Modeling Classification

HI 743

Ryan Gallagher

Department of Health Informatics and Administration
Zilber College of Public Health
University of Wisconsin - Milwaukee

February 27, 2025

Overview

1. Classification Intro

2. Logistic Regression

2.1 Logistic Regression
2.2 Interpretation
2.3 Output
2.4 Model Evaluation

3. Multiple Logistic Regression

3.1 Multiple Logistic Regression
3.2 Example
3.3 Interpretation

4. Multinomial Logistic Regression

4.1 Multinomial Logistic Regression
4.2 Example
4.3 Interpretation

What is Classification?

Classification is a predictive modeling task where the **response variable is categorical**.

- Unlike regression, classification predicts discrete labels instead of continuous values.
- Example: Predicting whether a patient has a disease (Yes/No) based on symptoms.
- Applications: Medical diagnosis, fraud detection, spam filtering, etc.

Examples of Classification Problems:

- **Medical Diagnosis:** Does a patient have diabetes? (Yes/No)
- **Fraud Detection:** Is a bank transaction fraudulent? (Fraud/Not Fraud)
- **Genetics:** Classifying a genetic mutation as harmful or benign.
- **Customer Churn:** Will a customer continue using a service? (Stay/Leave)

Example: Default Data Set

Objective: Predict whether an individual will default on a credit card payment.

- Response Variable: `Default` (Yes/No)
- Predictors: Income, Balance, Student Status

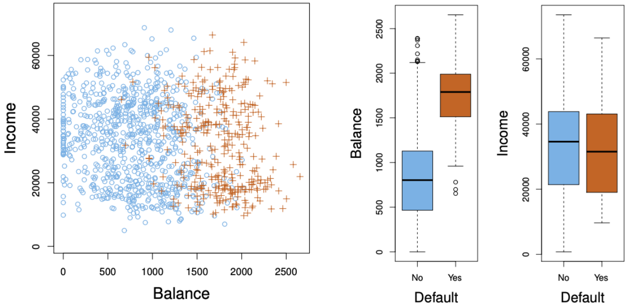


Figure: Credit card default dataset visualization.

Why Not Linear Regression?

- Linear regression assumes a continuous response variable, making it unsuitable for predicting discrete class labels.
- Predictions from a linear model can extend beyond the valid probability range of $[0, 1]$, which is problematic for classification.

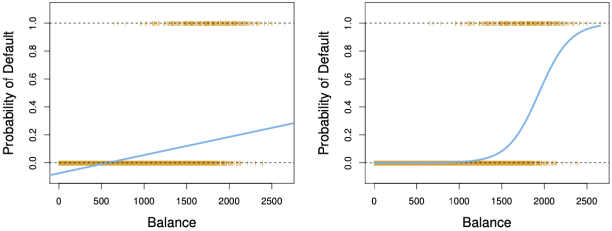


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default` *(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Logistic Regression

Logistic regression models the probability that a given observation belongs to a particular class. The logistic function ensures predictions stay within $[0, 1]$:

$$P(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{1}$$

Unlike linear regression, logistic regression predicts probabilities instead of continuous values. $P(Y = 1|X)$ reads as "The probability of Y equal to 1 given X".

Logistic Regression

The model uses the log-odds (logit) transformation to relate predictors to the probability of an event occurring:

$$log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right)=\beta_0+\beta_1 X \tag{2}$$

Maximum likelihood estimation (MLE) is used to estimate the coefficients, rather than least squares. The coefficients indicate how much the *log-odds* of the outcome change with a one-unit increase in the predictor.

Interpreting Logistic Regression

- A positive coefficient β_1 implies an increase in the predictor increases the probability of the event.
- A negative coefficient implies a decrease in probability.
- The odds ratio e^{β_1} represents how the odds change with a one-unit increase in the predictor.
- Example: If $\beta_1 = 0.3$, then $e^{0.3} \approx 1.35$, meaning the odds increase by 35%.

Understanding Odds and Odds Ratios

- **Odds:** The odds of an event occurring is defined as:

$$\text{Odds} = \frac{P(Y=1)}{1 - P(Y=1)} \tag{3}$$

- If $P(Y=1) = 0.75$, then the odds are $\frac{0.75}{0.25} = 3$, meaning the event is three times as likely to occur than not.

Understanding Odds and Odds Ratio

- **Odds Ratio (OR):** Compares the odds of an event occurring for different values of a predictor variable.
- The odds ratio is given by:

$$\mathrm{OR} = \frac{P(Y = 1|X + 1)/(1 - P(Y = 1|X + 1))}{P(Y = 1|X)/(1 - P(Y = 1|X))} \tag{4}$$

- If $OR > 1$, the event is more likely as the predictor increases; if $OR < 1$, the event is less likely.
- Example: If $OR = 2.5$, the event is 2.5 times more likely for each unit increase in the predictor.

Understanding Odds and Odds Ratio

		OUTCOME	
		Disease (Case)	No Disease (Controls)
EXPOSURE	Exposed	a	b
	Unexposed	c	d

$$Odds\ of\ Exposure\ in\ Cases = \frac{Number\ of\ Cases\ with\ Exposure}{Number\ of\ Cases\ without\ Exposure} = \frac{a}{c}$$

$$Odds\ of\ Exposure\ in\ Controls = \frac{Number\ of\ Controls\ with\ Exposure}{Number\ of\ Controls\ without\ Exposure} = \frac{b}{d}$$

$$Odds\ Ratio = \frac{Odds\ of\ Exposure\ in\ Cases}{Odds\ of\ Exposure\ in\ Controls} = \frac{a/c}{b/d} = \frac{a * d}{b * c}$$

Logistic Regression Model Output

Fitted Model:

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = -5.2 + 0.04X \tag{5}$$

Interpretation:

- Intercept (-5.2): Baseline log-odds of diabetes when glucose $= 0$.
- Coefficient (0.04): A one-unit increase in glucose increases log-odds of diabetes by 0.04.
- Normal glucose levels range between $70 - 99$, meaning individuals within this range generally have lower predicted probabilities of diabetes.

Calculating Predicted Probabilities

Example: Patient with Glucose Level = 150

$$\text{Log-Odds} = -5.2 + (0.04 \times 150) = 0.8 \tag{6}$$

$$\text{Probability} = \frac{e^{0.8}}{1 + e^{0.8}} \approx 0.69 \tag{7}$$

Interpretation: This patient has a 69% probability of having diabetes.

- A patient with glucose level $X = 85$ (mid-normal range) would have:

$$\text{Log-Odds} = -5.2 + (0.04 \times 85) = -1.8 \tag{8}$$

$$\text{Probability} = \frac{e^{-1.8}}{1 + e^{-1.8}} \approx 0.14 \tag{9}$$

- Meaning, a patient in the normal glucose range has a much lower probability of diabetes.

Thresholding for Classification

- A threshold (e.g., 0.5) is applied to classify patients.
- If $P(Y = 1|X) > 0.5$, classify as diabetic ($Y = 1$).
- If $P(Y = 1|X) \leq 0.5$, classify as non-diabetic ($Y = 0$).
- Example: With a probability of 0.69, the patient is classified as diabetic.
- Patients within the normal glucose range typically have probabilities well below 0.5, supporting non-diabetic classification.

Classification Model Evaluation - Basic Metrics

Metrics for Evaluating Model Performance:

- **Accuracy:** Measures the proportion of correctly classified cases.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

- **Precision:** Measures how many predicted positives are actually correct.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

- **Recall (Sensitivity):** Measures how many actual positives were correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{12}$$

Classification Model Evaluation - Advanced Metrics

Additional Performance Measures:

- **F1-Score:** Harmonic mean of precision and recall, balancing false positives and false negatives.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

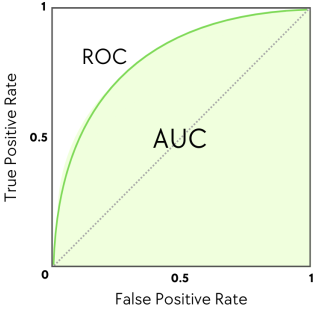
- **Specificity:** Measures the proportion of true negatives correctly identified.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{14}$$

AUC-ROC Curve for Classification

AUC-ROC Curve: Measures how well the model distinguishes between classes.

- The Area Under the Curve (AUC) quantifies performance:
    - $\text{AUC} = 1$: Perfect classifier.
    - $\text{AUC} = 0.5$: Random guessing.
- The ROC curve plots the true positive rate (recall) against the false positive rate at different threshold values.



Multiple Logistic Regression

Extending Logistic Regression to Multiple Predictors

- In multiple logistic regression, we model the probability of a binary outcome using multiple predictors.
- The model equation is:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{15}$$

- This allows us to account for multiple factors simultaneously when predicting an outcome.
- Model evaluation methods are the same as from simple Logistic Regression.

Example: Predicting Credit Default

Using Multiple Predictors:

- Consider a dataset where we predict whether a person defaults on a loan.
- Predictors: Credit balance, income, and student status.
- Fitted model:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = -10.869 + 0.0057 \times \text{Balance} + 0.0030 \times \text{Income} - 0.6468 \times \text{Student} \tag{16}$$

- Interpretation:
  - A one-unit increase in balance increases the log-odds of default.
  - Being a student decreases the probability of default, holding other factors constant.

Interpreting Coefficients

Understanding the Impact of Each Predictor:

- **Odds Ratio:** The exponentiated coefficient e^{β} represents the multiplicative change in the odds.
- Example:

$$e^{0.0057} \approx 1.0057 \tag{17}$$

- This means that for each additional dollar in balance, the odds of default increase by approximately 0.57%.
- Similarly, a student has lower odds of default by a factor of:

$$e^{-0.6468} \approx 0.523 \tag{18}$$

- This means students are about 48% less likely to default than non-students, controlling for other factors.

Multinomial Logistic Regression

Extending Logistic Regression to More than Two Classes (Outcomes)

- When the response variable has more than two categories, multinomial logistic regression is used.
- Unlike binary logistic regression, we model multiple class probabilities simultaneously.
- The model equation for class k is:

$$P(Y = k|X) = \frac{e^{\beta_{k0}+\beta_{k1}X_1+\cdots+\beta_{kp}X_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}X_1+\cdots+\beta_{lp}X_p}} \tag{19}$$

Example: Predicting Disease Type

Classifying Patients into One of Three Conditions:

- Response Variable: Y (Disease type: Stroke, Drug Overdose, Epileptic Seizure).
- Predictors: Age, Blood Pressure, Medical History.
- A fitted model:

$$\log\left(\frac{P(Y = \text{Stroke}|X)}{P(Y = \text{Epileptic Seizure}|X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{20}$$

- Interpretation:
  - A one-unit increase in X_1 (e.g., blood pressure) changes the log-odds of having a stroke relative to an epileptic seizure.
  - Probabilities for all categories sum to one.

Interpreting Coefficients

Understanding the Impact of Each Predictor:

- Coefficients describe log-odds of one category relative to the baseline category.
- The exponentiated coefficient e^{β} represents the change in odds for a one-unit increase in the predictor.
- Example:

$$e^{0.3} \approx 1.35 \tag{21}$$

- This means that for each unit increase in the predictor, the odds of belonging to a given class (e.g., Stroke) relative to the baseline increase by 35%.