# Regularized Linear Regression (Ridge, Lasso, Elastic, Early Stopping)

Chiagoziem Cyriacus Ugoh
*Electronics Engineering Department (of Hamm-Lippstadt University of Applied Sciences)*

Lippstadt, Germany
chidexmailbox@gmail.com

***Abstract*—Our societies are full of checks and balances. In our political systems, parties balance each other (in theory) to find solutions that are at neither extreme of each other's views. Professional areas, such as financial services, have regulatory bodies to prevent them from doing wrong and ensure that the things they say and do are truthful and correct. When it comes to machine learning, it turns out we can apply our own form of regulation to the learning process to prevent the algorithms from over-fitting the training set. We call this regulation in machine learning regularization [1].**

***Index Terms*—**

## I. Introduction

Regularization (also sometimes called shrinkage) is a technique that prevents the parameters of a model from becoming too large and "shrinks" them toward 0. The result of regularization is models that, when making predictions on new data, have less variance [2][3].

## II. Linear Regression

Linear regression is the simplest and most widely used statistical technique for predictive modeling. It basically gives us an equation, where we have our features as independent variables, on which our target variable is dependent upon [4]. So what does the equation look like? Linear regression equation looks like this:

Here, we have Y as our dependent variable, X's are the independent variables and all thetas are the coefficients. Coefficients are basically the weights assigned to the features, based on their importance. For example, if we believe that sales of an item would have higher dependency upon the type of location as compared to size of store, it means that sales in a tier 1 city would be more even if it is a smaller outlet than a tier 3 city in a bigger outlet. Therefore, coefficient of location type would be more than that of store size [5][6][7].

Identify applicable funding agency here. If none, delete this.

$$Y = \theta_1 X_1 + \theta_2 X_2 + \dots \theta_n X_n$$

Fig. 1. Linear Regression Equation

## III. Regularization Technique

### A. Regularization

While we can apply regularization to most machine learning problems, it is most commonly used in linear modeling, where it shrinks the slope parameter of each predictor toward 0. Three particularly well-known and commonly used regularization techniques for linear models are as follows[8][9][10][11][12]:

. Ridge regression
. Least absolute shrinkage and selection operator (LASSO)
. Elastic net

These three techniques can be thought of as extensions to linear models that reduce over-fitting. Because they shrink model parameters toward 0, they can also automatically perform feature selection by forcing predictors with little information to have no or negligible impact on predictions.

### B. Ridge regression

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values[13][14]. The cost function for ridge regression:

$$Min(||Y - X(theta)||^2 + ||theta||^2) \quad (1)$$

Lambda is the penalty term. given here is denoted by an alpha parameter in the ridge function. So, by changing the values of alpha, we are controlling the penalty term. The higher the values of alpha, the bigger is the penalty and therefore the magnitude of coefficients is reduced[15][16].

It shrinks the parameters. Therefore, it is used to prevent multicollinearity It reduces the model complexity by coefficient shrinkage

Ridge Regression Models:

For any type of regression machine learning model, the usual regression equation forms the base which is written as [17]:

$$Y = XB + e \tag{2}$$

Where Y is the dependent variable, X represents the independent variables, B is the regression coefficients to be estimated, and e represents the errors are residuals.

Once we add the lambda function to this equation, the variance that is not evaluated by the general model is considered. After the data is ready and identified to be part of L2 regularization, there are steps that one can undertake[18].

### C. Least absolute shrinkage and selection operator (LASSO)

Machine learning models tend to overfit the data they are trained in leading to high variance. This problem of overfitting can be dealt with the process of regularization, reducing the complexity of the function. LASSO or Least Absolute Shrinkage and Selection Operator uses l1 regularization does variable selection(choosing the Independent variable arbitrarily)and regularization[19]. It is very similar to how ridge regression works, which uses the sum of the square of weights to constrain the regression model, while Lasso uses the absolute sum of weights. This is basically the residual sum of squares added with lambda times the absolute sum of weight. Multiple linear regression analysis has many independent variables, so there is a correlation between two or more independent variables. This independent variable that correlates with each other is called multicollinearity [20]. The LASSO (Least Absolute Shrinkage and Selection Operator) method regression can help to reduce multicollinearity and increase the accuracy of linear regression models [21]. LASSO is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produce [20]. The lasso is the best-studied, most basic, shrinkage operator technique [22]. LASSO shrinks the coefficients (parameters) which correlate to zero or close to zero [22], resulting in estimators with smaller variants and a more representative final model [22]. The Lasso method became known after the LAR algorithm in 2004. The solution paths of LAR are piecewise linear and thus can be computed very efficiently [23]. Modification of LAR (Least Angle Regression) for LASSO produces a more efficient algorithm for estimating the LASSO coefficient estimator solution. The LASSO method can shrink the ordinary least squares method coefficient to zero so that it can select the fixed variable. the model produced by the LASSO method is simpler and indirectly free from multicollinearity [25].

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|),$$

Fig. 2. Elastic Net Mathematical Model

### D. Elastic net

The main purpose of Elastic Net Regression is to find the coefficients that minimize the sum of error squares by applying a penalty to these coefficients. Elastic Net combines L1 and L2 (Lasso and Ridge) approaches[2]. As a result, it performs a more efficient smoothing process. In another source, it is defined as follows: Elastic Net first emerged as a result of critique on Lasso, whose variable selection can be too dependent on data and thus unstable. The solution is to combine the penalties of Ridge regression and Lasso to get the best of both worlds[14].

Features of Elastic Net Regression It combines the L1 and L2 approaches. It performs a more efficient regularization process. It has two parameters to be set,  and .

The elastic net method improves on lasso's limitations, i.e., where lasso takes a few samples for high dimensional data, the elastic net procedure provides the inclusion of "n" number of variables until saturation. In a case where the variables are highly correlated groups, lasso tends to choose one variable from such groups and ignore the rest entirely[23][24].

Elastic Net aims at minimizing the following loss function:

### E. Early Stopping

Early stopping is a form of regularization based on choosing when to stop running an iterative algorithm. Focusing on non-parametric regression in a reproducing kernel Hilbert space, we analyze the early stopping strategy for a form of gradient-descent applied to the least-squares loss function[26][27]. In a general learning algorithm, the dataset is divided as a training set and test set. After each epoch of the algorithm, the parameters are updated accordingly after understanding the dataset. Finally, this trained model is applied to the test set. Generally, the training set error will be less compared to the test set error. This is because of overfitting whereby the algorithm memorizes the training data and produces the right results on the training set. So the model becomes highly exclusive to the training set and fails to produce accurate results for other datasets including the test set. Regularization techniques are used in such situations to reduce overfitting and increase the performance of the model on any general dataset. Early stopping is a popular regularization technique due to its simplicity and effectiveness[28][29][30].

Regularization by early stopping can be done either by dividing the dataset into training and test sets and then using cross-validation on the training set or by dividing the dataset into training, validation and test sets, in which case cross-validation is not required. Here, the second case is analyzed[15]. In early stopping, the algorithm is trained using the training set and the point at which to stop training is determined from the validation set. Training error and validation

error are analysed. The training error steadily decreases while validation error decreases until a point, after which it increases. This is because, during training, the learning model starts to overfit to the training data. This causes the training error to decrease while the validation error increases. So a model with better validation set error can be obtained if the parameters that give the least validation set error are used. Each time the error on the validation set decreases, a copy of the model parameters is stored[19]. When the training algorithm terminates, these parameters which give the least validation set error are finally returned and not the last modified parameters.

In Regularization by Early Stopping, we stop training the model when the performance of the model on the validation set is getting worse-increasing loss or decreasing accuracy or poorer values of the scoring metric. By plotting the error on the training dataset and the validation dataset together, both the errors decrease with a number of iterations until the point where the model starts to overfit[22]. After this point, the training error still decreases but the validation error increases. So, even if training is continued after this point, early stopping essentially returns the set of parameters which were used at this point and so is equivalent to stopping training at that point. So, the final parameters returned will enable the model to have low variance and better generalization. The model at the time the training is stopped will have a better generalization performance than the model with the least training error[28]. Early stopping can be thought of as implicit regularization, contrary to regularization via weight decay. This method is also efficient since it requires less amount of training data, which is not always available. Due to this fact, early stopping requires lesser time for training compared to other regularization methods. Repeating the early stopping process many times may result in the model overfitting the validation dataset, just as similar as overfitting occurs in the case of training data[12].

The number of iterations taken to train the model can be considered as a hyperparameter. Then the model has to find an optimum value for this hyperparameter(by hyperparameter tuning)for the best performance of the learning model[14].

### *F. Linear Regression Loss Function*

There are different ways of evaluating the errors. For example, if you predicted that a student's GPA is 3.0, but the student actual GPA is 1.0, the difference between the actual and predicted GPAs is 1.0 – 3.0 = -2.0. However, there can't be a negative distance, can it be? So what can we do?

Well, you can either take the absolute difference, which is just 2.0. Alternatively, you can take the squared difference , which is 2.0(squared)=4.0. If you can't decide which one to use, you can add them together, it is not the end of the world, so it will be 1.0 + 4.0 = 5.0. Well, each of these distance calculation techniques (aka distance metrics) result in a differently behaving linear regression model. To escape the ambiguity about the distance between the actual and the predicted value, we use the term residual, which refers to the error, regardless of how it is calculated. The function we want to normalize when we are fitting a linear regression model is called the loss function, which is the sum of all the squared residuals on the training data, formally called Residual Sum of Squares (RSS)[8][9][10].

### *G. Conclusion*

Managerial decision making, organizational efficiency, and revenue generation are all areas that can be improved through the utilization of data-based insights. Currently, these insights are being more readily sought out as technological accessibility stretches further and competitive advantages in the market are harder to acquire. One field that seeks to realize value within collected data samples is predictive analytics[2]. By leveraging mathematical/statistical techniques and programming, practitioners are able to identify patterns within data allowing for the generation of valuable insights. Regression is one technique within predictive analytics that is used to predict the value of a continuous response variable given one or many related feature variables [6]. Algorithms of this class accomplish this task by learning the relationships between the input (feature) variables and the output (response) variable through training on a sample dataset. How these relationships are learned, and furthermore used for prediction varies from algorithm to algorithm. The practitioner is faced with options for regression modeling algorithms, however, linear regression models tend to be explored early on in the process due to their ease of application and high explainability[13].

## REFERENCES


[1] Leo Breiman. Heuristics of instability and stabilization in model selection. The Annals of Statistics, 24(6), 1996.

[2] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. The Annals of Statistics, 32:407–499, 2004.

[3] Bradley Efron and Robert Tibshirani. An Introduction to the Bootstrap. Chapman and Hall, London, 1998.

[4] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. glmnet: Lasso and elastic- net regularized generalized linear models, 2009. R package version 1.1-4, http: //CRAN.R-project.org/package=glmnet.

[5] Ludwig Fahrmeir, Thomas Kneib, and Stefan Lang. Regression: Modelle, Methoden und Anwendungen. Springer, Berlin, 1. edition, 2007.

[6] Jianqing Fan and Runze Li. Variable selection via non-concave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96:1348– 1360, 2001.

[7] Ludwig Fahrmeir and Gerhard Tutz. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer, New York, 2. edition, 2001. 25, 26, 28, 56

[8] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12:55–67, 1970. 8, 9

[9] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, 2. edition, 2009. 13, 14

[10] Justin Lokhorst, Bill Venables, Berwin Turlach, and Martin Maechler. lasso2: L1 constrained estimation aka 'lasso', 2009. R package version 1.2-10, http: //www.maths.uwa.edu.au/ berwin/software/lasso.html. 39, 74

[11] Olivier Ledoit and Michael Wolf. Honey, I shrunk the sample covariance matrix. Journal of Portfolio Management, 30(4):110–119, 2004. 19

[12] Rainer Opgen-Rhein and Korbinian Strimmer. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. Statistical Applications in Genetics and Molecular Biology, 6(9), 2007. 19

[13] Sebastian Petry. Shrinkage regression with polytopes, 2009. http://www. statistik.lmu.de/ petry/Hoehenried.pdf. 5, 16, 20, 22

[14] Mee Young Park and Trevor Hastie. glmpath: L1 Regularization path for generalized linear models and cox proportional hazards model, 2007. R package version 0.94, http://CRAN.R-project.org/package=glmpath. 74

[15] R Development Core Team. R: A Language and Environment for Statistical Com- puting. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0, http://www.R-project.org. 36

[16] Florian Reithinger. Zusammenhangsstrukturen, 2006. Vorlesungsskript, Mul- tivariate Verfahren SS06, http://www.statistik.lmu.de/ flo/ss06/ strukturen.pdf. 18

[17] Juliane Scha fer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology, 4(32), 2005. 19

[18] Helge Toutenburg. Lineare Modelle: Theorie und Anwendungen. Physica-Verlag, Heidelberg, 2. edition, 2003. 8

[19] Jan Ulbricht. Variable selection in generalized linear models. Dissertation, Ludwig- Maximilians-Universita t, Mu nchen, 2010. 5, 29, 30, 32, 33, 73

[20] Sanford Weisberg. Applied Linear Regression. Wiley, New York, 2. edition, 1985. 13, 16

[21] Joe Whittaker. Graphical Models in Applied Multivariate Statistics. John Wiley, Chichester, 1990. 18

[22] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society B, 67:301–320, 2005. 5, 11, 20, 35, 73

[23] Hui Zou and Trevor Hastie. elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA, 2008. R package version 1.0-5, http://www.stat.umn.edu/ hzou. 39

[24] Hui Zou. The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101(476):1418–1429, 2006. 17

[25] Zhang, Long, and Kang Li. 2015. "Automatica Forward and Backward Least Angle Regression for Nonlinear Zhou, Xiao-ping, and Xiao-cheng Huang. 2018. "Reliability Analysis of Slopes Using UD-Based Response Surface Methods Combined with LASSO." Engineering Geology 233(June 2017): 111–23.System." Automatica 53: 94– 102.

[26] Chen, Hongmei, and Yaoxin Xiang. 2017. "ScienceDirect The Study of Credit Scoring Model Based on Group Lasso." Procedia Computer Science 122: 677–84.3

[27] Dyar, M D et al. 2012. "Spectrochimica Acta Part B Comparison of Partial Least Squares and Lasso Regression Techniques as Applied to Laser-Induced Breakdown Spectroscopy of Geological Samples." Spectrochimica Acta Part B: Atomic Spectroscopy 70: 51–67

[28] Gauthier, Philippe-aubert, William Scullion, and Alain Berry. 2017. "Sound Quality Prediction Based on Systematic Metric Selection and Shrinkage : Comparison of Stepwise , Lasso , and Elastic-Net Algorithms and Clustering Preprocessing." Journal of Sound and Vibration 400: 134–53.

[29] A Mixed Integer Programming Approach." EXPERT SYSTEMS WITH APPLICATIONS 42(1): 325–31. Permai, Syarifah Diana, and Heruna Tanty. 2018. "ScienceDirect ScienceDirect Linear Regression Model Using Bayesian Approach for Energy Linear Regression Model Using Bayesian Approach for Energy Performance of Residential Building Performance of Residential Building." Procedia Computer Science 135: 671–77.

[30] Kazemi, A, A Mohamed, H Shareef, and H Zayandehroodi. 2013. "Electrical Power and Energy Systems Optimal Power Quality Monitor Placement Using Genetic Algorithm and Mallow ' s Cp." International Journal of Electrical Power and Energy Systems 53: 564–75.