Logistic Regression for BMI Classification in Women Dataset

Simon Green

December 26, 2025

Introduction

This experiment implements logistic regression to classify Body Mass Index (BMI) categories using the built-in *women* dataset which includes height (inches) and weight (pounds) for 15 individuals as key input features (McNeil, 1977; James et al., 2013). BMI is calculated and binarized into low ($BMI < 23$, label 0) or high ($BMI \geq 23$, label 1) categories, creating a balanced split (10 low, 5 high) (CDC, 2025). The aim is to assess if a linear classifier can separate these categories effectively in the feature space, with expected high accuracy given the low dimensionality and direct BMI relationship.

Methodology

Data preprocessing computed BMI via the formula $BMI = \frac{\text{weight} \times 703}{\text{height}^2}$, and assigned labels at a threshold. The dataset was shuffled with a fixed seed for reproducibility, preventing order-based biases, and split 70/30 (11 training, 4 test samples) to balance learning and evaluation on limited data (James et al., 2013). Logistic regression was selected over neural networks for its simplicity, interpretability, and suitability for binary outcomes with few features, implemented via *glm* function with binomial family. This models the log-odds of high BMI as a linear function of height and weight, providing coefficients that reveal feature impacts. Predictions thresholded probabilities at 0.5 for classification, justifying the choice as it yields probabilistic outputs for better uncertainty assessment than hard classifiers.

Results

The model achieved 100% training accuracy but 75% on the test set, indicating strong fit to seen data yet moderate generalization, likely due to the tiny sample size amplifying variance. Coefficients showed a large positive intercept (2845.67), negative height effect (-92.26), and positive weight effect (23.00), with extreme standard errors signaling perfect separation in training—where the linear boundary fully divides classes, causing estimates to approach infinity. Suggest that this experiment is a good exercise in methods, but is not realistic in terms of real population but rather representative of sample data and is overfitted. Precision and recall were perfect on training but imply potential false positives/negatives on test, underscoring limitations in small datasets. Future improvements could include regularization with *glmnet* to curb separation issues or cross-validation for better hyperparameter tuning. Overall, the experiment confirms logistic regression's efficacy as an interpretable baseline for health metrics, meeting the introduction's goals while revealing scale-related issues.

References

CDC (2025). About adult bmi. Accessed: 2025-12-26.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R.* Springer, New York, NY, 1 edition.

McNeil, D. R. (1977). *Interactive data analysis: A practical primer.* John Wiley & Sons, New York.