Section 8: Classification with R

STA 141A – Fundamentals of Statistical Data Science

Instructor: Akira Horiguchi

Fall Quarter 2025 (Sep 24 – Dec 12)
MWF, 9:00 AM – 9:50 AM, TLC 1215
University of California, Davis

Overview

Based on Chapter 4 of ISL book James et al. (2021).

- For more R code examples, see R Markdown files in `https://www.statlearning.com/resources-second-edition`

1. Why not linear regression?

2. Odds and log odds

Example (two categories)

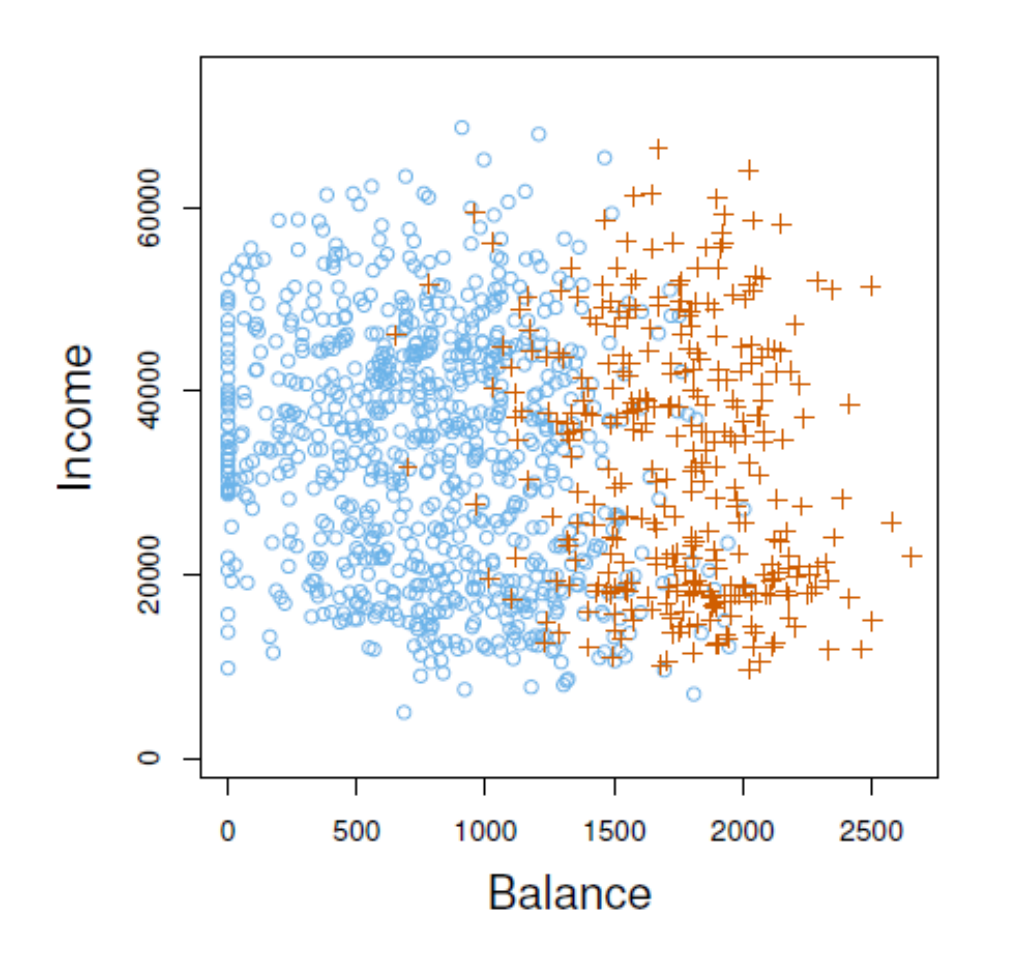

Figure 1: Image by James et al. (2021), based on the `Default` data set in R. The annual incomes and monthly credit card balances of a number of individuals, where the individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.

Examples

What are the predictors and responses in each example?

1. A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of these medical conditions does the person have based on the symptoms given?

2. An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.

3. On the basis of DNA sequence data for a number of patients with and without a given disease, one would like to figure out which DNA mutations are disease-causing and which are not.

The concept

Classification: the task of predicting *qualitative*/*categorical* responses

- Each response y_i is one of finitely many predetermined categories.
- *Classifying* an observation: assigning/predicting that observation to a certain category/class.
- In contrast, regression deals with "continuous" numeric response values.

As in regression, in the classification setting

- We have a set of training observations $(x_1, y_1), \ldots, (x_n, y_n)$ that we can use to build a classifier.
- We want our classifier to perform well not only on the training data, but also on test observations that were not used to train the classifier.

Why not linear regression?

No natural ordering

In example 1 above, a person arrives at the emergency room with a set of symptoms. We would like to treat the person based on three reasonable medical conditions: `Appendicitis`, `Food poisoning`, `Gastritis`.

- We could code each medical condition Y as:

$$Y = \begin{cases} 1, & \text{if } \texttt{Appendicitis}\,, \\ 2, & \text{if } \texttt{Food poisoning}, \\ 3, & \text{if } \texttt{Gastritis}. \end{cases}$$

  This coding implies an ordering on the outcomes, insisting that the difference between `Appendicitis` and `Food poisoning` is the same as the difference between `Food poisoning` and `Gastritis`.

- We could also code:

$$Y = \begin{cases} 1, & \text{if } \texttt{Gastritis}, \\ 2, & \text{if } \texttt{Appendicitis}, \\ 3, & \text{if } \texttt{Food poisoning}. \end{cases}$$

  Equally reasonable, but would lead to very different predictions on test observations.

Natural ordering

What if categories had a natural ordering, such as `mild`, `moderate`, and `severe`?

- Issue: the distance between *ordinal* categories is generally unknown.
- In general there is no natural way to convert a qualitative response variable with *more than two levels* into a quantitative response that is ready for linear regression.

Only two levels

Can we use linear regression for a *binary* (two levels) response?

- In the Default data set, the two response values can be coded as

$$Y = \begin{cases} 1, & \text{if } \texttt{Default}, \\ 0, & \text{if } \texttt{Not default}. \end{cases}$$

- We could then fit a linear regression to this binary response:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \times \texttt{Balance} + \hat{\beta}_2 \times \texttt{Income}$$

  and then predict Default if $\hat{Y} > 0.5$ and Not default otherwise.
- What if we also want to estimate e.g.,

$$P\big(\texttt{Default} \mid \texttt{Balance} = 4000, \texttt{Income} = 80000\big)$$

  i.e., the probability of defaulting given certain values of Balance and Income?
- Issue: $\hat{Y}$ can be smaller than zero or larger than one.

Only two levels

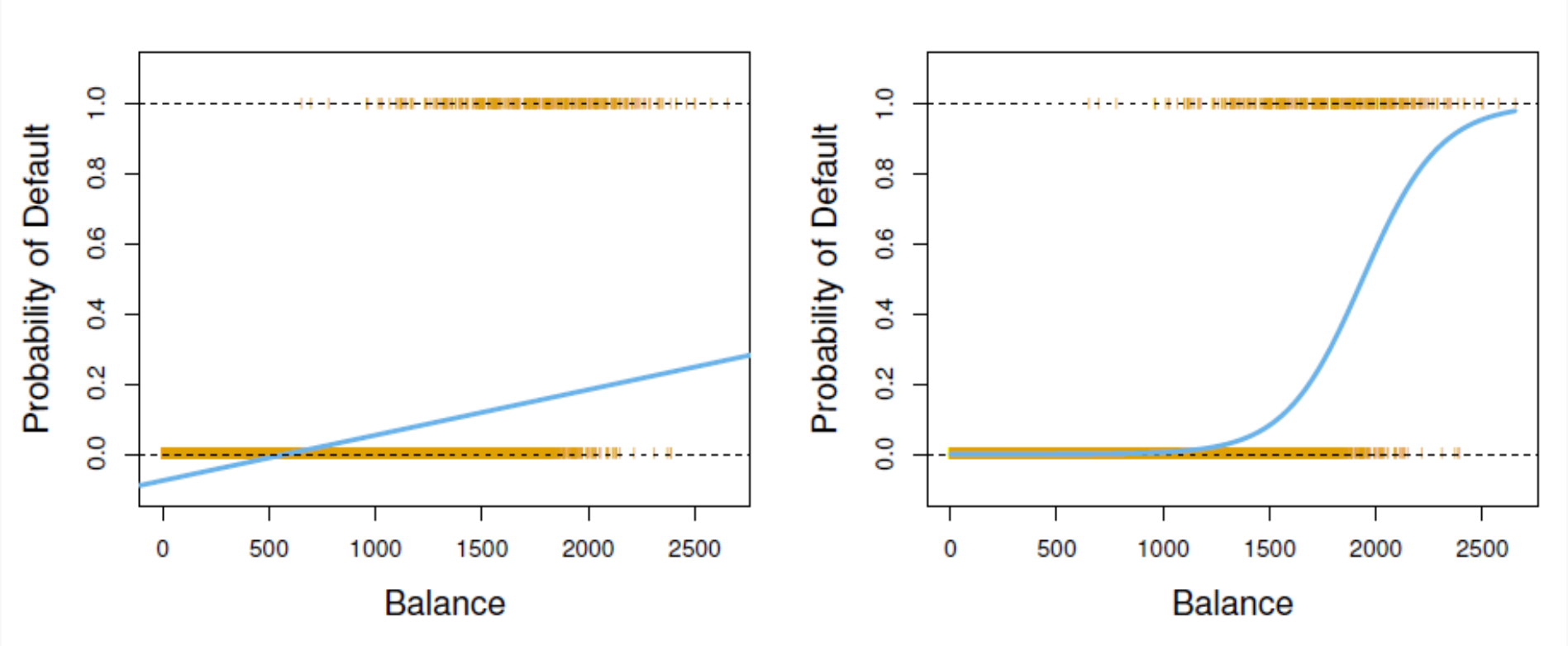

Figure 2: Image by James et al. (2021), based on the `Default` data set in R. Left: The estimated probability of default using *linear regression*, where the orange ticks indicate the values "0" for `No`, and "1" for `Yes`. Right: Predicted probabilities of default using *logistic regression*, where all probabilities lie between 0 and 1.

- Let's explore logistic regression.

Odds and log odds

Odds and log odds

Let $P(A)$ be the *probability* that event A occurs. Then $P(A) \in [0, 1]$.

- The *odds* of A occurring is defined as

$$\frac{P(A)}{1 - P(A)} \tag{1}$$

  which can be a value in $[0, \infty)$.

- The *log odds* of A occurring is defined as

$$\log\left(\frac{P(A)}{1 - P(A)}\right) \tag{2}$$

  which can be a value in $(-\infty, \infty)$.

Will be useful for interpreting a logistic regression model.