple regions and a regression function is fitted to each one. Such an approach is the *OPLRA* (Optimal Piecewise Linear Regression Analysis) model (Yang et al., 2016) which is a mathematical programming approach that optimally partitions the data into multiple regions and fits a linear regression functions minimising the Mean Absolute Error between prediction and truth. However, using many regions to describe the data can lead to overfitting and bad results. In this work an extension of the *OPLRA* model is proposed that deals with the problem of selecting the optimal number of regions as well as overfitting. To achieve this result, information criteria such as the Akaike and the Bayesian are used that reward predictive accuracy and penalise model complexity.

Keywords: Mathematical programming, Regression analysis, Optimisation, Information criterion, Machine learning

1 Introduction

Regression analysis is a predictive modeling technique that estimates the relationship between variables. Given a multivariate dataset, this modeling technique will try to formulate the correlation between the set of dependent variables, called predictors, and the independent variable, called response. The final goal of the analysis is to create a mathematical model that describes that relationship.

There are various methods for regression available in the literature that approach the topic in different ways. One of the most widely known methods is linear regression that establishes a relationship between the response and the predictors by fitting a simple straight line. Other more sophisticated approaches include Support Vector Machine Regression (SVM) (Smola and Schölkopf, 2004), K-nearest neighbors (KNN) (Korhonen and Kangas, 1997), Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991) and Random Forest (Breiman, 2001). In the field of mathematical programming there is the Automated Learning of Algebraic Models for Optimisation (ALAMO) (Cozad et al., 2014; Wilson and Sahinidis, 2017) and a segmented regression approach called Optimal Piecewise Linear Regression Analysis (*OPLRA*) (Yang et al., 2016).

In machine learning, constructing an accurate predictive model involves fitting it to a set of training data and then fine-tuning its parameters in such a way that this model will be able to make reliable predictions on new untrained data. Tuning these parameters and deciding on the complexity of the final model is essential in order to avoid overfitting.

Overfitting is a common concern when constructing a predictive model. When fitting a regression model to a set of data, it is possible to create such a complex structure that the final model will also predict the noise of the data. That means the model is not able to describe the overall population in the dataset and as a result has poor performance. Figure 1 illustrates this problem.

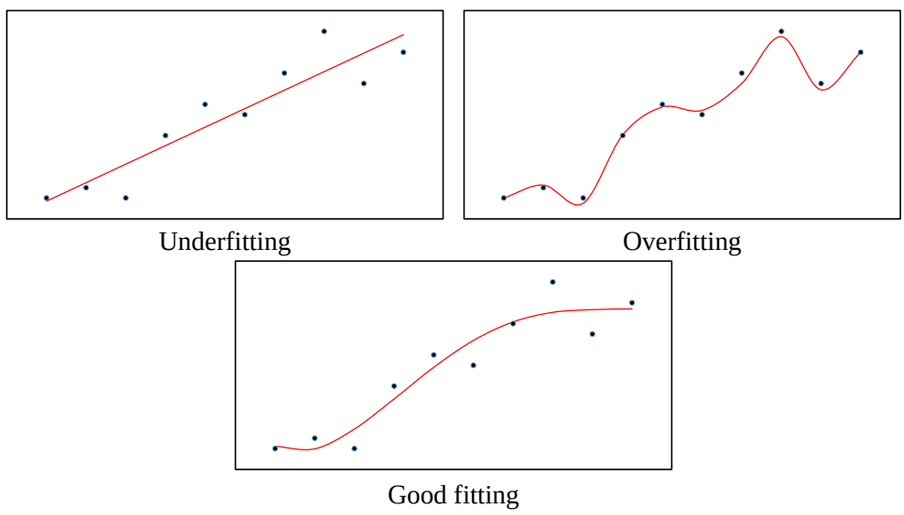

Figure 1: Visual representation of over- and under-fitting in regression

By describing the noise that exists in the data, the resulting model is very sensitive and is affected by small fluctuations in the data (Hawkins, 2004). On the opposite

end there is underfitting, the process of constructing a very simple model that is not capable of capturing the information that exists in the data, hence creating the tradeoff between variance and bias. In machine learning, this problem is about trying to create an algorithm that will have good predictive performance and it will be able to generalise well. The term bias usually refers to underfitting because it can cause a model to miss the relation between the data. Variance on the other hand, is associated with overfitting and modeling the noise that exists in the data.

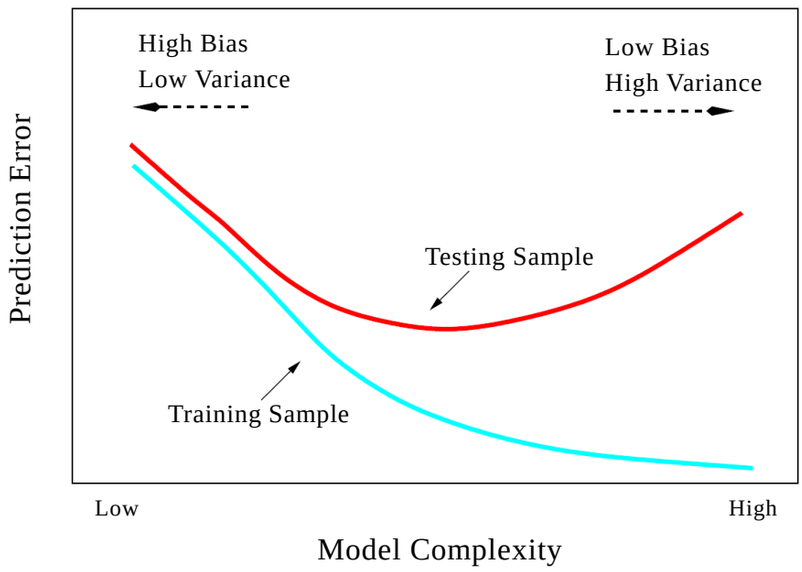

Figure 2: Test and training error as a function of model complexity (Hastie et al., 2008)

Figure 2 illustrates the impact that model complexity has on performance. Low model complexity leads to a model with high bias meaning that there is a large prediction error in both training and testing. High model complexity leads to a model with high variance resulting in a big performance gap between the training and testing phase.

Deciding on the final complexity of a model requires trial and error and fine-tuning, until the final model has been chosen. During this process, every time a parameter is changed, a new model is constructed. So in the end there exists a set of candidate models to choose from and the objective is to determine which one best approximates the data. Tackling this problem of model selection can be achieved using information criteria. Information criteria are measures of the relative goodness of fit of a statistical model. There are various information criteria in literature with two of the most popular being the *Akaike Information Criterion* (AIC) and the *Bayesian information criterion* (BIC) (Wagenmakers and Farrell, 2004).

1.1 Information Criteria

These two criteria have been established as two of the most frequently used in the literature for model selection problems, with a wide variety of applications. A few examples include the wine industry (Snipes and Taylor, 2014) where different models are

compared in an attempt to explore the relationship between ratings and prices of wines, cancer research where AIC was used to develop a prognostic model in patients with germ cell tumors who experienced treatment failure with chemotherapy (International Prognostic Factors Study Group, 2010)

These two criteria have also been used in a number of statistical and machine learning methods such as outlier detection (Lehmann and Lösler, 2016) and feature selection. Kimura and Waki (2018) proposed a branch-and-bound search algorithm formulated as a mixed integer nonlinear programming problem, minimising the value of the AIC, to perform feature selection. Sato et al. (2016) have also used both the AIC and BIC as measures of the goodness-of-fit to perform feature selection for logistic regression models.

The general formulation of the *AIC* and the *BIC* is as follows (Wagenmakers and Farrell, 2004):

$$AIC = -2 \cdot \ln(\hat{\mathcal{L}}) + 2K$$

$$BIC = -2 \cdot \ln(\hat{\mathcal{L}}) + K \cdot \ln(n)$$

where:

ln	is the natural logarithm
$\hat{\mathcal{L}}$	is the value of the log-likelihood function at its maximum point
K	is the number of parameters in the model
n	number of samples in the data

The AIC establishes a relationship between the Kullback-Leibler measure and maximum likelihood estimation method (Fabozzi et al., 2014). It is an estimate of the relative distance between the truth and the model that approximates it. The criterion is based on the idea that no model exists that perfecly describes the truth so the best we can do is approximate it. Given a set of candidate models the criterion can identify the model that performs the best (Burnham and Anderson, 2003).

The BIC however arises from a Bayesian viewpoint and belongs to a class of criteria that are "dimension-consistent" which differ from those that are estimates of the Kullback-Leibler measure. The formulation of the BIC is very similar to the AIC but the main difference is that the BIC is derived to provide a consistent estimator of the dimension of the data (Burnham and Anderson, 2003).

In regression analysis, if all the candidate models assume normally distributed errors with a constant variance, then the criteria can be reformulated as (Burnham and Anderson, 2003):

$$AIC = n \cdot \ln\left(\frac{RSS}{n}\right) + 2K \tag{1}$$

$$BIC = n \cdot \ln\left(\frac{RSS}{n}\right) + K \cdot \ln(n) \tag{2}$$

where:

RSS	is the residual sum of squares

1.2 Contribution of this work

Piecewise regression methods using linear expressions have the advantage of simplicity and model interpretability because of the linear expressions between the predictors and the response, but identifying the position of the break points is not an easy task. The break points are points at which the data will be split in order to define a new region. In the end, the final regression model will consist of a number of regions and break points with a linear expression fitted to each one. The segmented package, which is part of the R (R Development Core Team, 2016) library, is able to perform piecewise analysis (Muggeo, 2008, 2003). This package fits segmented regression expressions to a set of data but the user has to specify the number of regions as well as estimates for the position of the break points. The method then iterates until the final break points have been identified.

The *OPLRA* method (Yang et al., 2016) is a mathematical programming-based regression method that performs piecewise linear regression. This method is formulated as a Mixed Integer Linear Programming (MILP) problem that partitions a single variable into segments and fits linear functions to them. A big advantage of this method is the ability to simultaneously determine the position of each break point and calculate the regression coefficients. So the user is not required to give estimates for the break points. However, this method requires the number of regions for the partitioning of the data. A heuristic approach was proposed that could identify that number by iteratively solving multiple MILP models, increasing the number of regions every time.

In this work, an extension of the *OPLRA* model is proposed that includes information criteria, such as the AIC and the BIC, to fully automate the entire regression process and select the optimal number of regions. This novel approach requires a multivariate dataset as input and then can simultaneously decide on the optimal number of regions, the optimal position of the break points and can estimate the regression coefficients while directly minimising the value of the information criteria. This has been achieved by introducing new binary variables to the model that are able to 'activate' regions so that samples can be allocated to them.

By accommodating information criteria and reformulating the model to automatically decide on the number of regions for partitioning the data, the algorithm has the advantage of building piecewise linear models without demanding any user input. Also, by using the AIC and BIC the approach handles the trade-off between bias and variance and is able to choose the optimal number regions needed in order to build an accurate but not very complex model.

2 Optimisation Approaches

In this work, two new methods are proposed that extend the *OPLRA* mathematical programming model and include information criteria. The first approach is an iterative method that solves multiple MILP models and uses the AIC and the BIC to identify the optimal number of regions for piecewise regression. In the second approach, a single level MILP model is constructed that is able to optimally decide on the number of

regions and the optimal position of the break points while minimising the value of the criterion.

The original ORPLA mathematical model is presented in appendix A. This section contains the symbols for all the sets, parameters and variables that are necessary to formulate equations 12-19. A brief explanation of all the equations is also provided in that section to explain how each equation works and regression is performed.

2.1 Iterative approaches

In this section, the AIC and BIC are used to select the optimal number of regions. As stated previously, the criteria can be used in regression to select the best model in a set of candidate models. For this approach, the set of candidate models differ in the number of regions selected. With each added region, more parameters are introduced to the model since a mathematical function is fitted to each new region.

Using equations 1 and 2 and modifying them to fit the notation used in this work, equation 3 and 4 are derived:

$$AIC = |S| \cdot \log\left(\frac{\sum_s D_s^2}{|S|}\right) + 2(|M|+1)|R| \tag{3}$$

$$BIC = |S| \cdot \log\left(\frac{\sum_s D_s^2}{|S|}\right) + (|M|+1)|R| \cdot \ln(|S|) \tag{4}$$

$|S|$ is the total number of samples in the dataset. The term $\sum_s D_s^2$ is the residual sum of squares. Finally, the number of parameters is $(|M|+1)|R|$, where $|M|$ is the total number of predictors in the dataset and $|R|$ is the total number of regions. Each model consists of $|M|$ coefficients (one for each predictor) and one parameter for the intercept of each region. It is obvious that the more regions we add to the model, the more parameters we introduce resulting in a higher model complexity. In this section, two iterative variants have been introduced that use the same optimisation model but process the results with two different criteria. The two approaches are the *Piecewise Regression with Iterative Akaike information criterion* (***PRIA***) and the *Piecewise Regression with Iterative Bayesian information criterion* (***PRIB***). The two models can be summarised as follows:

minimise objective function 19

subject to constraints 12 - 18

post-process the results of the optimisation with equation 3 (***PRIA***)

post-process the results of the optimisation with equation 4 (***PRIB***)

The proposed methods still use an iterative approach to select the optimal number of regions. As a first step, the algorithm will try to identify a variable in order to partition the data based on that variable and create regions. So we fix the number of regions to R=2 and solve the OPLRA model multiple times, each time changing the partitioning variable and minimising the absolute deviation. In the end, the variable that yields the lowest error becomes the partitioning variable. It is worth noting that the information criteria could be have been used in order to identify which predictor variable should

be used to partition the data. But because the number of regions is constant, the only parameter in equations 3 and 4 that changes is the absolute deviation. So, using the information criteria for this part is not necessary.

The next step in the process is to identify the optimal number of regions. Since the partitioning variable has been selected, we solve the PRIA/PRIB model iteratively, adding an extra region with each iteration. At the end of each iteration we check the values of the criteria (either AIC or BIC depending on which method the user selected) and compare them with the values of the previous iteration. If there is an improvement (AIC or BIC decreases), then another region is added and a new iteration begins. Otherwise, the iterations are terminated and the final number of regions is the number of the iteration that had the minimum AIC/BIC value.

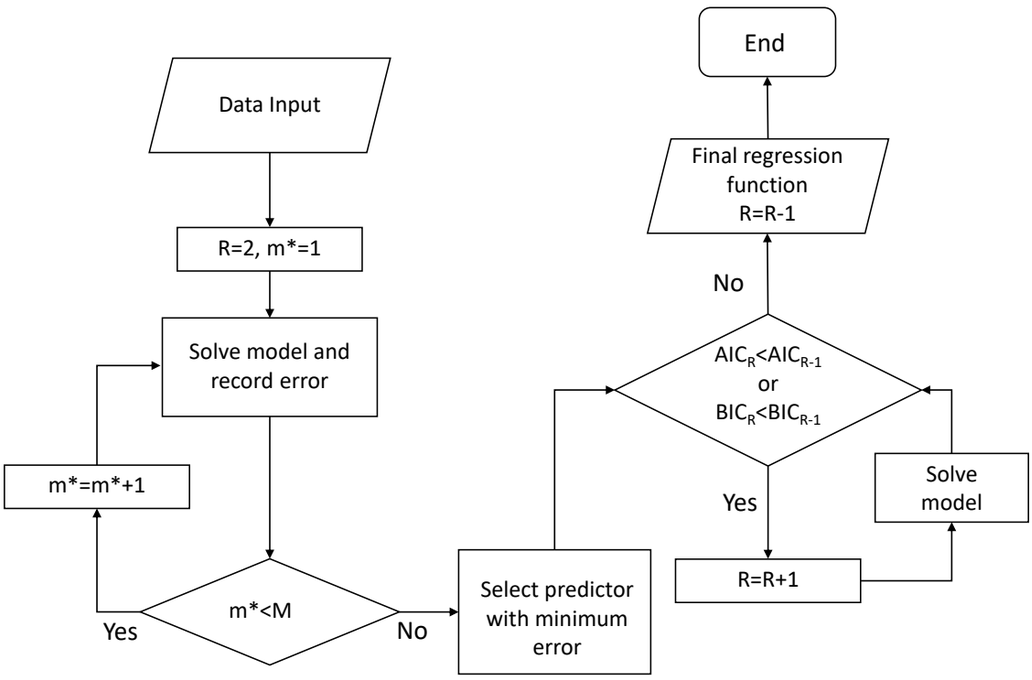

Figure 3: Flowchart for the proposed iterative approaches

The heuristic approach described in figure 3 is very similar to the one used in the original *OPLRA* model (Yang et al., 2016). The major difference is the use of the AIC and BIC in the final loop. By using them, the original stopping criteria is replaced. The main reason for this change is to avoid over or under fitting. The criterion is now solely responsible for the optimal number of regions.

Also, in the original work a heuristic approach was proposed for identifying the optimal number of regions for piecewise regression. To achieve that goal the authors introduced a user-specified parameter, called β, as a threshold to stop the iterations and converge to a solution. The elimination of the user-specified β parameter is a another improvement over the previous work. In order to assign a value to this parameter, Yang et al. (2016) performed a sensitivity analysis based on some specific datasets. This action might have an effect on overall performance and how well the method generalises, since the model was tailored around these datasets. With this new change there are no

user-specified variables and the model is independent of any data, hence allowing for possibly improved predictive accuracy.

However, this new modification to the model still has the drawback of solving the problem iteratively. This means that an MILP problem has to be solved for every new region that is added to the model. And with each added region the model becomes more complex and more computationally expensive to solve. To tackle this problem, a new approach has to be developed that solves the problem of identifying the number of regions and also fits a regression function simultaneously.

2.2 Single level MILP approaches

In this section, a single-level MILP model is constructed that is able to optimally partition the data into multiple regions, decide on the position of the break points and calculate the regression coefficients simultaneously. Instead of solving the model multiple times to identify the number of optimal regions through a heuristic approach, this time a single model is constructed and the information criteria are used as the objective function of the optimisation model.

More specifically, a new binary variable is introduced to the model to determine the number of selected regions. This variable is a decision variable that formulates whether or not a region has been selected. If a region is selected then samples can be allocated to that region according to equations 13 and 14. The objective function of this model is the minimisation of either the AIC or BIC. To overcome the non-linear nature of the criteria and formulate the problem as an MILP approach, some adjustments have to be made. Those include new equations in order to approximate the logarithmic nature of the criteria with linear expressions and the use of Mean Absolute Error values to reward predictive accuracy instead of squared error values. The new additions to the model are presented below:

Indices

i	number of breaking points , $i = 1, 2, ...N$

Variables

AIC	*Akaike Information Criterion* value
BIC	*Bayesian Information Criterion* value
λ_i	SOS2 variables that describe which discrete points will be used for the linear approximation
G	The final result of the approximation

Binary variables

E_r	1 if region r is selected; 0 otherwise

Parameters

γ_i	The discrete points for the linearisation
β_i	The 'output' of the discrete points

Constraints

The following constraint ensures that sample s belongs to region r only if that region is selected. F_s^r is a binary variable that has the value of 1 if sample s belongs to region r. For more information we address the reader to appendix A:

$$F_s^r \leq E_r \qquad \forall\ r, s \tag{5}$$

The following constraint ensures that if region r is not selected, then all of the following regions will not be selected as well:

$$E_{r+1} \leq E_r \qquad \forall\ r = 1, 2, ..., R-1 \tag{6}$$

The next set of equations are responsible for the linear approximation of the logarithm in the AIC and BIC. We introduce λ_i variables which are a SOS2 set (special ordered set of type 2). That means that at most two variables within this ordered set can take on non-zero values. Those two values have to be for adjacent variables in that set. Parameter γ_i is used to discretise the domain of the equation that needs to be linearised. Since there is a mere numerical relation between the two, the following equation is not part of the optimisation model.

$$\beta_i = \ln(\gamma_i) \qquad \forall i$$

The new equations that are introduced to the model are presented below. In these equations another simplification is applied by using absolute error values instead of RSS:

$$\sum_s D_s = \sum_i \gamma_i \cdot \lambda_i \tag{7}$$

$$G = \sum_i \beta_i \cdot \lambda_i \tag{8}$$

$$\sum_i \lambda_i = 1 \tag{9}$$

Equation 7 is known as the 'reference row' and is used to describe the independent variable. In this case, the independent variable that we want to approximate is the sum of absolute errors $\sum_s D_s$. Equation 8 is called the 'function row' and is used to calculate the value of the response that was described in equation 7. Equation 9 ensures that the sum of all the λ_i variables will equal to one. So λ acts as a weight factor that describes which two discrete points have been used for the approximation.

Finally, formulating the objective function depends on the criterion that is chosen in order to perform model selection. For the AIC approach:

$$\min AIC = |S| \cdot G - |S| \ln(|S|) + 2(|M| + 1) \cdot \sum_r E_r \tag{10}$$

Whereas for the BIC approach:

$$\min BIC = |S| \cdot G - |S| \ln(|S|) + \ln |S| \cdot (|M| + 1) \cdot \sum_{r} E_r \tag{11}$$

As stated in section (1), the difference between the two criteria is on the penalty imposed for the complexity of the model. In this approach, the partitioning feature is once again identified by solving the original model proposed in the previous section for all of the features while fixing the number of regions to two. The feature with the smallest prediction error is selected for the partitioning. The next step is to select the maximum number of regions R for the model. The binary variable E_r that is introduced will decide the optimal number of regions that will be selected and constraint (5) will ensure that all of the samples belong to those regions. Overall, the proposed MILP model can be split into two sub-models depending on the objective function. The first model is for *Piecewise Regression with Optimised Akaike information criterion* (***PROA***) and the second is *Piecewise Regression with Optimised Bayesian information criterion* (***PROB***). The two models can be summarised as follows:

minimise objective function 10 (***PROA***) or objective function 11 (***PROB***)

subject to constraints 12 - 18 and 5 - 9

The proposed extended approach solves a single MILP model instead of multiple MILP models, to identify the optimal number of regions. With the new additions the method now deals with overfitting by using two well known and established information criteria as objective function.

Method	**Description**	**Equations**
OPLRA	Original *OPLRA* model (Yang et al., 2016)	min 19, s.t 12-18
PRIA	*OPLRA* model with AIC post-process	min 19, s.t 12-18, post 3
PRIB	*OPLRA* model with BIC post-process	min 19, s.t 12-18, post 4
PROA	Single level MILP with AIC objective function	min 10, s.t. 12-18 and(5 - 9
PROB	Single level MILP with BIC objective function	min 11, s.t. 12-18 and 5 - 9

Table 1: A summary of the of the otpimisation based approaches

Table 1 gives a brief summary of the optimisation based regression approaches. The original model *OPLRA* is used from literature, whereas the rest are the new approaches that are proposed in this work, all of which use information criteria. All of the equations that are part of the models are reported as well. We can see that the *OPLRA* approach has the same optimisation model with *PRIA* and *PRIB*, with the difference in the post processing of the results, whereas *PROA* and *PROB* include new equations to formulate the single MILP approach.

3 Computational part

3.1 Illustrative example

An example from literature is used to demonstrate what the final regression functions of the various methods would look like. This example is about the octane rating of fuel. This specific dataset investigates the octane rating of petrol during a manufacturing process in a refinery. The rating of the fuel is measured as a function of 3 raw materials, named A1, A2 and A3, and a variable that is quantifies the manufacturing conditions of the refinery, named Q (Wood, 1973).

By applying the *PRIA* method that was described in figure 3, we can extract the final regression model. The first step of the process is to identify the partitioning feature that yields the minimum fitting error. Once this feature is known, the method starts to add more regions to the data, until the optimal number is found.

After identifying the correct partitioning feature, instead of applying an iterative approach we can also apply a single level MILP approach, such as *PRIA*, to directly minimise an information criterion and identify the optimal number of regions as well as the position of the break points simply by solving one optimisation model. Table 2 illustrates the difference in the final models from these two proposed approaches.

Prior to performing the analysis, we first apply feature scaling. Each dataset has a number of predictor variables and an output. We perform *feature scaling* with the following equation:

$$\frac{A_{s,m} - \min_s A_{s,m}}{\max_s A_{s,m} - \min_s A_{s,m}}$$

That means that the predictors of the datasets are now within the range of [0,1]. The main advantage of scaling is not having predictors with great numeric ranges that can potentially dominate those in smaller ranges. As a results, all of the break points that will be determined by the model will also be within that same range.

Method	Regression functions
PRIA	$Y = \begin{cases} -5.13 \cdot A_1 + 0.11 \cdot A_2 - 0.71 \cdot A_3 + 1.95 \cdot Q + 95.71, & 0 \leq A_3 \leq 0.58 \\ -7.57 \cdot A_1 - 0.97 \cdot A_2 - 3.70 \cdot A_3 + 3.96 \cdot Q + 98.83, & 0.58 < A_3 \leq 0.71 \\ -8.13 \cdot A_1 - 1.80 \cdot A_2 - 1.82 \cdot A_3 + 2.05 \cdot Q + 99.00, & 0.71 < A_3 \leq 0.92 \\ 23 \cdot A_1 + 2.95 \cdot A_2 + 13.90 \cdot A_3 + 6.48 \cdot Q + 59.15, & 0.71 < A_3 \leq 1 \end{cases}$
PROA	$Y = \begin{cases} -5.13 \cdot A_1 + 0.11 \cdot A_2 - 0.71 \cdot A_3 + 1.95 \cdot Q + 95.71, & 0 \leq A_3 \leq 0.58 \\ -7.79 \cdot A_1 - 1.19 \cdot A_2 - 0.1.07 \cdot A_3 + 3.11 \cdot Q + 97.71, & 0.58 < A_3 \leq 1 \end{cases}$

Table 2: Final regression functions for some of the proposed methods

As we can see, both methods were able to identify exactly the same linear function for the first region. This linear expression is a function of all 4 variables with the corresponding regression coefficients and an intercept. However, the iterative methods identifies a total of 4 regions compared to just 2 of the single level MILP approach.

Even though both models use exactly the same criterion, in this case the AIC, we can see that there are differences in the results. In order to determine which of the proposed methods have good predictive performance more testing has to be done. In the next section a number of examples are used to test and compare the proposed methods to other regression methods.

To test the proposed methods a number of real world datasets have been used. The datasets reported in table 3 are derived from different online sources. More specifically the pharmacokinetics and earthquake data are available through the `datasets` package in R, bodyfat and sensory data are available through StatLib (Vlachos, 2005), distillation data from `OpenMV.net` and the rest from the `UCI` machine learning repository (Dheeru and Karra Taniskidou, 2017).

Dataset	**No.samples**	**No.variables**	**Output variable**
Pharmacokinetics	132	4	Drug concentrarion
Bodyfat	252	14	Bodyfat percentage
Distillation	253	26	Vapour pressure
Yacht Hydrodynamics	308	6	Residuary resistance
Sensory	576	11	Wine score
Cooling efficiency	768	8	Cooling load
Heating efficiency	768	8	Heating load
Earthquake	1000	4	Magnitude
Concrete	1030	8	Compressive strength
White wine quality	4898	11	Quality

Table 3: Regression datasets examined in this work

The datasets that are taken from the `UCI` repository are also used in the original work (Yang et al., 2016). The yacht hydrodynamics set predicts the residuary resistance of sailing yachts for evaluating the ships' performance and for estimating the required propulsive power. The energy efficiency dataset (Tsanas and Xifara, 2012) assesses the heating and cooling load requirements of different buildings as a function of 8 parameters. The concrete dataset (Yeh, 1998) tries to predict the compressive strength of concrete as a structural material. The wine dataset (Cortez et al., 2009) tries to predict the quality of white wine according to some of it's properties.

As mentioned in section (A), the original *OPLRA* model used a heuristic approach to identify the number of regions and introduced a parameter as a stopping criterion. This user specified parameter was set at 0.03 after a sensitivty analysis was performed with the same `UCI` datasets. So in a way, the algorithm was tailored to those datasets. In order to test the accuracy and robustness of the proposed extension, new datasets are introduced in this work.

The pharmacokinetics dataset contains data from a study of the kinetics of the anti-asthmatic drug theophylline. Twelve subjects were given oral doses of the drug and the aim is to predict the final theophylline concentration of each subject by measuring parameters such as weight and time. The earthquake data gives the location of seismic events that occurred near Fiji since 1964. The bodyfat dataset uses features such as age, weight and height to measure the percentage of bodyfat in a subject. The sensory dataset has data for the evaluation of wine quality by a total of 6 judges. The distillation dataset is comprised of measurements from a distillation column with the final result, vapour pressure, being the quality variable that was measured in the lab.

3.2 Validation of the methods

Having developed the algorithm for regression and having resolved the issue of region selection by using two different information criteria, it is now vital to evaluate the method and all of the constructed models. The simplest way to evaluate a model is to split the original data into two subsets, one for training and one for testing. The training set will be used to construct the regression model which will be evaluated by using the testing set. The reason for doing so is to try and measure how well the model generalises to new, previously unseen data (Müller and Guido, 2016).

Cross-validation is a statistical method of evaluating the performance of models that is more thorough and reliable than simply splitting the data into two sets. The most common form of cross-validation is *k-folds* where the data is split into k subsets of equal size. Then the method uses one of these sets for testing and the rest for training. The method stops when all of the k sets have been used as the testing set. Parameter k is user-specified and is usually set to either 5 or 10 (Müller and Guido, 2016).

In this work, 5-fold cross-validation is selected to evaluate the performance of the proposed model. 10 runs will be performed and the Mean Absolute Error (MAE) between model prediction and the true data will be calculated for each fold. The final score is the average of all the runs. All of the proposed mathematical programming models are implemented in the General Algebraic Modeling System (GAMS) (GAMS Development Corporation, 2016) and are solved using the CPLEX solver with optimality gap set at 0 and a time limit of 200s for the iterative approaches and 400s for the single-level MILP approaches. The R programming language (R Development Core Team, 2016) is used for the *k folds* cross-validation procedure. The `caret` package (Kuhn, 2008) that is available in R, contains tools for data splitting, pre-processing, feature selection and more. In this work, the package is used to create random partitions of the samples and perform *k folds* to evaluate the predictive accuracy of the models.

A number of methods from literature are also implemented in this work for comparison purposes on the same datasets. The methods include *KNN* regression (Korhonen and Kangas, 1997), *Random Forest* regression, *MARS* (Friedman, 1991) and *Support Vector* regression (SVR) (Smola and Schölkopf, 2004). All of those methods are implemented in the R programming language using the `FNN` (Beygelzimer et al., 2013), `randomForest` (Liaw and Wiener, 2002), `earth` (Milborrow, 2018) and `e1071` (Meyer et al., 2017) packages respectively. The same 10 runs of 5-fold cross-validation are performed to

evaluate and compare with the proposed methods.

4 Results

	Yacht	**Cooling**	**Heating**	**Concrete**	**Wine**
OPLRA	0.689	**1.275**	**0.805**	4.845	0.551
PRIA	0.680	1.337	0.820	4.840	0.553
PRIB	0.699	1.342	0.909	4.922	0.567
PROA	**0.678**	**1.275**	0.806	4.838	0.555
PROB	0.688	1.351	0.906	4.920	0.566
KNN	5.788	2.237	2.063	8.924	0.577
SVM	3.673	1.820	1.456	4.864	0.518
RandFor	2.454	1.326	0.861	**4.029**	**0.439**
MARS	1.079	1.340	0.826	4.932	0.569
	Bodyfat	**Sensory**	**Distil**	**Pharma**	**Earthquake**
OPLRA	1.273	0.632	2.650	1.613	**7.238**
PRIA	0.785	0.633	1.110	1.352	7.426
PRIB	0.763	0.652	1.127	1.387	7.357
PROA	0.631	0.626	**1.025**	**1.288**	**7.238**
PROB	1.341	0.636	1.105	1.325	7.256
KNN	2.869	0.642	1.966	1.981	8.464
SVM	1.391	0.613	1.128	1.834	7.250
RandFor	1.532	**0.562**	1.153	1.677	7.978
MARS	**0.389**	0.616	1.147	1.420	7.389

Table 4: Cross-validation results using MAE

Table 4 contains the Mean Absolute Error results of all the runs of the 5-fold cross-validation. For comparison purposes, we will first examine how the new methods perform against the previous work (Yang et al., 2016) and seek possible improvements. Next, we will compare the new methods against established methods from literature. For each examined dataset, the regression analysis that performs the best is marked with bold. The *PROA* model consistently performs better than the *OPLRA* model.

More specifically this approach has the lowest average error on 7 out of the 10 examined datasets. It is interesting to note that this approach achieved a better score in all of the new datasets but also for 3 out of the 5 original datasets, even though *OPLRA* was tailored around them.

The *PRIA* is also able to compete against *OPLRA* and manage to achieve a better score for multiple datasets. However, when comparing the methods that use the BIC the results are different. Those methods all have worse performance than the ones that used the AIC. Burnham and Anderson (2003), performed a number of simulations and both of the criteria were used to perform model selection. During those simulations it was discovered that the BIC selected models suffered from underfitting and had

poor performance. However, if the number of samples is very large then the AIC could potentially lead to cases of overfitting. But it is not always clear when a dataset is large since it depends on numerous aspects, the main of which is the process that created the data. So deciding which criterion to use is up to the researcher. In this work, the approaches that use the BIC have worse performance than the equivalent ones with the AIC. So we can assume that the strict penalty that the BIC enforces, can indeed lead to an underfitting model that is not able to properly capture the relationship that exists in the data.

It is very important to test and compare the accuracy of the proposed methods with established methods from literature. Table 4 also compares the proposed methods with the ones from literature. Since the *PROA* model was the best performer amongst the proposed methods, we seek this method to have good performance compared to the established methods. We can see that overall, the proposed *PROA* method has the lowest error in only 5 of the datasets. However, examining the results closer, it is obvious that the method performs well since the error scores are always very close to the ones that have the best overall performance.

	Yacht	**Cooling**	**Heating**	**Concrete**	**Wine**
OPLRA	1.966	4.087	**2.27**	47.371	0.595
PRIA	1.743	4.103	2.45	47.297	0.6
PRIB	1.734	4.203	3.199	47.643	0.611
PROA	**1.457**	3.957	2.273	46.383	0.605
PROB	1.502	4.324	2.62	47.41	0.611
KNN	98.848	10.08	8.939	122.043	0.54
SMV	44.229	7.073	5.101	44.654	0.476
RndFor	15.979	3.405	2.86	**30.104**	**0.366**
MARS	3.209	**3.215**	3.025	39.189	0.524
	Bodyfat	**Sensory**	**Distil**	**Pharma**	**Earthquake**
OPLRA	9.75	0.643	4.127	6.99	**96.594**
PRIA	8.325	0.649	3.22	3.099	116.407
PRIB	7.953	0.684	3.451	3.158	102.501
PROA	7.776	0.634	**2.052**	**2.735**	**96.594**
PROB	11.881	0.66	2.8	2.95	97.312
KNN	12.91	0.649	9.688	6.301	143.349
SMV	6.408	0.602	5.045	6.547	98.891
RndFor	5.071	**0.492**	3.175	4.25	119.577
MARS	**2.001**	0.59	3.305	3.158	99.081

Table 5: Cross-validation results using MSE

Table 5 contains the results of the 10 runs of cross validation but this time instead of MAE the reported values are the Mean Squared Error values (MSE). Regression algorithms such as Random Forests, MARS and KNN minimise the residual sum of squares. Consequently, it might not be fair to compare MAE values for those methods. This is obvious from the results of table 5 where there a noticeable difference in performance,

with Random Forest and MARS providing scores that are more competitive.

To demonstrate this we are going to develop a graph comparing the overall performance of each dataset. In this graph, the method that performed the best gets awarded 10 points, while the one that performed the worst gets 1 point. Everything else is within this range. The final performance score is the average across all of the available datasets.

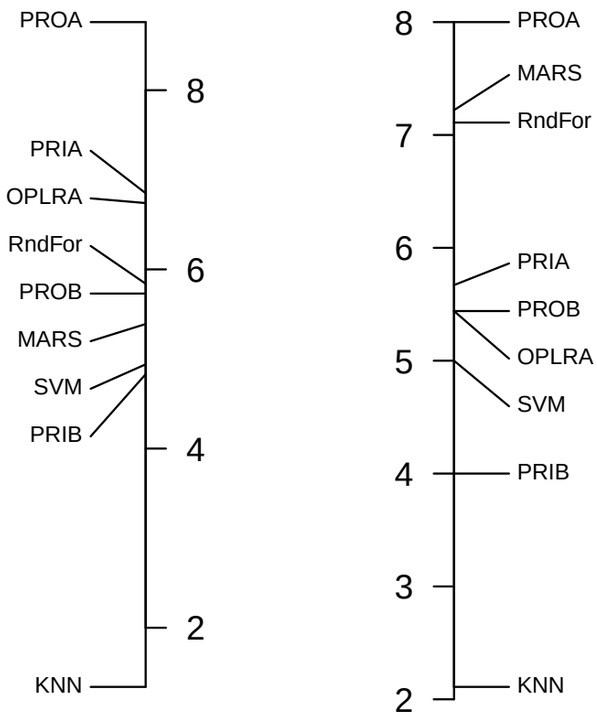

Figure 4: Visualisation of the performance of the methods. The left side is a comparison of MAE values. The right side is a comparison of MSE values.

Looking at figure 4 it is easier to compare the overall performance of the methods. We can see that the *PROA* model has by far the best overall score. This means that the method is consistent when it comes to predictive accuracy and outperforming other methods most of the time. The *OPLRA* and *PRIA* are almost tied while the methods that use the BIC have clearly worse performance that the other proposed methods. In fact, the *PRIB* method is not performing well compared to other established methods such as *SVM*, *Random Forest* and *MARS*. Furthermore, there is a clear difference when using MAE and MSE values. When using MSE as a comparison metric, Random Forest and MARS have very competitive performance scores and are in better than almost all of the proposed methods, except for *PROA*.

To perform statistical analysis we will use the Welch's t-test. This is a two-sample test which is used to test the hypothesis that two populations have equal means and is reliable when the samples have unequal variances (Welch, 1947). For more information we refer the reader to appendix B. If we have evidence to reject this hypothesis using that test, then we can conclude that the difference between the two means is significance.

In this work, the *PROA* method will be compared to the methods from literature. According to tables 4 and 5 and figure 4, the *PROA* method has better overall predictive performance out of all the mathematical programming based methods.

For each dataset, the two different populations that will be compared are the values of the 10 cross validation runs between the *PROA* method and one of the established ones. If by performing the Welch's t-test we have evidence to reject the hypothesis that the means of the two samples are the same, we will be able to conclude that there is a statistical significance between the two sample means, and the best method is the one that has the minimum average error.

Table 6 contains the variance of the 10 runs of cross validation for the *PROA* method and the literature methods. These are necessary to perform the Welch's t-test for unequal variances.

	Yacht	**Cooling**	**Heating**	**Concrete**	**Wine**
PROA	6.76E-04	9.61E-04	1.69E-04	3.36E-03	4.90E-05
KNN	8.84E-03	4.84E-04	6.40E-05	1.44E-03	1.00E-06
SVM	2.92E-03	9.00E-04	1.96E-04	1.68E-03	4.00E-06
RandFor	2.92E-03	3.60E-05	2.25E-04	4.36E-03	4.00E-06
MARS	6.76E-04	1.00E-04	2.25E-04	5.76E-04	1.00E-06
	Bodyfat	**Sensory**	**Distil**	**Pharma**	**Earthquake**
PROA	6.35E-02	1.44E-04	1.04E-01	5.93E-03	1.88E-02
KNN	1.16E-03	9.00E-06	5.76E-04	1.02E-03	3.14E-03
SVM	9.00E-04	4.90E-05	7.84E-04	1.44E-03	3.84E-03
RandFor	1.44E-03	2.50E-05	7.29E-04	1.52E-03	4.36E-03
MARS	1.02E-03	2.50E-05	1.52E-03	1.30E-02	6.08E-03

Table 6: The variances of the 10 runs of cross validation

So by calculating the variances of the cross validation runs and the mean values we can calculate the p-values from a t distribution. For this purpose, the embedded function `t.test()` of R programming language was used. The parameters were set for a two tailed t-test with unequal variances (Welch's test).

Table 7 contains the p-values for comparing the *PROA* methods to the rest of the established methods.

		Yacht	**Cooling**	**Heating**	**Concrete**	**Wine**
PROA	KNN	4.34E-19	1.63E-22	8.99E-29	1.99E-27	3.04E-06
PROA	RndFor	9.75E-20	5.09E-04	7.69E-08	2.09E-16	7.93E-14
PROA	SVM	1.12E-22	5.14E-19	1.28E-26	2.64E-01	1.04E-08
PROA	MARS	6.79E-18	6.11E-05	5.22E-03	4.84E-04	1.24E-04
		Bodyfat	**Sensory**	**Distil**	**Pharma**	**Earthquake**
PROA	KNN	2.70E-10	1.41E-03	6.45E-06	5.47E-12	6.59E-12
PROA	RndFor	9.63E-07	2.45E-09	2.41E-01	1.88E-09	1.05E-09
PROA	SVM	4.65E-06	1.01E-02	3.39E-01	2.71E-11	8.05E-01
PROA	MARS	1.41E-02	3.15E-02	8.72E-01	7.99E-03	8.85E-03

Table 7: P-values associated with the Welch's t-test

By setting a significance level of $\alpha = 0.01$ and comparing that value with the results of table 7, we can identify the examples for which the difference in the mean between *PROA* and the other methods is significant. The condition that needs to be satisfied is $p < \alpha$. To decide which method has better performance, we will examine the MAE values of those methods but only for those datasets that there is a significance between the means.

When comparing *PROA* and KNN, we can see that there is a significant statistical difference in mean for all the examined examples, with *PROA* being the best performer every time. On the other hand, when comparing to random forest there is one example, the distillation dataset, where the difference between the two means is not significant. However, for the rest 9 examples that is not the case with *PROA* being able to outperform random forest in 6 out of those 9 datasets in terms of absolute error.

The performance of SVM and *PROA* seems to be similar in terms of MAE because according to the results of table 7 there is a significant difference only in 6 out of 10 examples. But, overall SVM failed to perform better in terms of MAE for 5 out of those 6 examples. Finally, *PROA* and MARS have a difference in MAE for 7 datasets with *PROA* being the better performer in all of them.

5 Concluding Remarks

In this work, new extensions of the *OPLRA* mathematical programming model are proposed which address the topic of regression analysis and overfitting. In previous work a piecewise linear regression method was proposed (Yang et al., 2016) that partitioned the dataset into multiple regions on the predictor variable that would yield the minimum absolute error and a linear function would be fitted to each one. The resulting model is an MILP formulation that uses a heuristic iterative approach to converge to a solution and select the optimal number of regions. For this iterative approach a user specified parameter, called β, is introduced and used as a stopping criterion.

In an attempt to eliminate this parameter, two different approaches are proposed. These two approaches use the Akaike and Bayesian Information Criteria in an attempt to avoid cases of overfitting that might be caused by the previous heuristic approach. More specifically, the criteria are used as a metric to choose the optimal number of regions for the partitioning of the dataset. This way a balance between a 'good' model fit and model complexity is achieved, improving predictive accuracy. The two approaches tackle the problem of segmented regression in a different way. The first method solves the regression problem iteratively, minimising the absolute error. The AIC and BIC are the stopping criteria that decide when the algorithm has converged. However, the second approach is a single-level MILP formulation that simultaneously solves the problem of regression and partitioning, while directly minimising the values of the criteria.

To test the accuracy of the new approaches and the potential improvements over the original model, a total of 10 real-world datasets have been used. Evaluation of the model is achieved by applying 10 runs of 5-fold cross-validation on all the datasets.

The validation results are also compared to other regression methods from literature. According to the results in table 4 the single-level AIC MILP approach, called *PROA*, is clearly the best performer since it is outperforming other regression methods for half of the examined datasets and achieving a very competitive score for the remaining datasets. This statement is also evident during the Welch's statistical t-test. During this test, it was proven that for most datasets the difference in MAE scores between *PROA* and the other established methods is statistically significant. The BIC methods did not perform as well not only compared to the *OPLRA* model but also compared to other regression methods. The strict penalty that the BIC imposes compared to AIC leads to a small number of selected regions resulting in potential underfitting of the data.

Another key difference between the proposed approaches is the non iterative nature of the formulation. Since the approach can optimally select the number of regions by directly minimising the value of the criteria, it is expected that this approach will have better overall performance than the iterative ones. The difference in computational time is also important since there is only one model that needs to solved instead of solving a model for each region.

Appendix A Mathematical Formulation of OPLRA

In this section, the *OPLRA* mathematical programming model is described as formulated by Yang et al. (2016) in the literature. The model takes a multivariate dataset as input, splits it into multiple segments and fits a linear function to each segment while minimising the mean absolute error of the fitting. All of the indices, parameters and variables that are used in the formulation are explained as follows:

Indices

s	sample, $s = 1, 2, ..., S$
m	feature/independent input variable
r	region, $r = 1, 2, ..., R$
m^*	the feature where sample partitioning takes place

Parameters

A_s^m	numeric value of sample s on feature m
Y_s	output value of sample s
U_1, U_2	suitably large positive numbers
ϵ	very small number

Positive variables

$X_{m^*}^r$	break-point r on partitioning feature m^*
D_s	training error between predicted output and real output for sample s

Variables

W_m^r	regression coefficient for feature m in region r
B^r	intercept of regression function in region r
Pr_s^r	predicted output for sample s in region r

Binary variables

F_s^r	1 if sample s falls into region r; 0 otherwise

Equations and Constraints

Arranging the breaking points in an ordered way:

$$X_m^{r-1} \leq X_m^r \qquad \forall\ m = m^*,\ r = 2, 3, ..., R-1 \tag{12}$$

For a breaking point X_m^r to exist, at least two regions must be selected. The number of breaking points will always be one less than the number of regions.

In order to assign samples into the correct regions, binary variables are introduced to the model.

$$X_m^{r-1} - U_1 \cdot (1 - F_s^r) + \epsilon \leq A_s^m \qquad \forall\ s,\ r = 2, 3, ..., R,\ m = m^* \tag{13}$$

$$A_s^m \leq X_m^r + U_1 \cdot (1 - F_s^r) - \epsilon \qquad \forall\ s,\ r = 1, 2, ..., R-1,\ m = m^* \tag{14}$$

Parameter ϵ is added to the model to make sure that no values of the dataset will equal any of the breaking points.

Each sample of the dataset can be assigned to only one region.

$$\sum_r F_s^r = 1 \qquad \forall\ s \tag{15}$$

Variable Pr_s^r is the predicted response of the model.

$$Pr_s^r = \sum_m A_s^m \cdot W_m^r + B^r \qquad \forall\ s, r \tag{16}$$

The following two equations are used to formulate the absolute deviation between the real output and the predicted output of the model.

$$D_s \geq Y_s - Pr_s^r - U_2 \cdot (1 - F_s^r) \qquad \forall s, r \tag{17}$$

$$D_s \geq Pr_s^r - Y_s - U_2 \cdot (1 - F_s^r) \qquad \forall s, r \tag{18}$$

The objective function of the model is the minimisation of the absolute deviation error:

$$\min \sum_s D_s \tag{19}$$

The resulting model is formulated as an MILP problem that can be solved to global optimality.

In this literature work, Yang et al. (2016) proposed a heuristic procedure in order to identify the partitioning predictor variable and find the optimal number of regions. This was achieved by using an iterative approach and introducing a new parameter β, which was used as a threshold to the reduction percentage of the absolute error. If the reduction percentage of the error was above that parameter, then a new region was added and the model would solve again. The entire the process stops once convergence has been achieved.

Appendix B Welch's t-test

The t-test is formulated as (Ruxton, 2006):

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{20}$$

where

$\bar{X}_1, \bar{X}_2$	the mean of the 1^{st} and 2^{nd} sample respectively
s_1^2, s_2^2	the variance of the 1^{st} and 2^{nd} sample respectively
N_1, N_2	the size of the 1^{st} and 2^{nd} sample respectively

The degrees of freedom associated with this variance estimate is approximated as (Ruxton, 2006):

$$\nu \approx \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2 \cdot \nu_1} + \frac{s_2^4}{N_2^2 \cdot \nu_2}} \tag{21}$$

where

$v_1 = N_1 - 1$	the degrees of freedom associated with the 1^{st} variance
$v_2 = N_2 - 1$	the degrees of freedom associated with the 2^{nd} variance

Once the t-statistic and the degrees of freedom have been computed, the t distibution can be used to test the null hypothesis using a two-tailed test.

Acknowledgment

Funding from the UK Leverhulme Trust under grant number RPG-2015-240 is gratefully acknowledged. I would also like to express my gratitude to Marco Quaglio for his helpful suggestions on the statistical analysis part.

References

Beygelzimer, A., Kakadet, S., and Langford, J. (2013). Package FNN. *Available at* https://cran.r-project.org/web/packages/FNN/FNN.pdf.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Burnham, K. P. and Anderson, D. R. (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47:547–553.

Cozad, A., Sahinidis, N. V., and Miller, D. C. (2014). Learning surrogate models for simulation-based optimization. *AIChE Journal*, 60:2211–2227.

Dheeru, D. and Karra Taniskidou, E. (2017). UCI machine learning repository. [http://archive.ics.uci.edu/ml]. University of California, Irvine, School of Information and Computer Sciences.

Fabozzi, F. J., Focardi, S. M., Rachev, S. T., and Arshanapalli, B. G. (2014). *The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications*. John Wiley & Sons.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19:1–67.

GAMS Development Corporation (2016). General Algebraic Modeling System (GAMS) Release 24.7.1, Washington, DC, USA.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2 edition.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44:1–12.

International Prognostic Factors Study Group (2010). Prognostic factors in patients with metastatic germ cell tumors who experienced treatment failure with cisplatin-based first-line chemotherapy. *Journal of Clinical Oncology*, 28:4906–4911.

Kimura, K. and Waki, H. (2018). Minimization of akaike's information criterion in linear regression analysis via mixed integer nonlinear program. *Optimization Methods and Software*, 33:633–649.

Korhonen, K. T. and Kangas, A. (1997). Application of nearest-neighbour regression for generalizing sample tree information. *Scandinavian Journal of Forest Research*, 12:97–101.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28:1–26.

Lehmann, R. and Lösler, M. (2016). Multiple outlier detection: Hypothesis tests versus model selection by information criteria. *Journal of Surveying Engineering*, 142:04016017.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2:18–22.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Friedrich, L. (2017). Package e1071. *Available at* https://cran.r-project.org/web/packages/e1071/e1071.pdf.

Milborrow, S. (2018). Package earth. *Available at* https://cran.r-project.org/web/packages/earth/earth.pdf.

Müller, A. C. and Guido, S. (2016). *Introduction to Machine Learning with Python: A guide for data scientists*. O' Reilly Media, Inc.

Muggeo, V. M. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, 22:3055–3071.

Muggeo, V. M. R. (2008). Segmented:An R Package to Fit Regression Models with Broken-Line Relationships. *R news*, 8:20–25.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, 17:688–690.

Sato, T., Takano, Y., Miyashiro, R., and Yoshise, A. (2016). Feature subset selection for logistic regression via mixed integer optimization. *Computational Optimization and Applications*, 64:865–880.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14:199–222.

Snipes, M. and Taylor, D. C. (2014). Model selection and akaike information criteria: An example from wine ratings and prices. *Wine Economics and Policy*, 3:3–9.

Tsanas, A. and Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567.

Vlachos, P. (2005). StatLib-statistical datasets. *Available at* http://lib.stat.cmu.edu/datasets/.

Wagenmakers, E.-J. and Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11:192–196.

Welch, B. L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34:28–35.

Wilson, Z. T. and Sahinidis, N. V. (2017). The alamo approach to machine learning. *Computers & Chemical Engineering*, 106:785–795.

Wood, F. S. (1973). The use of Individual Effects and Residuals in Fitting Equations to Data. *Technometrics*, 15:677–695.

Yang, L., Liu, S., Tsoka, S., and Papageorgiou, L., G. (2016). Mathematical Programming for Piecewise Linear Regression Analysis. *Expert Systems with Applications*, 44:156–167.

Yeh, I. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28:1797 – 1808.