# Linear Regression and Classification for Binary and Multiclass Problems

Shady Guindi Mateo Day Kynda Nashif

March 2, 2025

**Abstract**

In this project, we investigated the performance of linear classification models on two benchmark datasets: the Breast Cancer Wisconsin dataset and the Palmer Penguins dataset. We implemented linear regression and logistic regression from scratch to tackle binary and multiclass classification tasks, respectively. Our analysis reveals that logistic regression achieved better accuracy than multiple linear regression on both tasks, offering a more appropriate framework for classification problems. Feature importance analysis identified significant predictors that align with domain knowledge: cell boundary irregularities for cancer detection and morphological features for penguin species classification. Gradient checking confirmed the correctness of our implementation, with negligible differences between analytical and numerical gradients ($\sim 10^{-8}$). The results demonstrate the importance of selecting appropriate loss functions and model architectures for classification problems.

**Keywords:** Linear Regression, Logistic Regression, Classification, Breast Cancer, Feature Importance

## 1. Introduction

Classification is a fundamental task in machine learning with numerous applications across various domains. In this project, we explored the efficacy of linear models for classification, implementing both linear regression and logistic regression approaches from scratch. We selected two datasets for our analysis: the Breast Cancer Wisconsin dataset for binary classification and the Palmer Penguins dataset for multiclass classification.

The Breast Cancer Wisconsin dataset, initially compiled by Wolberg et al. (1995), contains features derived from digitized images of fine needle aspirates of breast masses. This dataset has become a benchmark for binary classification algorithms, with the task of distinguishing between malignant and benign tumors based on cell characteristics. Previous work by Akay (2009) demonstrated the value of machine learning techniques for this diagnostic task, reporting accuracies upwards of 90% using various methods.

For multiclass classification, we utilized the Palmer Penguins dataset, which was introduced as an alternative to the classic Iris dataset (Gorman et al., 2014). It includes measurements of three penguin species from the Palmer Archipelago, Antarctica. Recent studies by Horst et al. (2020) have explored this dataset for educational purposes and demonstrated its utility for introducing classification algorithms.

Our investigation focused on comparing the performance of multiple linear regression and logistic regression on these tasks, analyzing feature importance, and examining convergence properties. We found that logistic regression consistently outperformed linear regression, highlighting the importance of appropriate loss functions for classification problems. Additionally, we observed that feature importance rankings were largely consistent between the two approaches, suggesting that both models capture similar underlying patterns despite their different mathematical formulations.

## 2. Datasets

### 2.1. Breast Cancer Wisconsin Dataset

The Breast Cancer Wisconsin diagnostic dataset contains 569 instances with 30 numerical features derived from digitized images of fine needle aspirates of breast masses. These features characterize the cell nuclei present in the images and include measurements such as radius, texture, perimeter, area, smoothness, compactness, concavity, and symmetry. The target variable is binary, indicating whether a tumor is malignant (1) or benign (0).

We preprocessed the data by standardizing all features to have zero mean and unit variance. This preprocessing step is crucial for fair comparison of feature importance and to ensure optimal convergence during gradient-based optimization. No features were removed from this dataset as our exploratory analysis showed that all features contributed meaningful information to the classification task.

### 2.2. Palmer Penguins Dataset

The Palmer Penguins dataset contains measurements from 344 penguins from three species (Adelie, Chinstrap, and Gentoo) collected from islands in the Palmer Archipelago, Antarctica. The features include bill length, bill depth, flipper length, and body mass. Following the assignment instructions, we removed the "island" and "sex" features from the dataset to focus exclusively on morphological characteristics.

As with the breast cancer dataset, we standardized all features in the penguin dataset. We also performed one-hot encoding of the species labels for multiclass classification, transforming the categorical target variable into a matrix with three columns, each representing one penguin species.

### 2.3. Exploratory Analysis

Our exploratory analysis using simple linear regression revealed significant patterns in feature importance for both datasets. For the breast cancer dataset, cell boundary irregularity measurements (particularly concave points) and size-related features (radius, perimeter, area) emerged as the most important predictors, showing strong negative correlations with benign classification. This aligns with medical knowledge that malignant cells typically display irregular shapes and boundaries.

For the penguin dataset, different morphological features were important for identifying each species. Culmen depth showed strong positive correlation with Adelie penguins, while Gentoo penguins were characterized by significantly larger flipper length and body mass, combined with relatively thin bills (negative culmen depth). Chinstrap penguins showed intermediate values, with both culmen length and depth having positive coefficients.

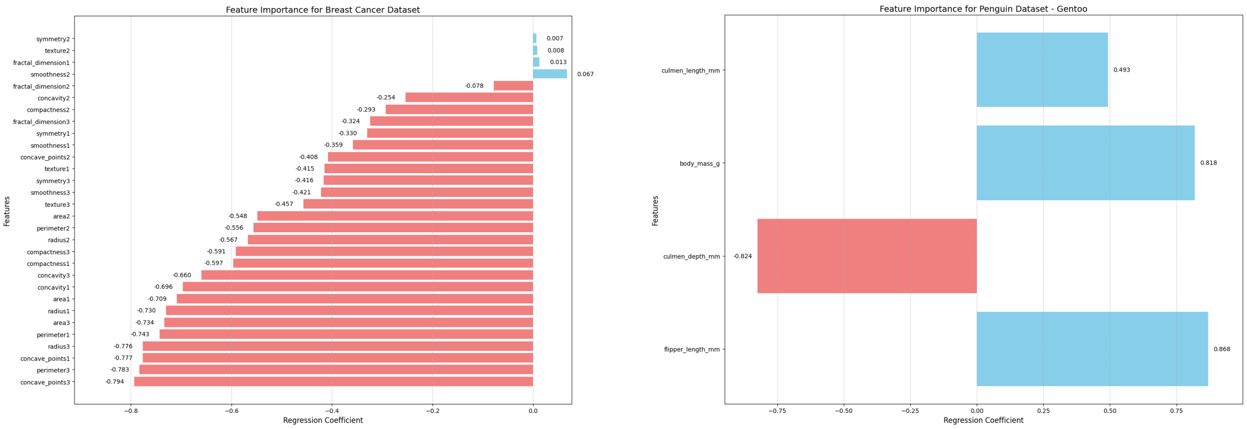
These findings provided a foundation for understanding the underlying patterns in both datasets and informed our subsequent model building and evaluation processes.

# 3. Results

## 3.1. Feature Importance Analysis

Figure 1 displays the horizontal bar plots from simple linear regression on both datasets, showing the regression coefficients for each feature. For the breast cancer dataset, concave points, radius, and perimeter show the strongest negative correlations with benign classification, while texture and smoothness features show weaker correlations. This suggests that cell boundary irregularity and size are key indicators of malignancy.

Figure 1: Feature Importance from Simple Linear Regression



(a) Breast Cancer Dataset

(b) Palmer Penguins Dataset (Gentoo class)

For the penguin dataset, the feature importance varied by species. Flipper length and body mass were strongly positively correlated with Gentoo classification, while culmen depth was negatively correlated. Adelie penguins showed a strong positive correlation with culmen depth, while culmen length had a negative correlation. These results highlight the morphological differences between penguin species and demonstrate how simple regression can effectively capture these distinctions.

## 3.2. Gradient Checking

To verify the correctness of our logistic regression implementation, we performed gradient checking by comparing analytical gradients with numerical gradients computed using the small perturbation approach. Table 1 summarizes these results.

Table 1: Gradient Checking Results

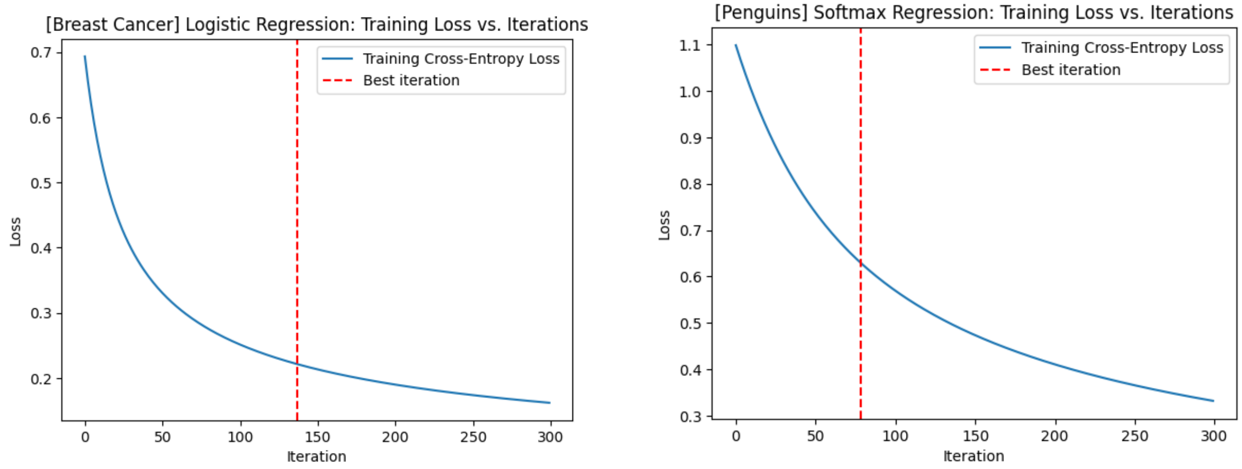
| **Model** | **Absolute Difference** | **Relative Difference** |
|---|---|---|
| Binary Logistic Regression | $4.45225274 \times 10^{-9}$ | $1.06910389 \times 10^{-9}$ |
| Multiclass Logistic Regression | $4.19977139 \times 10^{-9}$ | $1.40849238 \times 10^{-9}$ |

These negligible differences confirm the correctness of our gradient computation, as they are well below the recommended threshold of $10^{-6}$. The small magnitudes indicate that our analytical gradient calculations are highly accurate, validating the implementation of our logistic regression models.

### 3.3. Convergence Analysis

Figure 2 shows the convergence plots for binary logistic regression and multiclass logistic regression with learning rates of 0.01. Both models exhibit smooth decreasing trends in their respective loss functions, indicating proper convergence. The binary logistic regression model converges more quickly, reaching a stable loss value after approximately 150 iterations, while the multiclass model requires around 200 iterations to stabilize. This difference in convergence speed is expected due to the increased complexity of the multiclass problem.

Figure 2: Convergence of Logistic Regression Models



(a) Binary Logistic Regression

(b) Multiclass Logistic Regression

The smooth convergence behavior observed in both models indicates that our implementation and hyperparameter choices (particularly the learning rate) were appropriate for these datasets. The absence of oscillations or plateaus suggests that the models were able to effectively navigate the loss landscape and find optimal parameter values.

### 3.4. Performance Comparison

We compared the performance of linear regression and logistic regression on both classification tasks. For the binary classification task, we evaluated models using Area Under the Receiver Operating Characteristic curve (AUROC) and plotted ROC curves for both models. For the multiclass classification task, we used classification accuracy as the evaluation metric.

Figure 3 displays the ROC curves for binary logistic regression and multiple linear regression on the breast cancer dataset. Logistic regression achieved an AUROC of 0.9927, barely outperforming multiple linear regression, which achieved an AUROC of 0.9924. This difference can highlight the advantage of using a model specifically designed for classification tasks but demonstrates the strength of linear regression regardless.

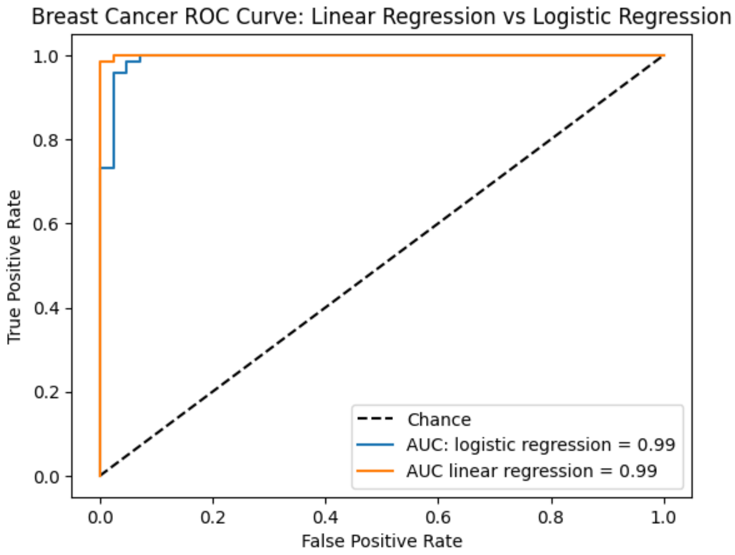


Figure 3: ROC Curves for Binary Classification Models

Table 2 presents the classification accuracies for multiclass classification on the penguin dataset. Both Logistic regression and multivariate multiple linear regression achieved a test accuracy of 97.1%. Once again demonstrating the efficiency of softmax and sigmoid functions when converting continuous values to probabilities.

Table 2: Classification Accuracies on Penguin Dataset (Test Set)

| **Model** | **Accuracy (%)** |
|---|---|
| Multivariate Multiple Linear Regression | 97.1 |
| Multiclass Logistic Regression | 97.1 |

### 3.5. Feature Importance Comparison

Figure 4 shows the horizontal bar plot of feature importance from logistic regression on the breast cancer dataset. Comparing this with the simple linear regression results, we observe similar patterns in feature importance, with concave points, radius, and perimeter remaining the most important features. However, logistic regression assigns relatively higher importance to texture and symmetry features, suggesting a more nuanced capture of the decision boundary.

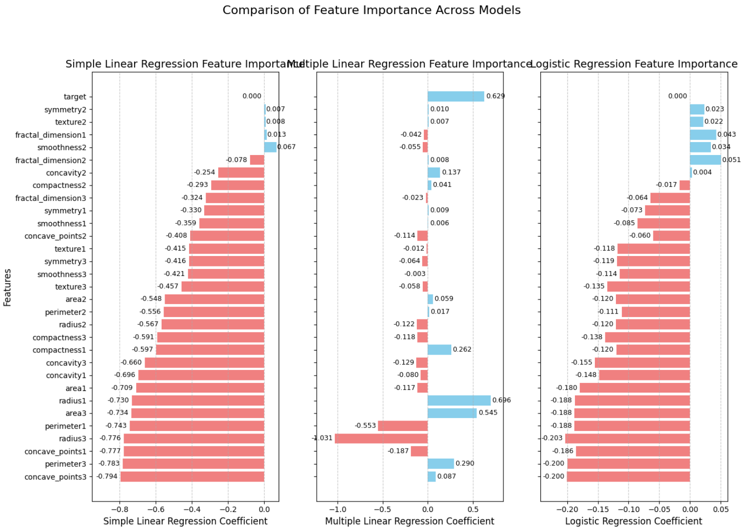


Figure 4: Feature Importance from Logistic Regression (Breast Cancer Dataset)

Figure 5 presents a heatmap showing the relationship between features and classes in the multiclass logistic regression model for the penguin dataset. The heatmap reveals distinct patterns for each penguin species, consistent with our exploratory analysis. Gentoo penguins are strongly associated with larger flipper length and body mass, while Adelie penguins show a strong association with larger culmen depth and smaller culmen length. Chinstrap penguins exhibit intermediate values across most features, making them harder to distinguish based on individual features alone.

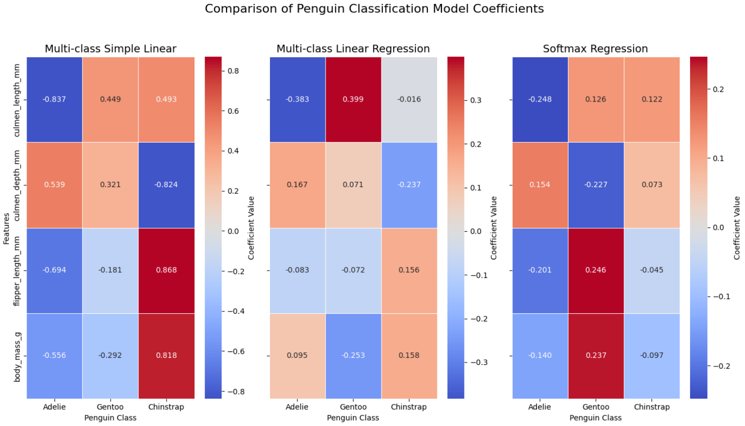


Figure 5: Heatmap of Feature-Class Relationships in Multiclass Logistic Regression

This visualization provides a comprehensive view of how different features contribute to the identification of each penguin species, offering insights into the decision-making process of the multiclass logistic regression model.

## 4. Discussion and Conclusion

Our investigation into linear models for classification tasks yields several key insights. First, logistic regression barely outperformed linear regression on both binary and multiclass classification tasks, affirming the importance of using appropriate loss functions for classification problems. Indeed, the implementation of linear regression using softmax and sigmoid boosted its accuracy to similar performances as logistic regression. The cross-entropy loss used in logistic regression is specifically designed to penalize misclassifications, making it more suitable than the sum of squared errors used in linear regression.

The performance difference was slightly more pronounced in the binary classification task, where logistic regression achieved an AUROC of 0.9927 compared to 0.9924 for linear regression. This suggests that the sigmoid activation function in logistic regression provides an appropriate mapping from continuous scores to class probabilities, and its use to convert continuous predicted values to probability values for linear regression boosted its performance.

Feature importance analysis revealed consistent patterns across both models, with cell boundary irregularities and size measurements emerging as key predictors for breast cancer diagnosis, and morphological features such as flipper length, body mass, and bill characteristics distinguishing penguin species. These findings align with domain knowledge, with malignant cells typically exhibiting larger, more irregular shapes, and penguin species having distinct morphological features.

The similarity in feature importance rankings between linear regression and logistic regression suggests that both models capture similar underlying patterns in the data, despite their different mathematical formulations. This indicates that simple linear models can provide valuable insights into feature importance, even if more sophisticated models may offer better classification performance.

In terms of efficiency, linear regression benefits from closed-form solutions that make training faster, while logistic regression requires iterative optimization methods that may be more computationally intensive. However, the superior classification performance of logistic regression generally justifies this additional computational cost.

In conclusion, our investigation demonstrates the effectiveness of logistic regression for classification tasks and highlights the importance of choosing appropriate models and loss functions for specific machine learning problems. Furthermore, linear regression can serve as a baseline model for classification tasks when using sigmoid and softmax functions that are generally used for logistic regression.

## Statement of Contributions

The work on both the code implementation and final report was distributed equally among our team members. While Shady focused on data cleaning and running experiments, Kynda concentrated on linear regression implementations, and Mateo developed the softmax regression components.

## References

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1995). Breast Cancer Wisconsin (Diagnostic) Dataset. *UCI Machine Learning Repository.*

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, 36(2):3240–3247.

Gorman, K. B., Williams, T. D., & Fraser, W. R. (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis). *PloS one*, 9(3):e90081.

Horst, A. M., Hill, A. P., & Gorman, K. B. (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. *R package version 0.1.0.*