TIRAMISU RESEARCH
Tiramisu Research Journal on Engineering and Technology 2025
Submission Mar 2025

ICE CREAM SALES PREDICTION BASED ON ENVIROMENT TEMPERATURE USING LINEAR REGRESSION

Moses Sinanta P.W.J.
Universitas Sebelas Maret
mosessinanta@student.uns.ac.id

Abstract

Environment temperature plays an important role in influencing consumer purchasing behaviour, especially in seasonal beverages like ice cream. Understanding this relationship allows businesses to optimize sales strategies to maximize revenue. This research aims to apply Linear Regression to predict ice cream sales profit based on environment temperature. A dataset containing temperature and ice cream sales profit was analysed, and a regression model was developed using the Least Squares Method. The model's performance was evaluated using Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2 score) to assess accuracy. The results suggest a strong relationship between environment temperature and sales, with 98.4% of the variations explained, as well as that with each 1°F increase lead to an about $1.19 rise in sales. The findings strengthens that Linear Regression is an effective tool for predicting sales trends based on temperature variations.

Key words: Linear Regression, predictive modelling, ice cream sales, environment temperature.

INTRODUCTION

Environment temperature plays an important role in shaping consumer behaviour, especially in the food and beverage industry. Hotter environment temperature leads to increased demand for cold beverages, making ice cream a popular choice during hot seasons. Understanding this relationship can help businesses anticipate changes in sales based on environmental conditions.

Accurate sales prediction is important for businesses to maximize revenue [5]. Traditional methods of forecasting often rely on intuition, but data-driven approaches offer more precise and reliable insights [9]. By using historical data and statistical models, businesses can make informed decisions to meet consumer demand effectively.

Linear Regression is a commonly used predictive modelling technique that establishes a relationship between a dependent variable and one or more independent variables [3]. In this research, it provides a simple yet effective framework for understanding how temperature influences ice cream sales. By analysing historical data, Linear Regression can help create a predictive model that estimates future sales based on variables such as environment temperature variations.

This research aims to apply Linear Regression to analyse the correlation between environment temperature and ice cream sales. By developing and evaluating a predictive model, we seek to determine the effectiveness of this approach in forecasting sales trends. The findings of this research can provide valuable insights for businesses looking to improve sales planning and decision-making strategies.

The structure of this research article can be explained as follows. The Theoretical Basis explains Linear Regression and evaluation metrics used in this research, followed by Methodology which explains the data processing and

model application. The Results and Evaluation explains the results from the model based on the evaluation metrics used, followed by the Conclusion section which summarizes the main findings and provides suggestions for further research.

THEORETICAL BASIS

Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables [3]. In its simplest form, Linear Regression is applied when there is only one independent variable influencing the outcome. For example, in this research, temperature serves as the independent variable, while ice cream sales represent the dependent variable.

Linear Regression can be written as:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where y represents the dependent variable, β_0 represents the intercept or the value of y when $x = 0$, β_1 represents the slope, x represents the independent variable, and ε represents the error term.

In predictive modelling, Linear Regression can also be written as:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$$

where $\hat{y}$ represents the predicted dependent variable, $\widehat{\beta_0}$ represents the predicted intercept and $\widehat{\beta_1}$ represents the predicted slope.

In predictive Linear Regression, the coefficients are estimated using the Least Square Method (LMS), an approach to estimate the best-fitting line for a given dataset. It works by minimizing the sum of the squared differences between the actual values and the predicted values from the regression model [4]. This is done by finding the regression coefficients that minimize the Sum of Squared Errors (SSE).

SSE can be written as:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where n represents the number of data. By taking the partial derivatives of SSE with respect to the coefficients and setting them to 0, we obtain the normal equations, which provide the optimal values of the regression coefficients.

Linear Regression operates under several key assumptions to ensure its reliability. These include linearity (a straight-line relationship exists between the dependent and independent variables) [1], the data follows a normal data distribution [6] and homoscedasticity (constant variance of residuals) [8]. Violations of these assumptions may affect the accuracy of the model's prediction.

Root Mean Squared Error (RMSE) is a commonly used metric for evaluating the accuracy of a regression model. It measures the average magnitude of prediction errors, with larger errors being penalized more heavily than smaller ones [7]. A lower RMSE value indicates that the model's predictions are closer to the actual observed values.

RMSE can be written as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

The Coefficient of Determination, or R^2 score, is a statistical measure that explains how well the independent variable accounts for variations in the dependent variable. It ranges $[0,1]$, where a value close to 0 indicates the model does not effectively explain the variance. On the other side, a value close to 1 indicate a strong relationship.

R^2 score can be written as:

$$R^2 \text{ score} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \mu_y)^2}$$

where μ_y represents the average value of y.

METHODOLOGY

The dataset used in this research [2] consists of historical records of environment temperature, and corresponding ice cream sales profit. The environment temperature were measured in Fahrenheit, and the ice cream sales profit were calculated in USD. This dataset provides a basis for training and evaluating a predictive model using Linear Regression.

Before applying the model, data pre-processing is performed to ensure accuracy and reliability. The dataset is checked for missing values, and any inconsistencies are addressed to maintain data quality. After pre-processing, the data is split into training and testing sets using

the 80-20 ratio (80% training set and 20% testing set) to evaluate the model's performance on unseen data.

Linear Regression is then applied to establish a relationship between temperatures and ice cream sales. The model is trained using the training dataset to learn the regression coefficients, which define the best-fitting line. Once trained, the model generates sales predictions based on input environment temperature values.

To evaluate the model's performance, RMSE and R^2 score are used. RMSE measures the average prediction error (in USD). On the other hand, R^2 score measures how well environment temperature explains sales variations.

RESULTS AND EVALUATION

The Linear Regression model derived the following equation to predict ice cream sales based on environment temperature:

$$\hat{y} = 1.190x - 33.661$$

The intercept of -33.661 suggests that at extremely low temperature, ice cream sales would be negative. This is unrealistic, but expected due to the model's linear nature. The slope of 1.190 suggests that for every 1°F increase in temperature, ice cream sales increase by about $1.19.

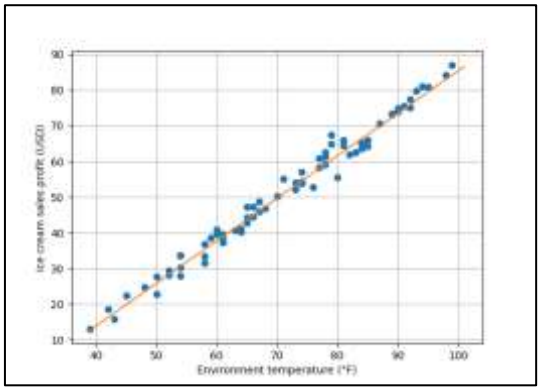

Figure 1 Ice cream sales profit vs environment temperature with Linear Regression model. The blue scatter represents the testing data points, and the orange line represents the predicted equation.

The model's performance was evaluated using RMSE and R^2 score. The RMSE of 2.301 suggests that, on average, the model's predictions deviate from actual sales by about $2.30, which is a relatively low error. The R^2 score of 0.984 suggests that 98.4% of the variation in ice cream sales is explained by the temperature, which shows a strong relationship.

Comparing the model's predictions to actual sales values, the Linear Regression closely follows the real data points. The small residuals indicate that the model makes highly accurate predictions, though minor deviations still exist. This is most likely due to external factors not included in the dataset or model.

Overall, the results suggest that environment temperature is a strong predictor of ice cream sales, with nearly perfect linear relationship. The regression equation shows that as environment temperature increases, sales consistently rise. This strengthens the idea that hotter weather drives higher demand for ice cream.

CONCLUSION

This research analysed the relationship between environment temperature and ice cream sales using Linear Regression, demonstrating a strong correlation between the two variables. The model achieved an RMSE of 2.301, showing accurate predictions, followed by R^2 score of 0.984, showing that environment temperatures explains most of the variation in sales. These findings suggest that businesses can effectively use environment temperature data to forecast sales and maximize revenue.

Future research could explore additional factors such as weekends, holidays and promotional events could be incorporated to improve the model's predictive accuracy. Using Multiple Linear Regression or other Machine Learning techniques like Decision Trees and Random Forests could further enhance performance by capturing more complex patterns in sales data. Expanding the dataset to include different geographic locations and seasonal variations would also provide a more comprehensive understanding of ice cream sales trends.

APPENDIX

The author would like to thank Noelle, who has given the author a lot of inspiration and has become a figure that the author admired.

This research documentation can be found at the Tiramisu Research website.

REFERENCES

[1] Lim, Chun Yee, et al. (2024). "Linearity Assessment: Deviation from Linearity and Residual of Linear Regression Approaches." *Clinical Chemistry and Laboratory Medicine (CCLM)*, 62(10), 1918-1927.

[2] Rephy. (2024). "Temperature and Ice Cream Sales." Accessed on 16 March 2025 from www.kaggle.com.

[3] James, Gareth, et al. (2023). "Linear Regression." *An Introduction to Statistical Learning: With Applications in Python*, 69-134. Cham: Springer International Publishing.

[4] Ding, Feng. (2023). "Least Squares Parameter Estimation and Multi-Innovation Least Squares Methods for Linear Fitting Problems from Noisy Data." *Journal of Computational and Applied Mathematics*, 426, 115107.

[5] Haselbeck, Florian, et al. (2022). "Machine Learning Outperforms Classical Forecasting on Horticultural Sales Predictions." *Machine Learning with Applications*, 7, 100239.

[6] Knief, Ulrich, and Wolfgang Forstmeier. (2021). "Violating the Normality Assumption may be the Lesser of Two Evils." *Behaviour Research Methods*, 53(6), 2576-2590.

[7] Hodson, Timothy O., Thomas M. Over, and Sydney S. Foks. (2021). "Mean Squared Error, Deconstructed." *Journal of Advances in Modelling Earth Systems*, 13(12), e2021MS002681.

[8] Đalić, Irena, and Svetlana Terzić. (2021). "Violation of the Assumption of Homoscedasticity and Detection of Heteroscedasticity." *Decision Making: Applications in Management and Engineering*, 4(1), 1-8.

[9] Yin, Xiaoting, and Xiaosha Tao. (2021). "Prediction of Merchandise Sales on E-Commerce Platforms Based on Data Mining and Deep Learning." *Scientific Programming*, 2021(1), 2179692.