Assignment-1: A Mathematical essay on linear regression

Ahmed Shmels Muhe
M.Tech In Data Science
Indian Institute of Technology Madras, IIT Madras
Chennai, India
ge22m009@smail.iitm.ac.in

***Abstract*—In this revised (V2) mathematical essay on linear regression, we look at how low income affects cancer diagnosis and treatment in the United States. We show that cancer incidence and mortality are related to socioeconomic status and provide quantitative and visual evidence. In our problem, we attempt to forecast the average death based on changes in multiple parameters. How do these variables, such as the number of people with medical insurance, affect the average death rate? What about the people's poverty situation? We discovered that most of the features have different ranges of numbers, which affects the results of our models, so we used some feature scaling techniques to avoid other fields of numbers that can cause the model to under-fit the data. To solve this, we go through various stages of work, such as data cleaning, processing, exploratory analysis, and modelling. In this revised version, I include the mathematics behind linear regression in Section 2. In Section 4, I tested the algorithm with 3 different regularisations (Lasso, ridge, and ElasticNet) and updated the writing and some of the outputs after some modifications.**

***Index Terms*—Linear Regression,under-fit, exploratory analysis.**

I. Introduction

Cancer survival rates differ according to socioeconomic status. In both developed and developing countries, lower socioeconomic status (SES) is associated with higher cancer incidence and poorer cancer survival than higher SES [1]. To investigate these correlations further, we used linear regression models in this study.

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data [3]. One or more variables are considered explanatory variables, and one is regarded as a dependent variable. The most common method for fitting a regression line is the method of least-squares [2]. This method calculates the best-fitting line for the observed data by minimising the sum of the squares of the vertical deviations from each data point to the bar.

In this study, we used linear regression to explore the correlation between cancer incidence, mortality rates, and socioeconomic and racial factors. We gather, clean, and prepare the data, performing exploratory analysis. Finally, we build statistical models and perform visualisations to provide quantitative and visual evidence of the relationships observed. In the next section, we highlight the fundamental principles underlying linear regression. Section 3 discusses the insights and observations drawn from the data and the models. Finally, in section 4, we outline the salient features of the study and present further avenues of possible investigation.

II. Linear Regression

Linear regression is a statistical method for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). Simple linear regression is used when there is only one explanatory variable; multiple linear regression is used when there is more than one. This term differs from multivariate linear regression, which predicts various correlated dependent variables rather than a single scalar variable. Linear regression models relationships using linear predictor functions, the unknown model parameters estimated from data. These are known as linear models. The conditional mean of the response given the explanatory variables (or predictors') values is commonly assumed to be an affine function of those values.

Linear regression has a wide range of applications. The majority of applications fall into one of two broad categories:

- Linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables
- Linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables

In this section let's see the mathematics behind liner regression, linear regression is the model's response is approximated as a linear function of the inputs:

$$y(x) = w^T x + \epsilon = \sum_{i=1}^{D} w_i s_i + \epsilon \tag{1}$$

where $w^T x$ represents the inner or scalar product between the input vector x and the model's weight vector w and ϵ is the residual error between our linear predictions and the true response.

We often assume that it has a Gaussian distribution. We denote this by $N_\epsilon(\mu, \sigma^2)$, where μ is the mean, and σ^2 is the variance. Mathematically, we can say:

$$p(y|x, \theta) = N(y|\mu(x), \sigma^2(x)) \tag{2}$$

In the simplest case, we assume μ is a linear function of x, so $\mu = w^T x$, and that the noise is fixed, σ^2 (x) = σ^2. In this case, θ = (w, σ^2) are the parameters of the model.

$$\mu(x) = w_0 + w_1 x = w^T x \tag{3}$$

where w_0 is the bias term, w_1 is the slope, and where we have defined the vector $x = (1, x)$. If W_1 is positive, it means we expect the output to increase as the input increases.

Linear regression can be made to model non-linear relationships by replacing x with some non-linear function of the inputs, $\phi(x)$. That is, we use

$$p(y|x, \theta) = N(y|w^T \phi(x), \sigma^2) \tag{4}$$

A. Maximum likelihood estimation

A common way to estimate the parameters of a statistical n model is to compute the MLE, which is defined as

$$\theta = argmax_\theta \log p(D|\theta) \tag{5}$$

It is common to assume the training examples are independent and identically distributed. This means we can write the log-likelihood as follows:

$$l(\theta) = \log p(D|\theta) = \sum_i = 1^N \log p(y_i|x_i, \theta) \tag{6}$$

Instead of maximizing the log-likelihood, we can equivalently minimize the negative log-likelihood:

$$NLL(\theta) = \sum_{i=1}^{N} \log p(y_i|x_i, \theta) \tag{7}$$

Now, if we apply the method of MLE to the linear regression and insert the definition of the Gaussian into the above, we find that the log-likelihood is given by

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{N} \log[(\frac{1}{2\pi\sigma^2})^{\frac{1}{2}} \exp^{(-1\frac{1}{2\sigma^2}(y_i - w^T x_i)^2)}] \\ &= -\frac{1}{2\sigma^2} RSS(w) - \frac{N}{2} \log(2\pi\sigma^2 \end{aligned} \tag{8}$$

where RSS stands for the residual sum of squares and is defined by

$$RSS(w) = \sum_{i=1}^{N} (y_i - w^T x_i)^2 \tag{9}$$

It can also be written as the square of the l2 norm of the vector of residual errors:

$$RSS(w) = ||\epsilon||_2^2 = \sum_{i=1}^{N} \epsilon_i^2 \tag{10}$$

where $\epsilon_i = (y_i - w^T x_i)$.

B. Derivation of the MLE

First, we rewrite the objective in a form that is more amenable to differentiation:

$$\begin{aligned} NLL(w) &= \frac{1}{2}(y - Xw)^T (y - Xw) \\ &= \frac{1}{2} w^T (X^T X) w - w^T (X^T y) \end{aligned} \tag{11}$$

where

$$X^T X = \sum_{i=1}^{N} x_i x_i^T = \sum_{i=1}^{N} \begin{pmatrix} x_i^2, 1 & \cdots & x_i, 1x_i, D \\ & \cdots & \\ & \cdots & \\ & \cdots & \\ x_i, Dx_i, 1 & \cdots & x_i^2, D \end{pmatrix}$$

is the sum of the squares matrix and

$$X^T Y = \sum_{i=1}^{N} x_i y_i.$$

Using results from the above equation, we see that the gradient of this is given by

$$g(w) = [X^T X w - X^T y] = \sum_{i=1}^{N} x_i (w^T x_i - y_i)$$

equating to zero we get,

$$X^T X w = X^T y \tag{12}$$

This is known as the normal equation. The corresponding solution wˆ to this linear system of equations is called the ordinary least squares or OLS solution, which is given by

$$wOLS == (X^T X)^{-1} X^T y$$

III. THE PROBLEM FORMULATION

The goal of this study is to determine whether low-income groups are more likely to be diagnosed with and die from cancer in order to assist a nonprofit with lobbying and fundraising. First, we show whether cancer incidence and mortality are related to socioeconomic status, and then we provide both quantitative and visual evidence that the nonprofit can use to further its mission.

A. Gather, Clean & Prepare data

The dataset used for this problem includes cancer incidence and mortality rates in various areas of the United States, as well as the median income divided by ethnicity in each area. The data also includes the number of people who have health insurance and the cancer incidence trend over the last five years.

1) Existing features in the data set: The merged dataset used in this study contains 25 columns and 3134 samples from various states. The following is an interpretation of the features:

1) All Poverty: Number of people of both genders below the poverty line. Similarly, M poverty is for males, and F Poverty is for females.
2) Area: Name of the area in which sampling is done.
3) State: State of the respective area, totally 51 in number.
4) FIPS: Zipcode of the area.
5) Med Income: Median Income of all ethnic groups in the area.
6) Med income White, Black, etc: Median Income of a particular ethnic group in the area.
7) All With Number of individuals having insurance in the area; Along these lines, All Without, Without Male, With Male, Without Female, With Female is defined.
8) Incidence Rate: Number of cancer cases detected per 100,000 people in the area.
9) Mortality Rate: Number of mortalities per 100,000 people.

To impute the missing values, two approaches are considered:

1) State-wise medians are filled in in place of the empty values such that the distribution does not get distorted, as would be the case if we imputed using means
2) Missing Values are dropped.

Multiple values in the columns contained noisy characters like '**' and '*'. These were replaced with missing values and '3 or fewer' was replaced with 3. Some numbers were preceded by a '#' symbol, which was removed. These points are not considered in the case of '*' because we do not know how many cases it is reliably (less than 16).

Finally, columns 'FIPS', 'fips_x', and 'fips_y' were dropped assuming Pincode holds no relation with incidence or mortality rates.

2) Additional features::

1) Total Population: The total population of the area is obtained by summing the number of males and females
2) Female Ratio: Number of females divided by the total number of people, similarly we can obtain the ratio of males in the area.
3) Female Poverty Ratio: Ratio of poor females among a total number of females, similarly Male Poverty Ratio can be obtained.
4) Female Insurance Ratio: Number of females insured divided by the total number of females, similarly, we can obtain the male insurance ratio.

```
Shape of dataframe  (3134, 25)
State                0.000000
AreaName             0.000000
All_Poverty          0.000000
M_Poverty            0.000000
F_Poverty            0.000000
FIPS                 0.000000
Med_Income           0.031908
Med_Income_White     0.063816
Med_Income_Black    38.608807
Med_Income_Nat_Am   52.967454
Med_Income_Asian    56.062540
Hispanic            21.729419
M_With               0.000000
M_Without            0.000000
F_With               0.000000
F_Without            0.000000
All_With             0.000000
All_Without          0.000000
fips_x               0.000000
Incidence_Rate      13.401404
Avg_Ann_Incidence    6.668794
recent_trend        14.901085
fips_y               0.000000
Mortality_Rate      10.370134
Avg_Ann_Deaths      10.370134
dtype: float64
```

Fig. 1. Percentage of missing values in each of the columns

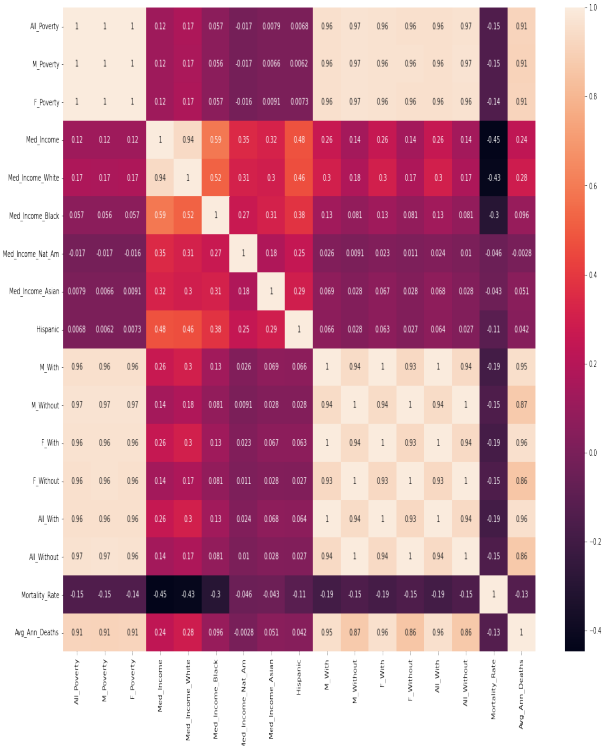

Fig. 2. Correlation Coefficient among all numeric features

IV. Exploratory, Analysis & Visualizations

In Figure 1, we see the correlation matrix forming clusters of features, indicating that groups of features are highly linearly related to others in that group. From this, we can already see that the median income and mortality rate have a negative moderate correlation coefficient, indicating that as the median income increases, mortality rates decrease.

In Figure 2, we plot the median income distributions for each of the ethnic groups. We observe that blacks have the

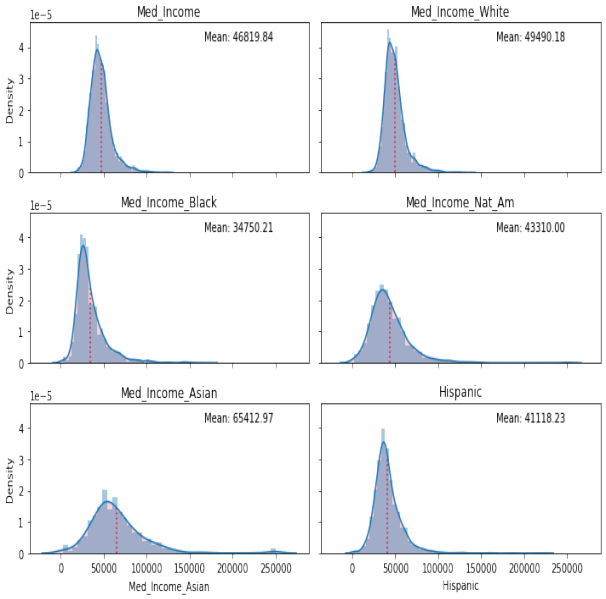

Fig. 3. Distribution of Median Incomes in all areas (a) all ethnic groups (b) Whites (c) Blacks (d) Native Americans (e) Asians (f) Hispanics

lowest mean median income' while Asians have the highest mean median income'. Further, native Americans and Asians have a more spread out distribution with higher variance, while whites, blacks, and Hispanics have lower variance in income, indicating that most blacks and Hispanics are paid low incomes, hinting at racial discrimination.

In Figure 3, we plot the mean mortality rate against the mean incidence rate of each of the 51 states. There is strong evidence of a linear relationship, confirming that as the rate of incidence increases, mortality rates also increase.

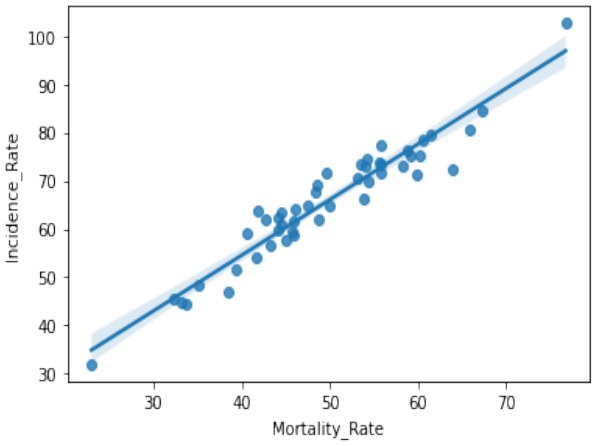

Fig. 4. Mean Incidence Rate vs Mean Mortality Rates for the 51 states

In Figure 4, it is seen that among all ethnic groups, subgroups with higher median income tend to show a falling trend in cases. This plot further shows us that among all ethnic groups, blacks have the lowest median income, even among the areas with falling cases.

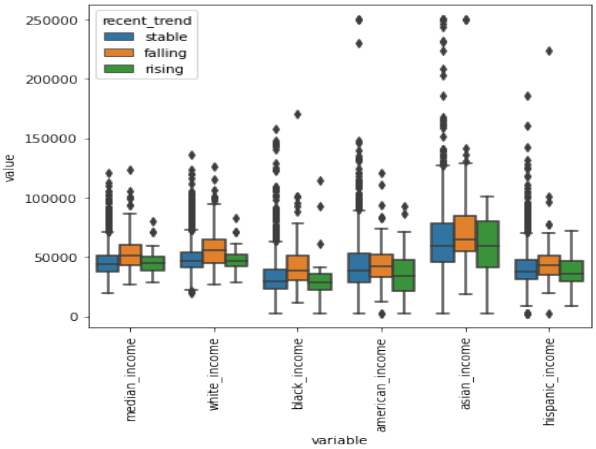

Fig. 5. Plot of the median income of various ethnic groups in regions with stable, falling and rising trend in the number of cases

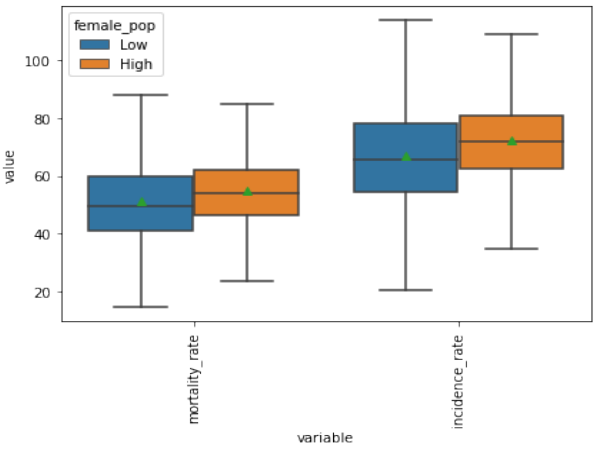

Fig. 6. Plot of Mortality and Incidence Rates for regions with low and high female population

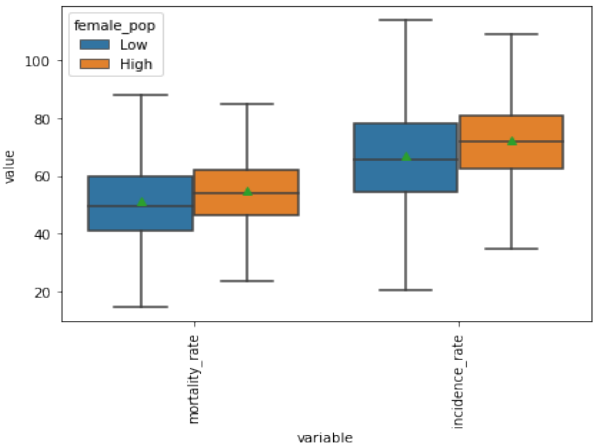

Fig. 7. Plot of Mortality and Incidence Rates for regions with low and high female population

In Figure 5, it is clearly visible that, on average, females are

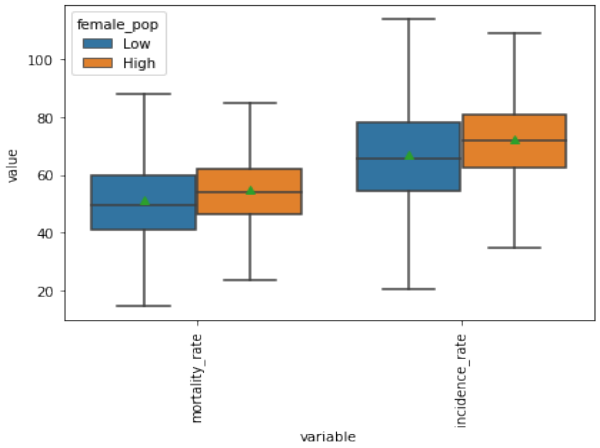

Fig. 8. Plot of Mortality and Incidence Rates for regions with low and high female population

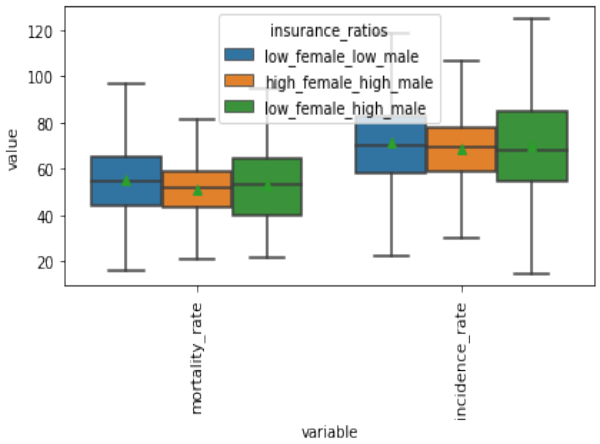

Fig. 9. Plot of Mortality and Incidence Rates among areas with low male and female insurance ratios in blue, high female and high male insurance ratios in orange and low female and high male insurance ratios in green. It must be noted that no samples with high female and low male insurance ratios exist.

more poverty-stricken than males. From figure 6, we observe that in regions, where poor females are more numerous as compared to poor males, incidence and mortality rates, are higher. Further, from Figure 7, it can be inferred that regions with a higher female population (above the median female population) have higher mean and median incidence and mortality rates. These three plots help us deduce that.

V. MODEL IMPROVEMENTS

The linear regression model is the way that we deal with continuous output in terms of predicting this output, but first, we need to train this model on some data, evaluate the model, and get the best parameters that minimise the cost function. We can train a linear regression model using gradient descent optimization or using what is called the normal equation, which helps us to get the best values of these weights directly, but this works only if we have a small number of features.

R^2 statistic: **0.21338**

Variable	Coefficients
All_Poverty	-7.99×10^{-5}
Med_Income	-5.41×10^{-4}
All_With	1.46×10^{-5}
All_Without	-1.05×10^{-5}
falling	1.18×10^{-4}
rising	-1.13×10^{-4}
Constant	78.58

Fig. 10. results using Ridge Regression Regularization

In our work here, we use gradient descent optimization to update these weights in such a way it decreases the cost function at the end and for model improvement Instead of removing features with a high level of collinearity, we will use regularized models. Regularized models address collinearity by shrinking coefficients towards 0, thus reducing model variance.

A. Ridge Regression

We modify our objective function by introducing a $l2norm$ penalty to penalise larger coefficients.

$$\phi(W) = \frac{1}{2}||XW - t||_2^2 + \lambda||W||_2^2 \tag{13}$$

where λ is the Regularization strength. Regularization improves the conditioning of the problem and reduces the variance of the estimates.Post-hyper-parameter tuning, we get λ to be, $\lambda = 10^6$.

B. Lasso Regression

We modify our objective function by introducing an $l1$ norm penalty to penalise more significant coefficients. The Lasso is a linear model that estimates sparse coefficients. It is helpful in some contexts due to its tendency to prefer solutions with fewer non-zero coefficients, effectively reducing the number of features upon which the given answer is dependent where λ is the Regularization strength Post hyper-parameter tuning, we get λ to be, $\lambda = 10^3$.

$$\phi(W) = \frac{1}{2}||XW - t||_2^2 + \lambda||W||_1 \tag{14}$$

C. ElasticNet Regression

ElasticNet is a linear regression model trained with both l1 and l2 norm regularization of the coefficients. This combination allows for learning a sparse model where few of the weights are non-zero like Lasso, while still maintaining the regularization properties of Ridge.

$$\phi(\boldsymbol{w} = \frac{1}{2}||\boldsymbol{X}\boldsymbol{w} - t||_2^2 + \lambda\rho||\boldsymbol{w}||_1 + \frac{\lambda(1-\rho)}{2}||\boldsymbol{w}_2^2 \tag{15}$$

R^2 statistic: **0.21326**

Variable	Coefficients
All_Poverty	-7.12×10^{-5}
Med_Income	-5.40×10^{-4}
All_With	1.46×10^{-5}
All_Without	-1.04×10^{-5}
falling	0
rising	0
Constant	77.99

Fig. 11. results using Lasso Regression Regularization

R^2 statistic: **0.21338**

Variable	Coefficients
All_Poverty	-7.98×10^{-5}
Med_Income	-5.40×10^{-4}
All_With	1.46×10^{-5}
All_Without	-1.04×10^{-5}
falling	0
rising	0
Constant	78.57

Fig. 12. results using ElasticNet Regression

where λ is the Regularization strength and ρ is the $l1$ ratio which controls the convex combination of $l1$ and $l2$ norm penalties. Post-hyper-parameter tuning, we get λ to be, $= 10^2 \& \rho = 0.1$

The ElasticNet Regression model gives us a score equal to that of the Ridge Regression model while being sparser at the same time. In accordance with Occam's Razor principle [8], we will choose the **ElasticNet Regression model to be our final model**.

VI. CONCLUSIONS

In this study, we observed that although not high, there is some correlation between socio-economic status and incidence and mortality (from the models as well as the exploratory analysis). Furthermore, females are poorer and more vulnerable to cancer incidence and mortality, so the non-profit organisation must raise awareness about the disease among them, work to create more employment opportunities for women, and strive for equitable pay among those who already work. In the future, non-linear models may be explored to improve their performance of models.

REFERENCES

[1] Shaffer, K. M., Jacobs, J. M., Nipp, R. D., Carr, A., Jackson, V. A., Park, E. R., ... & Temel, J. S. (2017). Mental and physical health correlates among family caregivers of patients with newly-diagnosed incurable cancer: a hierarchical linear regression analysis. Supportive Care in Cancer, 25(3), 965-971.

[2] Murugan, S., Kumar, B. M., & Amudha, S. (2017, September). Classification and prediction of breast cancer using linear regression, decision tree and random forest. In 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC) (pp. 763-766). IEEE.

[3] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis. John Wiley & Sons.

[4] Seber, G. A., & Lee, A. J. (2012). Linear regression analysis. John Wiley & Sons.

[5] Gross, J., & Groß, J. (2003). Linear regression (Vol. 175). Springer Science & Business Media.

[6] Bottou, L. (2012). Stochastic gradient descent tricks. In Neural networks: Tricks of the trade (pp. 421-436). Springer, Berlin, Heidelberg.

[7] Yan, X., Su, X. (2009). Linear regression analysis: theory & computing. world scientific.

[8] Ogutu, J. O., Schulz-Streeck, T., Piepho, H. P. (2012, December). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. In BMC proceedings (Vol. 6, No. 2, pp. 1-6). BioMed Central.