مجلة جامعة بني وليد للعلوم الإنسانية والتطبيقية
Bani Waleed University Journal of Humanities and Applied Sciences
تصدر عن– جامعة بني وليد – ليبيا
Website:https://jhas-bwu.com/index.php/bwjhas/index
المجلد العاشر ـ العدد الثالث ـ 2025 ـ الصفحات (178- 184)
ISSN3005-3900

Machine Learning using Simple Linear Regression

Abdulwahed Faraj Almarimi [1]*, Asmaa Moustafa Salem [2]
[1.2] Department of Computer , Faculty of Education, Bani Waleed University, Bani Walid, Libya
AbdulwahedAlmarimi@bwu.edu.ly

التعلم الآلي باستخدام التوقع الخطي البسيط

عبدالواحد فرج المريمي [1]*، أسماء مصطفى سالم [2]
[2.1] قسم الحاسوب ، كلية التربية ، جامعة بني وليد ، بني وليد ، ليبيا.

تاريخ الاستلام: 17-05-2025 تاريخ القبول: 29-06-2025 تاريخ النشر: 08-07-2025

Abstract:

This paper illustrates linear regression as a basic method of machine learning for predicting continuous outcomes. It is one of the most widely used models for prediction in statistics and machine learning. Supervised learning has become an area of much research activity in the field of machine learning. The linear regression algorithm is perhaps one of the most common, comprehensive and widely used statistical and machine learning algorithms, as it is used to find the linear relationship between two variables, we have a comprehensive discussion of the theoretical and mathematical basis for this algorithm Where our study included defining the linear regression and analyzing the data used from the Kaggle_platform for regression analysis and the steps used in describing methods and functions, including calculating the cost function, the Gradient descent function and the error rate in order to minimize the differences between expected values and real values, and we evaluated a set of data for analysis represented by population and income with a focus on modeling the relationships between dependent (Y) and independent (X) variables. This paper is a small contribution to the understanding and simplification of the theory and practice of machine learning using linear regression.

Keywords: linear regression, machine learning, supervised learning, simple linear regression.

ملخص:

توضح هذه الورقة البحثية الانحدار الخطي كطريقة أساسية في التعلم الآلي للتنبؤ بالنتائج المستمرة. وهو من أكثر النماذج استخدامًا للتنبؤ في الإحصاء والتعلم الآلي. وقد أصبح التعلم المُشرف مجالًا بحثيًا واسعًا في مجال التعلم الآلي. ولعل خوارزمية الانحدار الخطي من أكثر خوارزميات الإحصاء والتعلم الآلي شيوعًا وشمولًا واستخدامًا، إذ تُستخدم لإيجاد العلاقة الخطية بين متغيرين. وقد تناولنا في هذه الورقة مناقشة شاملة للأساس النظري والرياضي لهذه الخوارزمية. حيث تضمنت دراستنا تعريف الانحدار الخطي وتحليل البيانات المستخدمة من منصة Kaggle لتحليل الانحدار والخطوات المستخدمة في وصف الطرق والدوال، بما في ذلك حساب دالة التكلفة ودالة الانحدار التدرجي ومعدل الخطأ بهدف تقليل الاختلافات بين القيم المتوقعة والقيم الحقيقية، وقمنا بتقييم مجموعة من البيانات للتحليل ممثلةً بالسكان والدخل مع التركيز على نمذجة العلاقات بين المتغيرات التابعة (Y) والمستقلة .(X) تُعدّ هذه الورقة مساهمة صغيرة في فهم وتبسيط نظرية وممارسة التعلم الآلي باستخدام الانحدار الخطي.

الكلمات الدالة: الانحدار الخطي، التعلم الآلي، التعلم المُراقَب، الانحدار الخطي البسيط.

1- Introduction

Data has gained an important role in various fields of life in recent times, leading to a growing need for methods that enable us to use this data intelligently. This has led to the emergence of machine learning, a branch of artificial intelligence that enables computer systems to learn from past experiences and gradually improve their performance. One of the main methods in this field is supervised learning, which relies on providing the system with data that contains valid inputs and outputs, which helps the system understand the relationship between them. One of the models used in this study is simple linear regression, which is the basis of supervised machine learning, which is emerging as a simple and effective tool in a number of disciplines, especially due to its ease of implementation and interpretability [1]. To better understand these challenges, it is important to recognize the prominent role that regression analysis present in scientific research. It is one of the most widely used statistical methods due to its effectiveness in exploring and interpreting relationships between variables. It is widely used due to its simplicity and efficiency, especially in linear models. In many real-world scenarios, this method achieves performance comparable to more advanced and complex methods, while maintaining its simplicity and ease of interpretation of its results [2,3]. In this context, there is a procedure known as the model utility test, which is used to verify the presence of a statistically significant relationship between the independent and dependent variables. The null hypothesis in this test states that there is no meaningful relationship between the two variables. These tests are applied within regression models designed to help researchers predict the value of a particular variable based on other variables. These models are constructed using several techniques, including simple linear regression, multiple linear regression, nonlinear regression, and multivariate regression [4]. In this study, we will discuss one such type, simple linear regression. Over the past few decades, research in linear regression has undergone a significant shift in focus, with efforts directed toward developing computationally efficient algorithms applicable to high-dimensional data. Many researchers have developed improved linear regression algorithms, such as Ridge Regression and Lasso Regression, both of which have proven effective in addressing challenges such as multicollinearity and overfitting. Researchers have also contributed to the development of innovative linear regression models tailored to specific problems in various fields, achieving remarkable results in practical applications [5,6]. The goal of linear regression is to construct a mathematical relationship between two variables (x and y) using coefficients calculated through linear algebra tools. This enables the relationship to be described quantitatively, which can be used for prediction and analysis. The dataset was obtained from Kaggle, a popular platform for sharing datasets and data science competitions. The dataset contains independent and target variables used to apply linear regression models and analyze the relationship between them. This dataset was selected for its quality and ease of use in machine learning applications[7].

2- Theoretical Background in our Study:

In this study, we apply the most common way to do simple linear regression using the SLR algorithm in python programming language. We use data from the Kaggle platform [6] for the analysis. Table 1 shows the complete set of parameter settings in this study.

Table1. Parameter settings used in this study.

Description	Symbol	value
the number of data sets	M	100
the number of Features	X_0,X_1	X_0=ones, X_1= 100_m values
the actual value	y	100_m
the cost function	θ_0,θ_1	0,0

the learning rate	α	0.01
the number of repetitions	Iterations	1000,10000

3- Linear Regression Algorithms:

In this paper, we will introduce two types of linear regression and apply a model to the first type, which is simple linear regression, and we hope in the future to make a practical comparison of both types [6]:

3.1 Simple Linear Regression (SLR_X):

Simple Linear Regression (SLR_X) It is a statistical method used to determine the relationship between two variables. One of the reasons for calling it simple is that it uses only two variables: an independent variable symbolized by (x) to predict the value of the other variable called the dependent variable symbolized by (Y). It finds or draws a straight line (Best Fiting Line) used to represent the relationship between the variables, which is closest to the data approximately, by finding the value of the coefficients (θ_0) and (θ_1), which determines the line and reduces regression errors. The relationship is represented by a straight line through the simple linear regression equation (SLR_X)[1,2,3]:

$$y = \theta_0 + \theta_1 x_i \qquad (1)$$

Table2. Statistical sample of our data for a population compared to their income.

SLR_X algorithm						
N	X_0	X_1	Y	Description		
0	1	6.1101	17.592		X_1	Y
1	1	5.5277	9.130	count	100.0000	100.0000
2	1	8.5186	13.662	mean	8.159800	5.839135
3	1	7.0032	11.854	std	3.869884	5.510262
4	1	5.8598	6.823	min	5.026900	-2.680700
5	1	8.3829	11.886	max	5.707700	1.986900
6	1	7.4764	4.348	25%	6.589400	4.562300
7	1	8.5781	12.000	50%	8.578100	7.046700
8	1	6.4862	6.5987	75%	22.203000	24.147000

3.2 Multiple linear regression (MLR) :

Multiple linear regression (MLR) is a common statistical technique used to analyze the relationship between a dependent variable and a number of independent variables. This model is based on the idea that there is a linear correlation between the dependent variable and a set of independent variables. In other words, the dependent variable is expressed as a linear combination of independent variables, with the addition of an element that represents the error or randomness in the prediction.

This model is represented by the following formula:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n \qquad (2)$$

The importance of this model stems from its ability to determine the extent to which each independent variable affects the final outcome, making it a powerful analytical tool used in multiple fields such as economics, medicine, engineering, and data science. It is also relied upon to predict future outcomes, discover influential factors, and test statistical hypotheses regarding the importance of input variables.

3.3 Steps of the algorithm (SLR_X):

1- The idea behind this algorithm is to read data and make a prediction based on the previously entered data. This is a type of machine learning approach called supervised learning. Our data in this work is a two-dimensional list, one is the input X and the other the output Y.
2- Display data through a data visualization technique (Scatter plot) that shows the relationship between two numerical variables.
3- To start applying the Gradient Descent Function, we have first separate the data list, and adding a new column called (Ones) before the data representing X_0.
4- Separate X (training data) from y (target variable).
5- Convert from data frames to The matrices.

X = matrix(X.values)
y = matrix(y.values)
theta = matrix(array([0,0]))

6- Cost function calculation: It detects larger errors more than smaller errors, helping the model focus on reducing large errors between predictions and actual values. Using the following formula:

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_0(x^{(i)}) - y^{(i)})^2 \qquad (3)$$

7- Gradient Descent Function: is an optimization algorithm used to minimize the cost function and find the best-fiting line for the model.

Using the following formulas:

$$\theta_0 := \theta_0 - \alpha \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i) \qquad (4)$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m}\sum_{i=1}^{m}((h_\theta(x_i) - y_i)x_i) \qquad (5)$$

8- perform gradient descent to 'fiting' the model parameters.

- get best fiting line.
- draw the line.
- draw error graph.

4- Results and Discussion :

After choosing data from the platform [6] for a population compared to their income, we used a Python program to evaluate simple linear regression model, as we show steps in the system system $LRA_{x,xn}$ describes in Figure 1.

According to the system that was used in our study, the first three steps they are to prepare the simple linear regression model. And the other steps used to minimize the cost function and find the best-fiting line for the model illustrated by Figure 2 . As in our analysis we made two steps of repetitions : the first was 1000 and the second 10000 with the same number of learning rate (0.01) as shown in Figure 3.

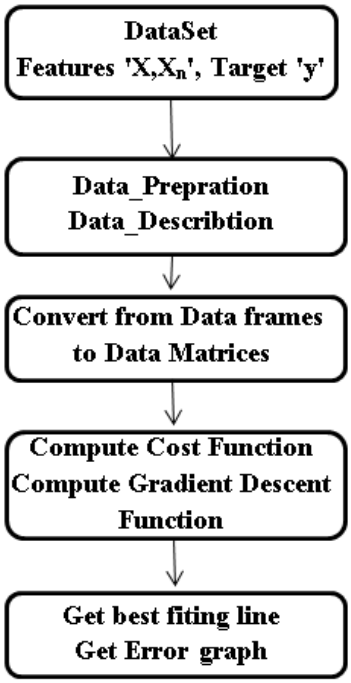

Figure1. System $LRA_{x,xn}$ for Predicted Target and get Error graph .On the results of the quality is based a predicted income vs. people and error vs. training Epoch.

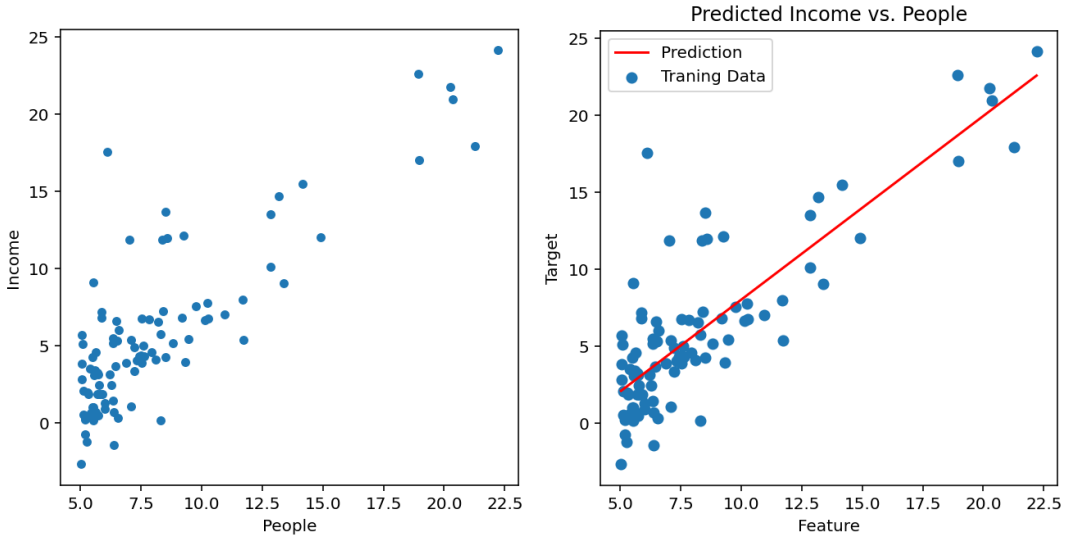

Figure2. The left panel shows the distribution of data and the visual relationship extracted directly from the data file before applying the model, where the horizontal axis represents the population and the vertical axis represents their income. It is noted that the greater the population, the greater the income. . The right panel shows results of The best fiting line to determine the relationship between population and income, where the blue dots (Training Data) represent the original input data, and the red line (Best fiting line), which is basically a representation of the function J(x), shows that the model was able to determine the relationship between population and income.

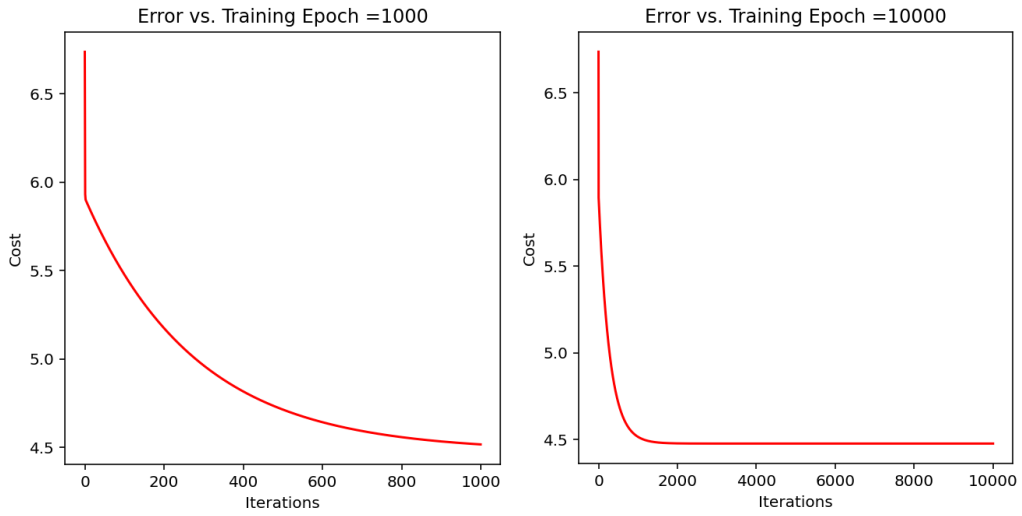

Figure3. The Results shows in the left panel the number of iterations = 1000 and the value of the error function (cost). Right panel shows the number of iterations = 10000.

We compute the equations 3 and 4 and repeated together 1000 times that minimizes the error between the predicted and actual values, this is the goal of the GD equation.

The final results were as the following:

θ_0= -3.89578082 , θ_1= 1.1930336 when the constant Iters=1000.

θ_0= -3.24140214, θ_1= 1.1272942 when the constant Iters=10000.

The following results show a sample of 50 values as calculated by Gradient Descent Function.

Table3. Sample of the first fifty values out of 1000.

6.73719046	5.93159357	5.90115471	5.89522859	5.89009494
5.88500416	5.87993248	5.87487909	5.86984391	5.86482687
5.85982789	5.85484692	5.84988389	5.84493874	5.83510181
5.83020991	5.82533562	5.82047889	5.81563965	5.81081784
5.80601341	5.80122627	5.79645638	5.79170367	5.78696808
5.78224955	5.77754801	5.77286341	5.76819568	5.76354477
5.75891061	5.75429313	5.74969231	5.74510803	5.74054027
5.73598897	5.73145406	5.72693549	5.72243319	5.71794711
5.71347718	5.70902336	5.70458558	5.70016379	5.69575792
5.69136792	5.68699373	5.6826353	5.67829257	5.66902336

The final value of Cost function in the position of iteration 999 is = 4.5159555321, also the final value of Cost function in the position of iteration 9999 is =4.47697132.

We note here through the experiment and the number of repetitions that choosing the number of repetitions of 1000 is very suitable for the data chosen for this study because there is no significant difference in calculating the gradient function as shown in the last values of the function.

Conclusion

We can conclude from this study that machine learning using linear regression is one of the successful methodologies used in machine learning, as we found that it has the ability to provide

us with clear criteria for model optimization and a powerful tool for modeling, prediction, parameter adjustment, pattern extraction, and relationships to achieve the best results. The results showed us that the simple regression algorithm is effective when there is a single independent variable, as well as the effectiveness of using stepwise descent in reducing the error function and improving model accuracy and iterative optimization accuracy. Our experiments also showed how to extract relationships from the data to get effective results. We also found that adjusting the learning rate (α) has a clear and significant impact on the model tool, the smaller the learning rate (0.01) leads to slow convergence but stable when the learning rate is large helps in reducing the time but needs careful selection to avoid overshoots, and increasing the iterations contributed significantly to improving the accuracy of the model as shown in Figure 3. We can say that the study of linear regression is a major starting point for any researcher or student in the field of machine learning.

Recommendation for further works:

Future research will also look at using other control models and other personalized prediction systems to further improve predictions, one such representation is Multiple Linear Regression and try to conduct research and experiments using larger and more complex data .

References

1. Fernandes P, R. Fonseca, P. FAL,& Sanches Fernandes LF(2023): Water quality predictions through linear regression - A brute force algorithm approach. MethodsX. 2023 Mar.
2. J.Hariji(2021): Simple Linear Regression Model and Multiple Linear Regression Model,Research Gate.0.13140/RG.2.2.17237.35044,2021.
3. L.YAN(2024): Predicting House Prices with a Linear Regression Model, Proceedings of the 2nd International Conference on Machine Learning and Automation, 0.54254/2755-2721/114/2024.18220,2024.
4. Z.fang (2024): Application of Linear Regression in GDP Forecasting, Proceedings of the 2nd International Conference on Management Research and Economic Development, 0.54254/2754-1169/72/20240659,2024.
5. Q. Kecheng(2024): Research on linear regression algorithm. MATEC Web of Conferences. 395. 10.1051/matecconf/202439501046. 2024.
6. A. Ahmad & M. Pandey(2025): Application of Linear and Non- Linea Regression in Research. International Books & Periodical Supply Service. ISBN:978-81-19105-38-0, E-ISBN:978-81-19105-44-1, January 2025.
7. Kaggle_platform- reference, Data Source, https://www.kaggle.com/datasets