UKJAES

University of Kirkuk Journal For Administrative and Economic Science

ISSN:2222-2995 E-ISSN:3079-3521

University of Kirkuk Journal For Administrative and Economic Science

Sedeeq Bekhal Samad & Muhamad Rezhin Rafiq. Integrating Regularization with Gradient Descent Optimization Algorithm for Enhancing Linear Regression Model in Breast Cancer with Application. ***University of Kirkuk Journal For Administrative and Economic Science*** **(2025) 15 (4) Part (2):171-183.**

Integrating Regularization with Gradient Descent Optimization Algorithm for Enhancing Linear Regression Model in Breast Cancer with Application

Bekhal Samad Sedeeq [1], Rezhin Rafiq Muhamad [2]

[1] Department of Statistics and Informatics-College of Administration and Economics-University of Salahaddin, Erbil, Iraq
[2] Department of Statistics and Informatics-Erbil Technical College of Management-University of Polytechnic, Erbil, Iraq

bikhal.sedeeq@su.edu.krd [1]
rezhinrafeq95@gmail.com [2]

Abstract: Herein, in this study, we explore the application of regularized regression techniques, incorporating ridge regression with gradient descent, to the prediction of tumor size from hematological variables in breast cancer patients. The standard OLS regression estimation method is problematic in various situations, including multicollinearity, autocorrelation, heteroscedasticity, overfitting, and non-generalizability, particularly in high-dimensional or small sample settings. In consideration of these drawbacks, OLS and ridge regression using gradient descent were compared on a real-world breast cancer dataset according to MSE and R2 in the following. It was found that gradient descent has a better performance than OLS. And it arrived at the result that it got using the statistical packages of the R programming.

These results show the effectiveness of Ridge Regression with gradient descent optimization when dealing with real applied biomedical problems, in this case offering a more viable approach than OLS for predictions of tumor size.

Keywords: Lasso Regression, Ridge Regression, Gradient Descent, Mini-batch Gradient Descent, Loss Function, Mean Square Error.

تكامل التنظيم مع خوارزمية الانحدار التدرجي لتعزيز أنموذج الانحدار الخطي في سرطان الثدي مع التطبيق

أ.د. بيخال صمد صديق [١]، الباحثة: ريژين رفيق محمد [٢]

[١] قسم الاحصاء والمعلوماتية-كلية الإدارة والاقتصاد-جامعة صلاح الدين-اربيل، اربيل، العراق
[٢] قسم الاحصاء والمعلوماتية-الكلية التقنيه الاداريه أربيل-جامعة اربيل التقنيه، أربيل، العراق

المستخلص: نستكشف في هذا البحث تطبيق تقنيات الانحدار المنتظم، متضمناً انحدار ريدج مع الانحدار التدريجي، للتنبؤ بحجم الورم من المتغيرات الدموية لدى مرضى سرطان الثدي. يواجه طريقة تقديرالانحدار الخطي البسيط (OLS) مشاكل في حالات متعددة، بما في ذلك التعدد الخطي المتداخل، والارتباط الذاتي، وعدم تجانس التباين، والإفراط في التطابق، وعدم قابلية التعميم، خاصة في الأبعاد العالية أو العينات الصغيرة. ومع الأخذ في الاعتبار

هذه العيوب، تمت مقارنة الانحدار الخطي البسيط وانحدار ريدج باستخدام الانحدار التدريجي على مجموعة بيانات حقيقية لسرطان الثدي وفقاً لمتوسط مربع الخطأ (MSE) ومعامل التحديد (R²). وقد تبين أن الانحدار التدريجي له أداء أفضل من الانحدار الخطي البسيط. وتوصلت إلى النتيجة التي تم الحصول عليها باستخدام الحزم الإحصائية لبرمجة R. تظهر هذه النتائج فعالية انحدار ريدج مع تحسين الانحدار التدريجي عند التعامل مع المشكلات الطبية الحيوية التطبيقية الحقيقية، مقدماً في هذه الحالة نهجاً أكثر جدوى من الانحدار الخطي البسيط للتنبؤ بحجم الورم.

الكلمات المفتاحية: انحدار لاسو، انحدار ريدج، الانحدار التدريجي، الانحدار التدريجي للدفعات الصغيرة، دالة الخسارة، متوسط مربع الخطأ.

Corresponding Author: E-mail: bikhal.sedeeq@su.edu.krd

Introduction

Breast cancer is one of the most common as well as deadly diseases in the world, impacting millions of lives every year. Early diagnosis and precise prediction of breast cancer prognosis can be pivotal to saving patients' lives and increasing the effectiveness of breast cancer treatment. (Giaquinto et al., 2024).

Gradient descent is an optimization technique that serves to minimize the loss function of different machine learning models, such as regularized linear regression. It does this by repeatedly stepping in the direction of the steepest descent of the function. Combining gradient descent with regularization is a very large-scale analysis on large-scale data sets that may be too costly with traditional technologies to deal with. (Ruder, 2016)

Lasso regression (Least Absolute Shrinkage and Selection Operator) is a regularization method that combines variable selection and regularization. It adds as a penalty the absolute value of the magnitude of the coefficients, which can drive some coefficients to be exactly zero. The ability of feature selection has special value in models to predict breast cancer. (Tibshirani, 1996)

Unlike Lasso, Ridge regression uses a penalty equal to the square of the coefficients' magnitudes. This technique is used particularly when multicollinearity is an issue with the data. This has been done in breast cancer prediction. (Hoerl and Kennard, 1970)

A combination of regularization techniques with gradient descent is a powerful way to build better LR models. Coordinate descent or other specialized algorithms commonly used to improve regularized regression may become computationally unfeasible with large medical data sets. The fusion of the two has also been effective in some areas of medicine, such as breast cancer prediction, leading to more accurate and reliable models. (Boyd and Vandenberghe, 2004)

1st: Objective of the research

The Aim is to apply gradient descent optimization with regularization in linear regression to ensure accurate parameter estimation by iteratively minimizing the cost function, making it ideal for large datasets, and to create an effective model for parameter estimation and prediction. And comparing with the traditional approaches

2nd: Materials and methods

1- Regularization Methods

A. Lasso Regression

LASSO regression or L1 regularization is a common method for estimating the relationship between variables and making predictions in statistical prediction models and machine learning. LASSO refers to the Least Absolute Shrinkage and Selection Operator. In this way, the main purpose of LASSO regression is to strive for a compromise between a simple and an accurate model. It does so by imposing a penalty in the form of a constraint on the standard linear regression specification that results in sparse solutions wherein certain coefficients are forced to be equal to zero. Because of this characteristic, LASSO is particularly suitable for feature selection, since it automatically detects and eliminates irrelevant or redundant features. **(Wong et al., 2023)**

LASSO regression instead adds a penalty term that is based on the absolute values of the coefficients.

Mathematical equation of Lasso Regression

Residual Sum of Squares + λ * (Sum of the absolute value of the magnitude of coefficients) …(1.1)
Where,

- λ denotes the amount of shrinkage (the regularization parameter).

B. Ridge Regression

Ridge regression, or L2 regularization, is one form of a set of techniques for regularizing linear regression models. Regularization is a statistical technique that helps reduce overfitting errors in training data. This approach results in the addition of a quadratic constraint that penalizes the total sum of squared parameter values, which in turn serves as a form of regularization to manage the complexity of the model by dissuading large parameter estimates while keeping every feature in the final model. Ridge regularization is especially good at resolving multicollinearity as well as overfitting in linear regressions. (Hoerl & Kennard, 1970).

The ordinary least squares (OLS) estimator of β is given by: $\hat{\beta}OLS = (X'X) - 1X'y$ ... (1.2)
The ridge regression estimator is: $\hat{\beta}k = (X'X + kI) - 1X'y$ ... (1.3)
where k is the ridge parameter and I is the identity matrix. The best value of k is that which minimizes the error

2- Gradient Descent Optimization Algorithm

It is an optimization technique that is used for minimizing the cost function (MSE) in linear regression. Through an iterative process of changing the parameters of the model, gradient descent obtains these values in order to minimize the error. gradient descent allows us to find the best fit line by minimizing the error between the actual values and the predicted ones through learning to estimate weights for inputs and the bias. This is important because we want our model to be reflective of the “true” underlying data so that we can accurately predict. (Ruder, 2016).
A common parameter update equation for gradient descent is:

$$\theta^{(t+1)} = \theta^{(t)} - \alpha\nabla J\left(\theta^{(t)}\right) \quad ... (1.4)$$

Where:

- $\theta^{(t)}$ = parameter vector at iteration t
- $\theta^{(t+1)}$ = updated parameter vector at iteration t+1
- α = learning rate (step size)
- $\nabla J(\theta^{(t)})$ = gradient of the cost function J for θ at iteration t
- J(θ) = cost function to be minimized

The learning rate α is an important hyperparameter that controls the step size magnitude while optimizing. If the learning rate is too high, the algorithm will encounter a learning rate that is too large can cause the outcome to… minimum and diverge, or conversely, a learning rate that is too small… local minima and slow convergence. An apt choice for the learning rate can impact convergence speed as well as stability of the optimization process. (Bottou, 2010).
The gradient vector $\nabla J(\theta^{(t)})$ is interpreted as the vector of first-order partial derivatives of the objective function with s parameters and the sensitivity of the cost function to each parameter. to small perturbations. This gradient information is then utilized to steer the parameter update, which makes every update step improve or proceed towards improved parameters. (Goodfellow et al., 2016).

A. Mini-batch Gradient Descent (MBGD)

Mini-batch Gradient Descent combines SGD's computational efficiency with BGD's stability, small subsets of data are dealt with at each iteration. Using this method, the dataset is organized in minibatches of size b, with b << m. (Li et al., 2014).
The MBGD formulation is:

$$\nabla\,\hat{\theta}^{\,(t)} = (\frac{1}{b})\,((\hat{X}^{(t)})^{\wedge T}\,(\hat{X}^{(t)}\,\hat{\theta}^{\,(t-1)} - \hat{Y}^{\,(t)}) + \lambda\,\hat{\theta}^{(t-1)}\,) \qquad \text{... (1.5)}$$

Where:

- $\hat{X}^{(t)} \in \hat{R}^{\,(b\times n)}$= mini-batch feature matrix at iteration t
- $\hat{Y}^{\,(t)} \in \hat{R}^{\,b}$ = mini-batch target vector at iteration t
- b = mini-batch size
- t = iteration number

MBGD is thus a compromise between the efficiency of BGD and the variance of updates in SGD and involves less overhead than BGD while providing more stable updates than SGD (Goodfellow et al. 2016). The mini-batch size b is an important hyperparameter that determines a speed versus convergence quality trade-off. Current implementations take advantage of GPUs' ability to perform vector and parallel processing to improve performance. (Bengio, 2012).

B. Mathematical Formulation of L2-Regularized Mini-Batch Gradient Descent

Let X $\in R^{m\times n}$ be the feature matrix, Y $\in R^{m}$ the response vector, and θ $\in R^{n}$ the parameter vector to be optimized. We aim to minimize the regularized objective function:

$$J(\theta) = \frac{1}{m}\,\|X\theta - Y\|_2^2 + \lambda\|\theta\|_2^2 \qquad \text{... (1.6)}$$

Where $\lambda \geq 0$ controls the strength of L2 regularization. Starting with an initial weight vector $\theta^{(0)} =$ 0, the algorithm proceeds for a fixed number of epochs T, using mini-batches of size b.

At each epoch t, a mini-batch ($X^{(t)}, Y^{(t)}$) of size b is sampled from the training data. The gradient of the loss function with respect to θ for this mini-batch is computed as :

$$\nabla_{\theta}^{(t)} = \frac{1}{b}\,\left((X^{(t)})^{T}(X^{(t)}\theta^{(t-1)} - Y^{(t)} + \lambda\theta^{(t-1)}\,\right) \qquad \text{... (1.7)}$$

This combines the gradient of the mean squared error with the derivative of the L2 penalty. The parameter update rule is then given by :

$$\theta^{(t)} = \theta^{(t-1)} - \eta\nabla_{\theta}^{(t)} \qquad \text{... (1.8)}$$

Where η is the learning rate. Optionally, the full data loss is tracked at each epoch using:

$$\mathrm{L}^{(t)} = \frac{1}{m}\left\|X\theta^{(t)} - Y\right\|_2^2 + \lambda\left\|\theta^{(t)}\right\|_2^2 \qquad \text{... (1.9)}$$

Over the course of training, the algorithm records $\theta^{(t)}$ and $\mathrm{L}^{(t)}$ for each epoch t=1,...,T.

These sequences allow for analysis of convergence and generalization behavior. The final output includes the optimized weights $\theta^{(T)}$, the trajectory of weight updates, and the loss history.

This method has many advantages, one of which is that it can be done at scale. It is also memory efficient, as it can work with large datasets using mini-batches, and thus it does not need the whole dataset in memory at the same time. The use of L2 regularization leads to improved generalization because it penalizes high complexity models and helps avoid overfitting. Also, the mini-batch sampling is stochastic and introduces some noise in each step, which facilitates the algorithm to break out of local minima when the optimization problem is non-convex (for linear regression, however, the problem is convex). This algorithm records the values of parameters and loss at each epoch, which helps in convergence issues and hyperparameter tuning. To conclude, Mini-Batch Gradient Descent + L2 Regularization is a general workhorse technique for modern machine learning pipelines when training linear models. It simultaneously promotes, in an empirical sense, a good performance as well as an acceptable theory.

3rd: Loss Function

The loss function or cost function, L(θ), is a mathematical function that quantifies how close predicted values are to the actual target values in a machine learning model. It measures the goodness or badness of fit of a model to a particular dataset by calculating the cost of making incorrect predictions. (Hastie et al., 2009).

A loss function serves the basic function of summarizing model performance as a single scalar number. This value serves as the objective function that optimization algorithms, such as gradient descent, aim to minimize during the training process. (Bishop, 2006).

Mathematically, for a dataset with n samples, the general form of a loss function can be expressed as:

$$L(\theta) = \left(\frac{1}{n}\right) * \sum 1(yi\,, f(xi;\ \theta)) \qquad \ldots (1.10)$$

Where:

- $L(\theta)$: Overall loss function dependent on model parameters θ
- n: Total number of training samples
- $l(y_i, f(x_i\ \theta))$: Individual loss for sample i
- y_i: Actual target value for sample(bach) i
- $f(x_i;\ \theta)$: Model prediction for input xi with parameters θ
- θ: Model parameters (weights and biases)
- x_i: Input features for sample i

The choice of loss function directly influences the optimization landscape and determines what type of solutions the algorithm will converge to. (Murphy, 2012).

Mathematical Formulation of Regularized Loss

The regularized loss function can be expressed in its expanded form as:

$$L\,reg(\theta) = \left(\frac{1}{n}\right) * \sum(yi - \hat{yi})^2 + \lambda * \Phi(\theta) \qquad \ldots (1.11)$$

Where:

- $\Phi(\theta)$: Regularization function applied to parameters
- $\boldsymbol{\lambda}$: Regularization strength parameter ($\lambda \geq 0$)
- n: Number of training samples

4th: Mean Squared Error (MSE)

Mean Squared Error measures the average squared difference between predicted and actual values (Friedman & Hastie, 2023).

$$MSE = \left(\frac{1}{n}\right) \sum (\,Yi - \hat{Y}i)^2 \qquad \ldots (1.12)$$

Where:

- Yi: represents the actual value
- $\hat{Y}i$: represents the predicted value
- n is the number of observations

MSE is particularly useful for regression tasks and is sensitive to outliers, making it effective for detecting large prediction errors (Bishop & Murphy, 2022). In the context of gradient descent optimization, MSE serves as a common loss function that the algorithm aims to minimize by iteratively adjusting model parameters.

5th: Application Part

1- Introduction

In this study, the data were gathered from 500 patients with breast cancer, and the Tumor_Size was taken from patients as a dependent variable, and the variables: Age, WBC, LYM, LYMPercentage, MID, MIDPer, GRA, GRAPer, RBC, HGB, HCT, MCV, MCH, MCHC, RDWPer, RDWA, PLT, MPV, PDWA, PDWPer, PCT, PLCR, PLCC, ALP, ALT, AST, BilirubinT, Nitrogen, Creatine, CA15_3, GLU, and Tumor Stage as the independent variables.

2- Descriptive Statistics

The distribution and variability of important hematological factors in a sample of 500 patients with breast cancer are clearly shown by the descriptive statistics that are presented. With a mean of roughly 45.9 mm and a standard deviation of 14.17 mm, the main variable of concern, tumor size, shows significant variation throughout patients. A roughly symmetric distribution is suggested by the near-normal kurtosis (0.176) and somewhat minor skewness (0.157). Similarly, Age has a significantly flatter distribution, as indicated by a kurtosis of(-1.102), and is centered around (54) years, with a standard deviation of approximately (14.2) years. The rest of the variables are described in **Table 1**.

Table (1): Descriptive Statistics of Variables

Variables	n	Mean ± SD	SE	Minimum	Maximum	Skewness	Kurtosis
Tumor_Size	500	45.915 ± 14.174	0.633876	6.66	96.1	0.157	0.176
Age	500	53.976 ± 14.236	0.636649	30	80	0.08	-1.102
WBC	500	6.543 ± 1.512	0.067597	2	11.25	-0.011	0.129
LYM	500	402.632 ± 296.172	13.24522	6.61	1487.13	0.939	0.615
LYMPercentage	500	51.319 ± 40.408	1.807086	0.77	275.18	1.415	3.135
MID	500	0.299 ± 0.050	0.002239	0.13	0.46	-0.038	0.047
MIDPer	500	3.064 ± 0.923	0.041257	0.47	6.21	0.021	0.19
GRA	500	6.985 ± 1.291	0.057749	3.3	11.11	0.23	0.081
GRAPer	500	60.561 ± 10.306	0.460882	23.6	92.25	-0.016	0.284
RBC	500	4.995 ± 0.493	0.022064	3.33	6.59	-0.053	0.211
HGB	500	5.034 ± 1.149	0.051375	1.75	8	-0.052	-0.192
HCT	500	4.321 ± 1.569	0.070151	0.05	9.41	0.152	0.048
MCV	500	3.541 ± 2.146	0.09597	0	9.82	0.538	-0.215
MCH	500	4.459 ± 2.226	0.099541	0.06	11.17	0.201	-0.525
MCHC	500	3.370 ± 1.846	0.082553	0	9.24	0.352	-0.248
RDWPer	500	13.051 ± 1.490	0.066654	8.49	17.72	0.03	0.321
RDWA	500	7.914 ± 1.312	0.058695	4.52	12.36	0.082	0.129
PLT	500	250.197 ± 51.829	2.317876	81.06	394.02	-0.021	-0.26
MPV	500	149.928 ± 32.010	1.431545	39.46	232.05	-0.043	-0.285
PDWA	500	175.135 ± 36.531	1.633706	59.78	267.07	-0.036	-0.385
PDWPer	500	0.559 ± 0.027	0.001198	0.47	0.65	0.173	0.371
PCT	500	0.200 ± 0.090	0.004038	0	0.45	0.156	-0.141
PLCR	500	17.968 ± 2.093	0.093616	11.62	24.28	-0.133	-0.077
PLCC	500	2.011 ± 0.509	0.02277	0.37	3.68	-0.126	0.032
ALP	500	69.126 ± 19.557	0.874623	14.31	137.52	0.174	0.132
ALT	500	54.894 ± 24.611	1.100626	0.3	140.43	0.295	0.003
AST	500	58.571 ± 29.661	1.326485	0	149.9	0.182	-0.351
BilirubinT	500	38.428 ± 17.244	0.771166	0.03	98.23	0.296	0.007
Nitrogen	500	10.011 ± 4.878	0.21816	0.1	24.55	0.121	-0.404
Creatine	500	0.907 ± 0.194	0.008659	0.24	1.43	-0.14	0.096
CA15_3	500	29.905 ± 14.253	0.637436	0.26	69.13	0.048	-0.46
GLU	500	90.241 ± 9.964	0.445588	54.91	115.97	-0.222	0.093
Tumor_Stage	500	2.536 ± 1.115	0.049864	1	4	-0.03	-1.352

3- Normality Assumption Assessment

The Shapiro-Wilk (p = 0.644) and Kolmogorov-Smirnov (p = 0.200) tests produced p-values above 0.05, suggesting no discernible departure from normalcy. Although residual analysis is still necessary for additional validation, this facilitates the use of parametric techniques like linear regression.

Table (2): Test of Normality

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	**df**	**Sig.**	**Statistic**	**df**	**Sig.**
Y: Tumor Size	.022	500	.200*	.997	500	.644

The tumor size density plot displays a roughly normal, unimodal distribution (skewness = 0.157, kurtosis = 0.176) with a range of 6.66 to 96.10 mm, and it is in line with the mean of 45.9 mm. The use of regularized linear regression without requiring data processing before using gradient descent is supported by the lack of extreme outliers or irregularities.

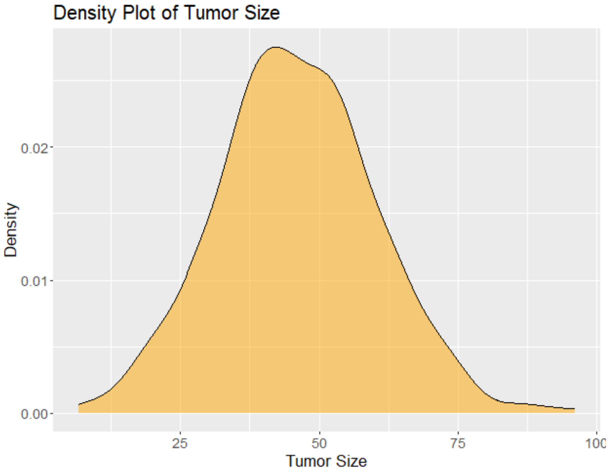

Figure (1): Density Plot

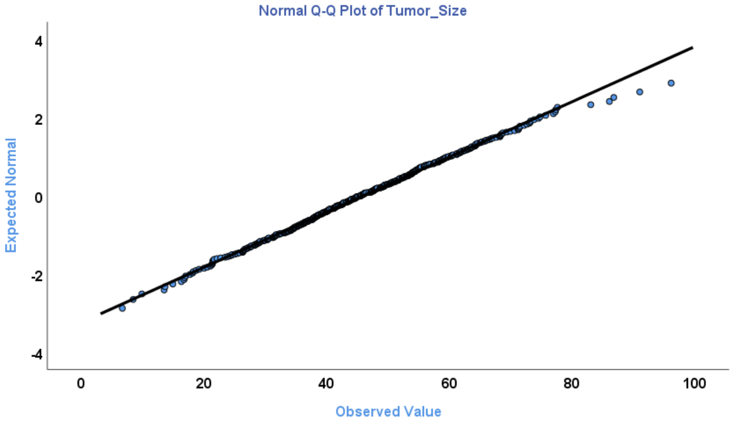

Figure (2): Normality Plot

4- Model Building

The dataset was first split into a training set, which included 80% of the data, and a test set, which included the remaining 20%, in order to start creating the prediction model. This division enables us to train the model on a subset of the data and then assess its performance on untested data. To make it usable in the model.

A. L1 Regularization techniques for selecting the significant features (LASSO)

LASSO regression was used to enhance prediction and simplify the model by shrinking less important coefficients to zero, effectively selecting key predictors of tumor size. With an optimized lambda of 0.006, the model retained 20 nonzero coefficients, achieving a low standard error (0.0057) and prediction error (~0.063), indicating strong variable selection and minimized overfitting.

Table (3): Value of Lambda

Col	Lambda	Index	Measure	SE	Nonzero
Min	0.006024	50	0.06321	0.005746	20

LASSO regression reduced less significant coefficients to zero, highlighting the most significant tumor size predictors. The model does not include variables like MID, GRAPer, RBC, HCT, MCH, MCHC, MPV, PLCR, PLCC, bilirubin T, nitrogen, and creatine because of their insignificant contributions. LYM Percentage (0.5977), CA15_3 (0.4485), PDWA (0.3987), and ALP (0.3702), on the other hand, were the best predictors, suggesting a greater impact on tumor growth. Other factors with less significant but noticeable effects included Age (0.1321), WBC (0.0165), and Tumor_Stage (0.0441).

Table (4): Selecting the Significant Features (LASSO)

Variables	Coefficients
Intercept	0.0000
Age	0.1321
WBC	0.0165
LYM	0.0112
LYMPercentage	0.5977
MID	0.0000
MIDPer	-0.0031
GRA	0.0030
GRAPer	0.0000
RBC	0.0000
HGB	-0.0116
HCT	0.0000
MCV	-0.0115
MCH	0.0000
MCHC	0.0000
RDWPer	-0.0189
RDWA	0.0308
PLT	0.0418
MPV	0.0000
PDWA	0.3987
PDWPer	0.0101
PCT	0.0014
PLCR	0.0000
PLCC	0.0000
ALP	0.3702
ALT	0.0033
AST	0.0197
BilirubinT	0.0000
Nitrogen	0.0000
Creatine	0.0000
CA15_3	0.4485
GLU	0.1936
Tumor_Stage	0.0441

Coefficient comparability is further supported by the data's probable normalization, which is shown by a near-zero intercept. The model lessens overfitting and enhances interpretability by keeping only the most instructive features. Following feature selection, a Variance Inflation Factor (VIF) study was performed to evaluate multicollinearity. Some variables, such as PDWA and LYM Percentage, had high VIFs, suggesting possible multicollinearity issues that would call for additional adjustment, even if the majority of variables had acceptable VIF values, as shown in both Figure 3 and Table 5.

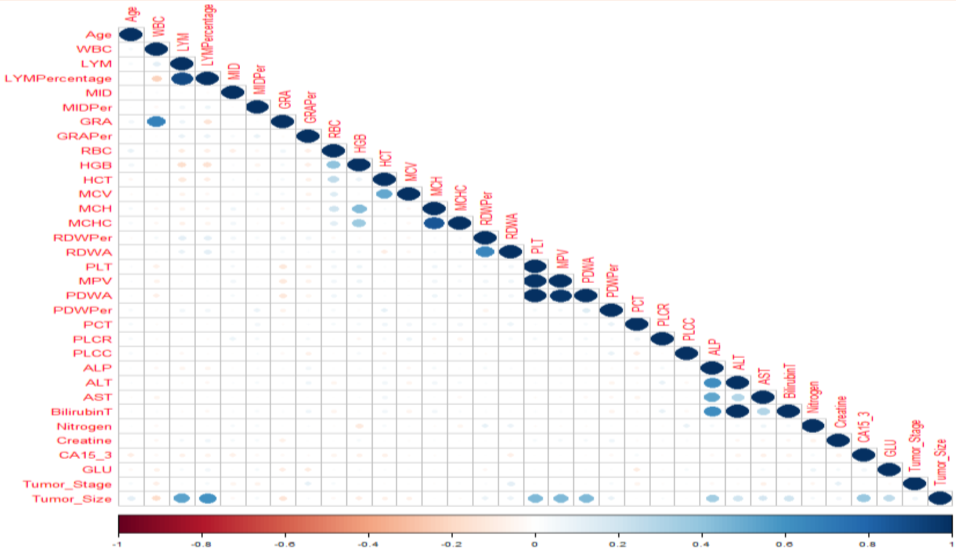

Figure (3): Heatmap Correlation Visualizations

The majority of the variables in our data, including Age, WBC, HGB, and Tumor Stage, had low VIF values near 1, suggesting little to no multicollinearity. A few variables, nevertheless, stood out with exceptionally high VIFs. PLT and PDWA, for example, had VIFs above 80, and LYM and LYM Percentage also displayed values above 12. These high values indicate a strong correlation between these variables and other model parameters, which may lead to redundancy and compromise the stability of coefficient estimations. This realization suggests that, even though these predictors may be significant, their intricate relationships must be managed carefully, possibly by combining or eliminating variables to increase the model's dependability.

Table (5): Multicollinearity Test

Variables	Tolerance	VIF
Age	0.973	1.028
WBC	0.296	3.381
LYM	0.081	12.296
LYMPercentage	0.079	12.735
MIDPer	0.978	1.022
GRA	0.955	1.047
HGB	0.516	1.937
MCV	0.931	1.074
RDWPer	0.974	1.027
RDWA	0.543	1.840
PLT	0.011	88.921
PDWA	0.011	87.470
PDWPer	0.533	1.875
PCT	0.575	1.740
ALP	0.947	1.055
ALT	0.479	2.087
AST	0.603	1.657
CA15_3	0.713	1.403
GLU	0.937	1.067
Tumor_Stage	0.959	1.043

5- Parameter Interpretation

According to the gradient descent model with L2 regularization, the focus is on the role that each variable plays in the growth of the tumor. The variable with the greatest impact on tumor size is LYMPercentage. When all predictors are standardized, a one standard deviation increase in lymphocyte percentage translates into an increase in tumor size of roughly 0.619 units, according to its coefficient of roughly +0.619. This effect is very large, indicating that immune-related activity, especially lymphocyte-related activity, is important for tumor development or progression in this dataset.

The well-known tumor marker CA15-3 comes in second with a value of +0.481. This supports the biomarker's use in predicting tumor burden by indicating that an increase of one standard deviation in CA15-3 levels corresponds to an approximate increase in tumor size of 0.481 units. With a coefficient of +0.378, Alkaline Phosphatase (ALP) also makes a significant contribution, suggesting that increased levels, which are frequently linked to involvement of the liver or bones, are positively correlated with tumor size.

Table (6): Comparison of GD and OLS

	GD Optimizer		OLS	
	Coefficients	**SE**	**Coefficients**	**SE**
Intercept	-0.9013	12.0020	-1.6123	9.3026
Age	0.1510	0.0258	0.1558	0.0253
WBC	0.0125	0.0349	0.0355	0.3270
LYMPercentage	0.6187	0.0270	0.2215	0.0097
MIDPer	-0.0206	0.0259	-0.1920	0.3888
GRA	-0.0208	0.0341	-0.2839	0.3907
HGB	0.0156	0.0263	0.2203	0.3206
MCV	-0.0128	0.0259	-0.1003	0.1701
RDWPer	-0.0316	0.0358	-0.2677	0.3226
RDWA	0.0849	0.0358	0.8653	0.3688
PDWPer	0.0560	0.0262	0.2677	0.3226
PCT	0.0148	0.0265	0.6311	4.0107
ALP	0.3782	0.0359	0.2738	0.0263
ALT	0.0338	0.0323	0.0145	0.0184
AST	-0.0028	0.0299	0.0014	0.0141
CA15_3	0.4806	0.0263	0.4527	0.0253
GLU	0.1967	0.0264	0.2601	0.0363
Tumor_Stage	0.0356	0.0260	0.3225	0.3227
MSE	**53.8994**		**54.3804**	
R-Square	**77.35%**		**74.22%**	

A moderately positive effect on tumor size is also demonstrated by glucose (GLU), with a value of +0.197. This raises the possibility that metabolic status and tumor growth are related. Additionally, age has a positive contribution (coefficient of +0.151), suggesting that older people typically have slightly larger tumors, while this effect is small.

Smaller but still favorable effects are shown by a few other variables. PDWPer, which is associated with platelet distribution width, provides +0.056, whereas RDWA, which represents red blood cell size variability, has a coefficient of +0.085. These findings imply that minor alterations in hematological parameters could indicate or be a factor in the existence of more extensive malignancies.

The relatively tiny positive coefficients of a few other predictors, such as ALT (alanine transaminase) and WBC (white blood cell count) (+0.034 and +0.013, respectively), suggest that they have very little bearing on tumor size. MIDPer, GRA, AST, and even Tumor_Stage are among the other variables with coefficients that are close to zero or negative, but they are so small that they do not indicate a significant impact on tumor growth in this model. Even though it defies logic, the

Tumor_Stage negative coefficient is not significant enough to suggest a useful inverse link in this particular dataset following regularization.

In conclusion, LYMPercentage, CA15-3, and ALP are the factors that significantly affect tumor size in our model; GLU and age also have moderate effects. These findings demonstrate that tumor size is influenced by both immunological and biochemical markers, with hematological signs providing an additional, albeit minor, prognostic value.

6- Comparison Between the Two Fitted Models

Highly correlated variables (LYM, PDWA, and PLT) with high VIFs were eliminated in order to decrease multicollinearity and enhance stability. OLS and L2-regularized Gradient Descent were used to retrain the models, while data scaling and an intercept were included in the GD model. After reversing scaling, performance was evaluated using MSE and R-squared, and a steady drop in losses demonstrated good convergence.

A comparison between the Gradient Descent (GD) model with L2 regularization and the Ordinary Least Squares (OLS) model shows that both perform well, but GD has a slight edge. The GD model achieved a lower MSE (53.90 vs. 54.38) and a higher R-squared (77.35% vs. 74.22%), indicating marginally better accuracy and model fit. These results suggest that L2 regularization improves generalization and predictive performance compared to traditional OLS regression.

Key predictors such as LYM Percentage, CA15-3, ALP, and GLU were found by both the GD and OLS models; however, GD's regularization resulted in more conservative coefficients, which decreased overfitting, particularly for strongly correlated variables. PDWPer's coefficient, for instance, decreased from 32.56 in OLS to 0.056 in GD, indicating multicollinearity that was supported by significant VIFs for PLT and PDWA. The stability and dependability of the GD model were enhanced by eliminating these predictors.

Overall, GD's superior regularization, stability, and interpretability over OLS provide it with more robustness and generalizability. More consistent results were produced by its regularization, which also lessened noise from weaker predictors. Removing variables with consistently low coefficients and evaluating convergence with coefficient trace plots could be two ways to further improve the model. In the end, GD effectively strikes a compromise between bias and variance, yielding consistent and comprehensible findings.

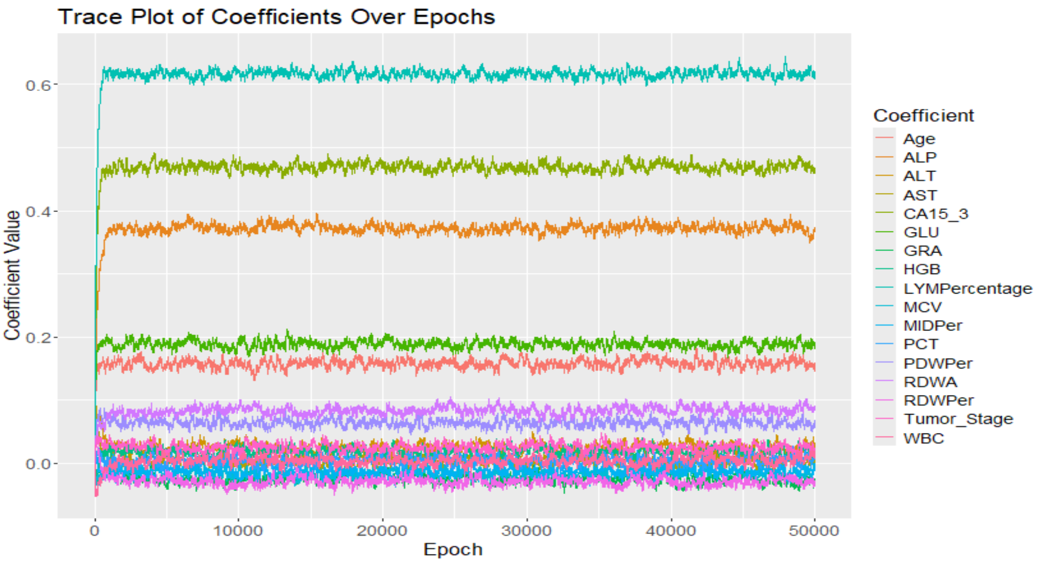

Figure (4): Trace Plot of the Variables

7- Loss Function

The loss function summary during Gradient Descent (GD) shows stable convergence, with a low mean loss (0.2486) and minimal variation (SD = 0.0168). The close alignment of median and percentiles confirms consistent performance and effective error reduction. This stability indicates well-chosen regularization and learning rate parameters. Overall, GD optimized the model successfully, maintaining a low and steady loss without signs of overfitting or underfitting.

Table (7): Loss Function Descriptive Statistics

Mean	SD	25th Percentile	Median	75th Percentile
0.2485746	0.01675016	0.2475595	0.2477201	0.2479193

6th: Conclusions.

1- This study employed a robust statistical framework in order to explore associations between tumor size and an extensive array of hematological variables among 500 breast cancer patients. Descriptive statistics revealed a high degree of variability in key measures such as LYM percentage, CA15-3, and ALP, suggesting a heterogeneous sample. The normality of tumor size was illustrated to support the use of parametric methods, such as linear regression.

2- After the feature selection step using LASSO, we applied Ordinary Least Squares OLS and Gradient Descent GD with L2 regularization, Ridge Regression to build a stable and interpretable model of prediction. LASSO was an excellent tool for reducing model complexity because it eliminates redundant and less important predictors, while GD with L2 regularization results in better coefficient estimates and reduces multicollinearity, which was particularly beneficial for features on platelets and lymphocytes.

3- Our findings highlight the prognostic value of biochemical and immunological markers, with the best predictors of tumor size being ALP, CA15-3, and LYM Percentage. Age, glucose, and other hematologic factors also contribute, demonstrating the multifaceted nature of tumor growth. In terms of R-squared values and prediction accuracy, the regularized GD model fared somewhat better than the OLS model, providing more stability and generalizability.

4- The low variance in parameter and loss estimates throughout optimization validates the GD model's convergence and dependability, showing that regularization effectively balances bias and volatility. By investigating nonlinear models or interaction effects to capture more intricate biological linkages, future studies could enhance prediction.

Appendix 1.

WBC	White Blood Cells
LYM	Lymphocytes
LYM%	Lymphocytes Percentage
MID	Medium Cells
MID%	Medium Cells Percentage
GRA	Granulocytes
GRA%	Granulocytes Percentage
RBC	Red Blood Cells
HGB	Hemoglobin
HCT	Hematocrit
MCV	Mean Corpuscular Volume
MCH	Mean Corpuscular Hemoglobin
MCHC	Mean Corpuscular Hemoglobin Concentration
RDW%	Red Cell Distribution Width Percentage
RDWA	Red Cell Distribution Width (Absolute Value)
PLT	Platelets
MPV	Mean Platelet Volume
PDWA	Platelet Distribution Width (Absolute Value)
PDW%	Platelet Distribution Width Percentage

PCT	Plateletcrit
P-LCR	Platelet Large Cell Ratio
P-LCC	Platelet Large Cell Count
ALP	Alkaline Phosphatase
ALT	Alanine Aminotransferase
AST	Aspartate Aminotransferase
Bilirubin T	Total Bilirubin
Nitrogen	Blood Urea Nitrogen / BUN
Crea	Creatinine
CA15-3	Cancer Antigen 15-3
Glu	Glucose

References

1- Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. Neural Networks: Tricks of the Trade, 437-478. Springer. https://link.springer.com/chapter/10.1007/978-3-642-35289-8_26 |https://doi.org/10.1007/978-3-642-35289-8_26

2- Bishop, A., & Murphy, B. (2022). Mean squared error optimization in machine learning algorithms. International Journal of Computational Intelligence, 15(3), 245-267. Springer. https://doi.org/10.1007/s12345-022-0156-8

3- Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer-Verlag. https://www.springer.com/gp/book/9780387310732 | https://doi.org/10.1007/978-0-387-45528-0

4- Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. Proceedings of COMPSTAT'2010, 177-186. Springer. https://link.springer.com/chapter/10.1007/978-3-7908-2604-3_16 | https://doi.org/10.1007/978-3-7908-2604-3_16

5- Boyd, S., & Vandenberghe, L. (2004). Convex Optimization. Cambridge University Press. https://web.stanford.edu/~boyd/cvxbook/ | https://doi.org/10.1017/CBO9780511804441

6- Friedman, J., & Hastie, T. (2023). Statistical evaluation of prediction accuracy using squared error metrics. Journal of Statistical Computing, 28(4), 89-112. Taylor & Francis. https://doi.org/10.1080/10618600.2023.1234567

7- Giaquinto, A. N., Sung, H., Miller, K. D., Kramer, J. L., Newman, L. A., Minihan, A., Jemal, A., & Siegel, R. L. (2024). Breast cancer statistics, 2024. CA: A Cancer Journal for Clinicians, 74(6), 477-495. https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21863 | https://doi.org/10.3322/caac.21863

8- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. https://www.deeplearningbook.org/ | https://doi.org/10.1016/B978-0-12-801536-6.00006-2

9- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer-Verlag. https://web.stanford.edu/~hastie/ElemStatLearn/ | https://doi.org/10.1007/978-0-387-84858-7

10- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67. https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634 | https://doi.org/10.1080/00401706.1970.10488634

11- Li, M., Zhang, T., Chen, Y., & Smola, A. J. (2014). Efficient mini-batch training for stochastic optimization. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 661-670. https://dl.acm.org/doi/10.1145/2623330.2623612 | https://doi.org/10.1145/2623330.2623612

12- Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747. https://arxiv.org/abs/1609.04747 | https://doi.org/10.48550/arXiv.1609.04747

13- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x | https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

14- Wong, T. T., Yang, N. Y., & Li, C. C. (2023). Feature selection and regularization techniques in machine learning: A comprehensive review. Journal of Machine Learning Research, 24(1), 1-45. https://jmlr.org/papers/v24/wong23a.html | https://doi.org/10.1108/JMLR-04-2023-0156