4.1/4.2 An Overview of Classification

Dr. Lauren Perry

Examples of Classification Problems

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 the three conditions does the individual have?
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, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

Classification Problems

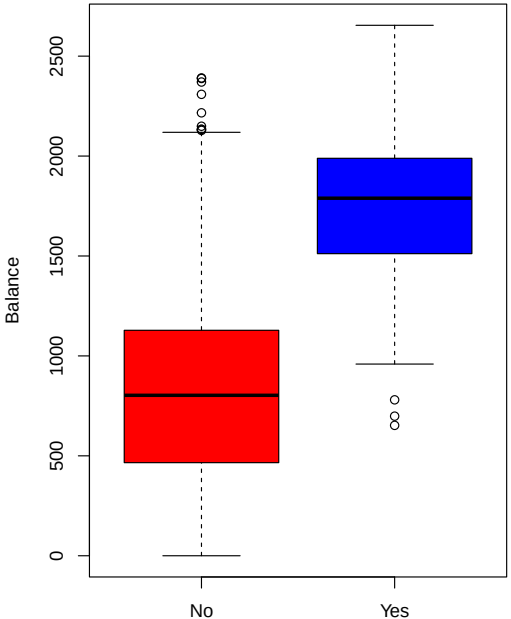
- The data is going to look very similar to the data used in regression.
- However, now Y will be *categorical*.

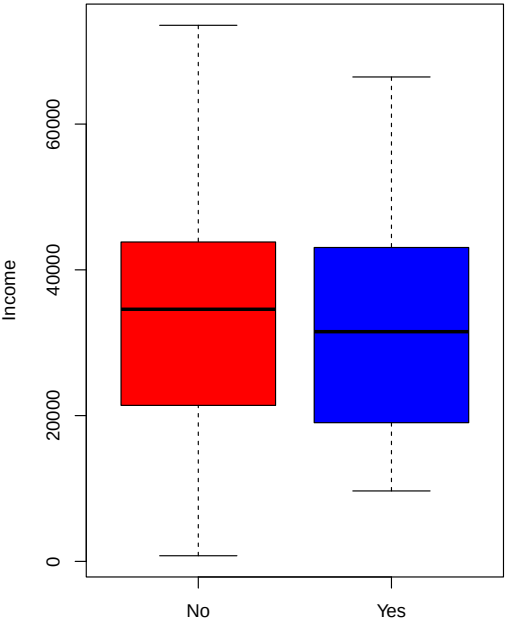
Classification: The `Default` Data

This is a simulated dataset with variables

- `default`, whether an individual defaulted
- `balance`, credit card balance
- `income`

Classification: The `Default` Data





Why not Linear Regression?

Why not just code the categorical output as numbers?

That is, for three diagnoses: `stroke`, `overdose`, and `epileptic seizure`, let

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if overdose} \\ 3 & \text{if epileptic seizure} \end{cases}$$

What problems might this cause? What are we assuming about the diagnoses that's different if we let

$$Y = \begin{cases} 1 & \text{if overdose} \\ 2 & \text{if stroke} \\ 3 & \text{if epileptic seizure} \end{cases}$$

What if Y only has two categories?

Can we use a dummy variable?

Suppose we are only interested in whether a patient has had a stroke.

$$Y = I(\text{stroke})$$

- This seems reasonable: we could just run the regression and predict
  - stroke if $\hat{Y} > 0.5$
  - no stroke otherwise
- This works out ok, but we'd like to be able interpret the output as the *probability* of a stroke.
  - However, in this scenario $\hat{Y}$ can take values outside of $[0, 1]$.