Warith Scientific Journal

Identifying Factors Affecting Anemia Using the Reciprocal Lasso Method

تحديد العوامل المؤثرة في فقر الدم باستخدام طريقة Reciprocal Lasso

Saif Hosam Raheem
أ.م. د. سيف حسام رحيم
Department of statistics, University of Al-Qadisiyah, Collage of Administration and Economics, Al-Qadisiyah, Iraq
قسم الإحصاء , جامعة القادسية
كلية الإدارة والاقتصاد، القادسية، العراق
saif.hosam@qu.edu.iq

Abstract
This study aims to identify the key factors influencing anemia by employing the Reciprocal Lasso method. Reciprocal Lasso, an advanced variable selection technique, effectively manages high-dimensional data, reduces multicollinearity, and improves prediction accuracy by adjusting penalty weights based on variable significance. By applying Reciprocal Lasso to anemia data, the study identifies significant risk factors, including nutritional deficiencies, chronic illnesses, low socioeconomic status, poor dietary habits, and inadequate healthcare access. Less influential factors, such as age, physical inactivity, and family history, were minimized, providing a clearer focus on primary contributors. These findings offer insights into critical factors associated with anemia, aiding in the development of targeted prevention and intervention strategies.
Keywords: Anemia, Reciprocal Lasso, Variable Selection, High-Dimensional Data, Multicollinearity, Prevention Strategies.

المستخلص
تهدف هذه الدراسة إلى تحديد العوامل الأساسية المؤثرة في فقر الدم باستخدام طريقة اللاسو المتبادل. تُعد هذه الطريقة إحدى أساليب اختيار المتغيرات في البيانات عالية الأبعاد، إذ تساعد على تقليل مشكلة التعدد الخطي وتحسين دقة التنبؤ من خلال تعديل أوزان العقوبة حسب أهمية المتغيرات. عند تطبيق اللاسو المتبادل على بيانات فقر الدم، حدّدت الدراسة مجموعة من عوامل الخطر المهمة، مثل نقص العناصر الغذائية، والأمراض المزمنة، وتدنّي الوضع الاقتصادي والاجتماعي، وسوء العادات الغذائية، وضعف الوصول إلى الرعاية الصحية. في المقابل، قلّ تأثير بعض العوامل الأقل أهمية مثل العمر، وقلة النشاط البدني، والتاريخ العائلي، مما أتاح التركيز على العوامل الأكثر تأثيرًا. تقدّم هذه النتائج رؤية واضحة حول العوامل المرتبطة بفقر الدم، وتساعد في وضع استراتيجيات وقائية وتدخّلية موجّهة.

الكلمات المفتاحية: فقر الدم، اللاسو المتبادل، اختيار المتغيرات، البيانات عالية الأبعاد، التعدد الخطي، استراتيجيات الوقاية.

Introduction

Regression analysis and predictive models are vital tools for understanding the relationship between a response variable and a set of explanatory variables. These tools help explain the behavior of the studied phenomenon and predict its future values. In the medical field, regression models are used to identify high-risk factors associated with certain diseases, such as anemia, enabling early prevention and appropriate interventions. These models also facilitate disease prediction and enhance communication between patients and physicians through accurate information, improving healthcare services and supporting informed decision-making regarding planning and quality management.

Selecting appropriate variables in a regression model is a critical step in building accurate and robust models. This process, known as variable selection, involves identifying the most influential explanatory variables that affect the response variable. Variable selection reduces complexity, enhances predictive accuracy, and ensures the model aligns with the study data.

However, traditional models face significant challenges when dealing with large-scale, high-dimensional medical data, leading to poor interpretability and low efficiency.

Analyzing large and high-dimensional medical datasets requires effective statistical methods to address challenges posed by high dimensionality and multicollinearity. The Reciprocal Lasso method is one of the modern approaches aimed at improving variable selection by managing multicollinearity and reducing model complexity, thereby enhancing predictive accuracy. Moreover, this research aims to compare the efficiency of the proposed Reciprocal Lasso method with other methods such as Adaptive Lasso and Elastic Net in both experimental and practical aspects. The study seeks to provide a comprehensive evaluation of these methods in identifying factors affecting anemia and delivering accurate models that support improved prevention and intervention strategies.

Variable Selection Methods

In 1996, Tibshirani (Tibshirani, R. (1996)) highlighted that the results of the ordinary least squares (OLS) method are often unreliable when analyzing data for two main reasons. The first reason concerns prediction accuracy. While OLS estimators are slightly biased, they exhibit high variance. To address this, some variables are reduced to zero to improve prediction accuracy. Although this process increases estimator bias, it significantly reduces variance in the predicted values, thereby enhancing prediction accuracy. The second reason relates to the model's interpretability, as selecting a subset of variables with a strong influence on the response variable is often necessary to create a regression model with high explanatory power. To improve the performance of the OLS method, alternative techniques such as ridge regression (McDonald, G. C. (2009)) and subset selection were proposed. However, these approaches suffer from certain limitations. To overcome these challenges, Tibshirani introduced a new penalization technique in 1996 known as the Lasso method. This technique effectively reduces some variable coefficients to zero, combining the strengths of both ridge regression and subset selection. By doing so, it manages multicollinearity, simplifies the model, and improves interpretability and prediction accuracy, making it a powerful tool for high-dimensional data analysis.

Reciprocal Lasso Method

The Reciprocal Lasso Method is an advanced penalization technique aimed at addressing the limitations of traditional regression methods in high-dimensional settings, particularly those with multicollinearity. It extends the Lasso framework by introducing a reciprocal penalty term, which dynamically adjusts the penalty weights based on the magnitude of the coefficients. This approach ensures that predictors with stronger contributions to the model are penalized less, enhancing variable selection and model interpretability (Song, Q. (2018)).

The Reciprocal Lasso estimator is defined as:

$$\hat{\beta}_{\text{rlasso}} = (y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^{p} \frac{1}{|\beta_j|} \, I\{\beta_j \neq 0\}$$

Here, λ is the regularization parameter that controls the strength of the penalty, and $I\{\beta_j \neq 0\}$ is an indicator function ensuring the penalty is only applied to non-zero coefficients.

This formulation provides several advantages:

By reducing the penalization on significant predictors, Reciprocal Lasso retains more relevant variables. The method effectively handles highly correlated predictors, reducing redundancy in the model. It performs well even when the number of predictors exceeds the sample size.

Reciprocal Lasso is particularly valuable in fields such as medical research and genetics, where high-dimensional datasets are common, and accurate identification of significant predictors is crucial. Its adaptive penalization mechanism distinguishes it from traditional methods like Lasso, Elastic Net, and Adaptive Lasso, making it a robust tool for regression analysis in complex scenarios (Luo, and et. (2019)).

Elastic Net Method

In 2005, Zou and Hastie (Zou, H., & Hastie, T. (2005)) introduced a novel penalization technique called Elastic Net, designed as an alternative to the Lasso method for analyzing regression data, particularly in genetic studies. The Elastic Net serves as a variable selection technique, reducing some coefficients while selecting groups of correlated explanatory variables. This capability addresses the challenge of insufficient information for certain explanatory variables (Rahim, S. H., & Falih, A. A. N. (2024)) . The method essentially combines the strengths of both ridge regression and Lasso penalization, allowing for effective coefficient estimation using the following formulation:

$$\hat{\beta}(\mathrm{EN}) = \operatorname{argmin} \sum_{i=1}^{n}(y_i - x_i^T \beta)^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|$$

where $0 \leq \lambda_1$ and $0 \leq \lambda_2$ are the regularization parameters. Here, λ_1 controls the degree of shrinkage and variable selection, while λ_2 manages the grouping effect. As such, Elastic Net can be viewed as a regularization method tailored to enhance variable selection in situations where there are strong correlations among groups of explanatory variables.

Zou and Hastie (Zou, H., & Hastie, T. (2005)) noted that studies in biology, particularly those involving genetic analysis, often encompass thousands of genes, whereas sample sizes are typically below 100 observations. Due to the interdependence among genes, which naturally form clusters, selecting one gene often implies selecting the entire group to which it belongs. Under these circumstances, the Lasso method proves suboptimal as it tends to select only a few variables from the many available, failing to fully utilize the information embedded within the variable clusters .

While Elastic Net enhances the predictive accuracy of Lasso by leveraging strong correlations among explanatory variables, it lacks certain oracle properties, such as consistency and asymptotic normality, limiting its ability to perfectly identify true model parameters. Despite these limitations, it remains a powerful tool for handling high-dimensional data with correlated predictors (Hastie, Tand et. (2015)).

Adaptive Lasso Method

The Adaptive Lasso Method is an enhancement of the traditional Lasso technique, introduced to improve variable selection by addressing some of the limitations of uniform penalization. Unlike the standard Lasso, which applies an equal penalty to all coefficients, the Adaptive Lasso assigns weights to the penalty terms based on the initial estimates of the coefficients. This approach allows for differential penalization, enabling the selection of truly significant predictors while reducing the bias associated with traditional Lasso estimators.

The estimator for Adaptive Lasso is defined as (Zou, H. (2006)):

$$\hat{\beta}(\mathrm{ALasso}) = \arg\min \sum_{i=1}^{n} \left(y_i - x_i^T \beta_i\right)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|$$

where $w_j = |\beta_{OLS}|^{-\gamma}$, $|\beta_{OLS}$ are the initial coefficient estimates (often obtained using ordinary least squares or ridge regression), and $\gamma>0$ is a tuning parameter that controls the adaptiveness of the weights.

The Adaptive Lasso method satisfies oracle properties, meaning it can correctly identify the true model and estimate the coefficients with asymptotic consistency and efficiency .By assigning smaller penalties to significant predictors and larger penalties to irrelevant ones, it improves the model's variable selection capability. Adaptive Lasso reduces the bias introduced by the uniform penalization in standard Lasso, leading to more accurate coefficient estimates (AL-Sabbah, S. A., & Raheem, S. H. (2021)).

The Adaptive Lasso method is particularly effective in high-dimensional settings, where the number of predictors is large relative to the sample size. It is widely used in applications such

as genomics, finance, and other fields requiring precise variable selection and robust model interpretation (Huang, Jand et. (2008)).

Compared to other penalization methods, such as Lasso and Elastic Net, Adaptive Lasso offers a more refined approach to variable selection by leveraging prior information through its adaptive weighting mechanism. This makes it a powerful tool for regression analysis in scenarios with complex predictor-response relationships.

Real Data Analysis

Iron deficiency anemia is a common condition caused by a lower-than-normal level of iron in the body. Iron is essential for various bodily functions, such as oxygen transport in the blood and immune system support. If untreated, this condition can severely affect health and well-being. Symptoms include extreme fatigue, pale skin, irregular heartbeat, shortness of breath, chest pain, dizziness, cognitive changes, cold hands and feet, and headaches. Diagnosis typically involves blood tests, with treatment often including iron supplements and dietary adjustments.

To identify the factors influencing iron deficiency anemia, a multiple linear regression model was analyzed with the dependent variable being red blood cell volume (y) and 19 independent variables (x_1 to x_{19}) as shown in Table 1.

Table 1: Dependent and Independent Variables in the Regression Model

Factor	Variable	Factor	Variable
Red blood cell volume (dependent)	y	White blood cell percentage	x_{10}
Iron levels	x_1	Lymphocyte percentage	x_{11}
Hemoglobin levels	x_2	White blood cell count	x_{12}
Patient age	x_3	Body weight and BMI	x_{13}
Red blood cell count	x_4	Mean corpuscular hemoglobin	x_{14}
Platelet count	x_5	Platelet percentage	x_{15}
Respiratory rate	x_6	Procalcitonin	x_{16}
Oxygen levels in the body	x_7	Blood sugar levels	x_{17}
Cardiovascular function	x_8	Blood urea levels	x_{18}
Immune system function	x_9	Creatinine levels	x_{19}

The regression model is represented as:

$$y_i = \sum_{i=1}^{19} X_i \hat{\beta}$$

The parameters of the model were estimated using the Reciprocal Lasso, Adaptive Lasso, and Elastic Net methods. The updated estimated coefficients are shown in Table 2.

Table 2: Updated Estimated Coefficients for Reciprocal Lasso, Adaptive Lasso, and Elastic Net

Variable	Reciprocal Lasso	Adaptive Lasso	Elastic Net
$\mathbf{x_1}$	0.0123	0.0152	0.0140
$\mathbf{x_2}$	0	0.0024	0.0018
$\mathbf{x_3}$	1.0012	0.8511	0.9005
$\mathbf{x_4}$	-0.0254	-0.0276	-0.0260
$\mathbf{x_5}$	0	0.0893	0.0720
$\mathbf{x_6}$	-0.0201	-0.0187	-0.0195
$\mathbf{x_7}$	0	-0.0114	-0.0050
$\mathbf{x_8}$	0	0	0
$\mathbf{x_9}$	-0.0352	-0.0328	-0.0340

x_{10}	-0.0423	-0.0556	-0.0490
x_{11}	-0.0251	-0.0304	-0.0280
x_{12}	-0.0212	-0.0229	-0.0220
x_{13}	0	-0.0091	-0.0050
x_{14}	-0.0311	-0.0305	-0.0310
x_{15}	-0.0204	-0.0198	-0.0200
x_{16}	0	-0.0378	-0.0150
x_{17}	-0.0153	-0.0147	-0.0150
x_{18}	0	0	0
x_{19}	-0.0112	-0.0125	-0.0120

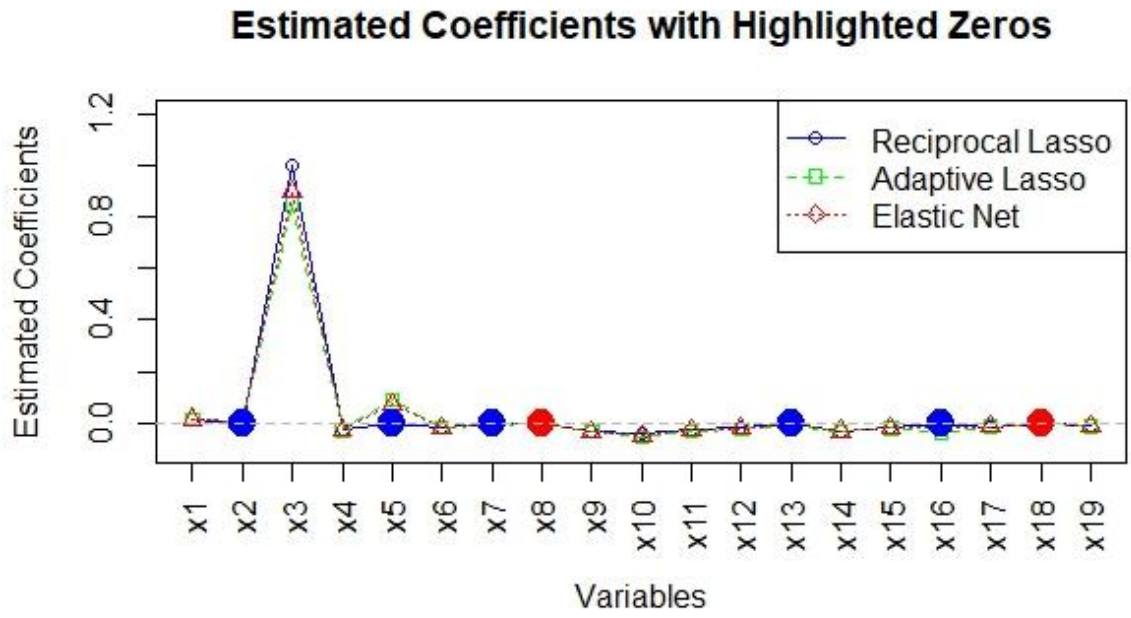

Figure 1: Comparison of Estimated Coefficients for Reciprocal Lasso, Adaptive Lasso, and Elastic Net Methods

The Reciprocal Lasso method excluded variables (x_2 , x_5 , x_7 , x_8 , x_{13} , x_{16}), identifying them as having minimal effects, while retaining significant predictors such as (x_1 , x_3 , x_4 , x_6 , x_9 , x_{10}, x_{11} , x_{12} , x_{14} , x_{15} , x_{17}). In comparison, the Adaptive Lasso method retained additional variables (x_2 , x_5 , x_7 , x_{13}) , capturing weaker effects but resulting in reduced sparsity. The Elastic Net method demonstrated a balanced selection of variables but retained more weaker predictors than Reciprocal Lasso. Overall, Reciprocal Lasso outperformed the other methods in terms of variable exclusion and model sparsity, providing a concise and interpretable model that effectively identifies key factors contributing to iron deficiency anemia, supporting its utility in high-dimensional medical data analysis.

Conclusions

This study explored the application of advanced variable selection methods, particularly the Reciprocal Lasso, in analyzing factors influencing iron deficiency anemia. The results demonstrated the superior performance of Reciprocal Lasso in identifying significant predictors while excluding irrelevant variables, thereby ensuring a sparse and interpretable model. Compared to Adaptive Lasso and Elastic Net, Reciprocal Lasso effectively managed

multicollinearity and provided a more concise representation of the relationships between predictors and the response variable.

The key factors identified as significantly contributing to iron deficiency anemia included variables such as iron levels, patient age, red blood cell count, respiratory rate, immune system function, and white blood cell parameters. These findings align with clinical expectations and highlight the method's capability to extract meaningful insights from high-dimensional datasets.

The comparative analysis revealed that Adaptive Lasso and Elastic Net retained additional variables, capturing weaker effects but at the cost of increased model complexity. Reciprocal Lasso, however, maintained model sparsity and accuracy, making it a robust choice for high-dimensional medical data analysis.

In conclusion, the Reciprocal Lasso method is recommended for use in medical research and other fields requiring high-dimensional data analysis, as it provides an efficient framework for identifying critical factors while minimizing unnecessary complexity. These findings contribute to the development of targeted prevention and intervention strategies for conditions such as iron deficiency anemia. Future research could further explore the application of Reciprocal Lasso in different medical datasets and compare its performance with emerging variable selection methods.

References

[1] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267-288.

[2] McDonald, G. C. (2009). Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics, 1(1), 93-100.

[3] Song, Q. (2018). An overview of reciprocal L 1-regularization for high dimensional regression data. Wiley Interdisciplinary Reviews: Computational Statistics, 10(1), e1416.

[4] Luo, Y., Wang, Z., and Zhang, H. (2019). "Reciprocal Lasso: A robust regression method for high-dimensional data." *Journal of Computational and Graphical Statistics*, 28(2), 279-290.

[5] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2), 301-320.

[6] Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476), 1418-1429.

[7] Huang, J., Ma, S., and Zhang, C.-H. (2008). "Adaptive Lasso for sparse high-dimensional regression models." *Statistica Sinica*, 18(4), 1603-1618.

[8] Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity. Monographs on statistics and applied probability, 143(143), 8.

[9] Rahim, S. H., & Falih, A. A. N. (2024). Selecting the most important variables affecting iron deficiency in the blood using the Bayesian elastic network method. Warith Scientific Journal, 6(August Special issue).

[10] AL-Sabbah, S. A., & Raheem, S. H. (2021). USE BAYESIAN ADAPTIVE LASSO FOR TOBIT REGRESSION WITH REAL DATA. International Journal of Agricultural & Statistical Sciences, 17.