

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.

Appendix Code

```
# Code generated with Grok
# Load the women dataset
data(women)

# Calculate BMI
women$bmi <- (women$weight * 703) / (women$height ^ 2)

# Assign binary labels: 0 for BMI < 23 (low), 1 for BMI >= 23 (high)
women$label <- ifelse(women$bmi < 23, 0, 1)

# Set seed for reproducibility
set.seed(42)

# Shuffle indices
indices <- sample(1:nrow(women))

# 70/30 train-test split (11 train, 4 test)
train_indices <- indices[1:11]
test_indices <- indices[12:15]

train_data <- women[train_indices, ]
test_data <- women[test_indices, ]

# Train logistic regression model
model <- glm(label ~ height + weight, data = train_data, family = binomial)

# Predict on training set
train_pred_prob <- predict(model, train_data, type = "response")
train_pred <- ifelse(train_pred_prob > 0.5, 1, 0)

# Training accuracy
train_accuracy <- mean(train_pred == train_data$label)
cat("Training Accuracy:", train_accuracy, "\n")

# Predict on test set
test_pred_prob <- predict(model, test_data, type = "response")
test_pred <- ifelse(test_pred_prob > 0.5, 1, 0)

# Test accuracy
test_accuracy <- mean(test_pred == test_data$label)
cat("Test Accuracy:", test_accuracy, "\n")

# Model summary for coefficients and interpretation
summary(model)
```

Appendix B Outputs

Training Accuracy: 1

Test Accuracy: 0.75

```
glm(formula = label ~ height + weight, family = binomial, data = train_data)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2845.67	6497124.41	0	1
height	-92.26	225294.59	0	1
weight	23.00	60207.54	0	1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1.4421e+01 on 10 degrees of freedom

Residual deviance: 3.3249e-10 on 8 degrees of freedom

AIC: 6

Number of Fisher Scoring iterations: 25

Appendix C R Studio

The screenshot shows the R Studio interface. The left pane contains the R script editor with the following code:

```
1 # Code generated with Grok
2 # Load the women dataset
3 data(women)
4
5 # Calculate BMI
6 women$bmi <- (women$weight * 703) / (women$height ^ 2)
7
8 # Assign binary labels: 0 for BMI < 23 (low), 1 for BMI >= 23 (high)
9 women$label <- ifelse(women$bmi < 23, 0, 1)
10
11 # Set seed for reproducibility
12 set.seed(42)
13
14 # Shuffle indices
15 indices <- sample(1:nrow(women))
16
17 # 70/30 train-test split (11 train, 4 test)
18 train_indices <- indices[1:11]
19 test_indices <- indices[12:15]
20
21 train_data <- women[train_indices, ]
22 test_data <- women[test_indices, ]
23
24 # Train logistic regression model
25 model <- glm(label ~ height + weight, data = train_data, family = binomial)
26
27 # Predict on training set
28 train_pred_prob <- predict(model, train_data, type = "response")
29 train_pred <- ifelse(train_pred_prob > 0.5, 1, 0)
30
31 # Training accuracy
32 train_accuracy <- mean(train_pred == train_data$label)
33 cat("Training Accuracy:", train_accuracy, "\n")
34
35 # Predict on test set
36 test_pred_prob <- predict(model, test_data, type = "response")
37 test_pred <- ifelse(test_pred_prob > 0.5, 1, 0)
38
39 # Test accuracy
40 test_accuracy <- mean(test_pred == test_data$label)
41 cat("Test Accuracy:", test_accuracy, "\n")
42
43 # Model summary for coefficients and interpretation
44 summary(model)
```

The right pane shows the console output:

```
> train_data <- women[train_indices, ]
> test_data <- women[test_indices, ]
> # Train logistic regression model
> model <- glm(label ~ height + weight, data = train_data, family = binomial)
> # Predict on training set
> train_pred_prob <- predict(model, train_data, type = "response")
> train_pred <- ifelse(train_pred_prob > 0.5, 1, 0)
> # Training accuracy
> train_accuracy <- mean(train_pred == train_data$label)
> cat("Training Accuracy:", train_accuracy, "\n")
Training Accuracy: 1
> # Predict on test set
> test_pred_prob <- predict(model, test_data, type = "response")
> test_pred <- ifelse(test_pred_prob > 0.5, 1, 0)
> # Test accuracy
> test_accuracy <- mean(test_pred == test_data$label)
> cat("Test Accuracy:", test_accuracy, "\n")
Test Accuracy: 0.75
> # Model summary for coefficients and interpretation
> summary(model)
Call:
glm(formula = label ~ height + weight, family = binomial, data = train_data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2845.67 6497124.41      0      1
height       -92.26 225294.59      0      1
weight        23.00  60207.54      0      1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1.4421e+01 on 10 degrees of freedom
Residual deviance: 3.3249e-10 on 8 degrees of freedom
AIC: 6

Number of Fisher Scoring iterations: 25

Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
>
```

Figure 1: R Studio Environment with Code and Output