Most statistical learning problems fall into one of two categories:

1. Supervised learning

   For each observation of the predictor measurement(s) x_i, $i = 1, ..., n$ there is an associated response measurement y_i.

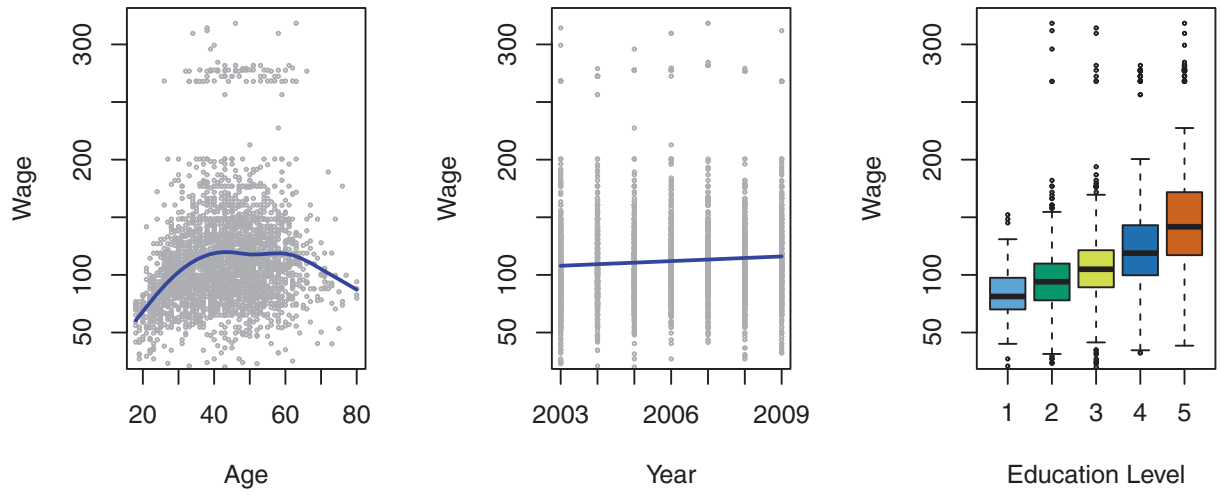
   - We wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or
   - better understanding the relationship between the response and the predictors (inference).

2. Unsupervised learning

   For every observation $i = 1, ..., n$, we observe a vector of measurements x_i but no associated response y_i.

   - We can seek to understand the relationships between the variables or between the observations. (One statistical learning tool that we may use in this setting is cluster analysis, or clustering.)

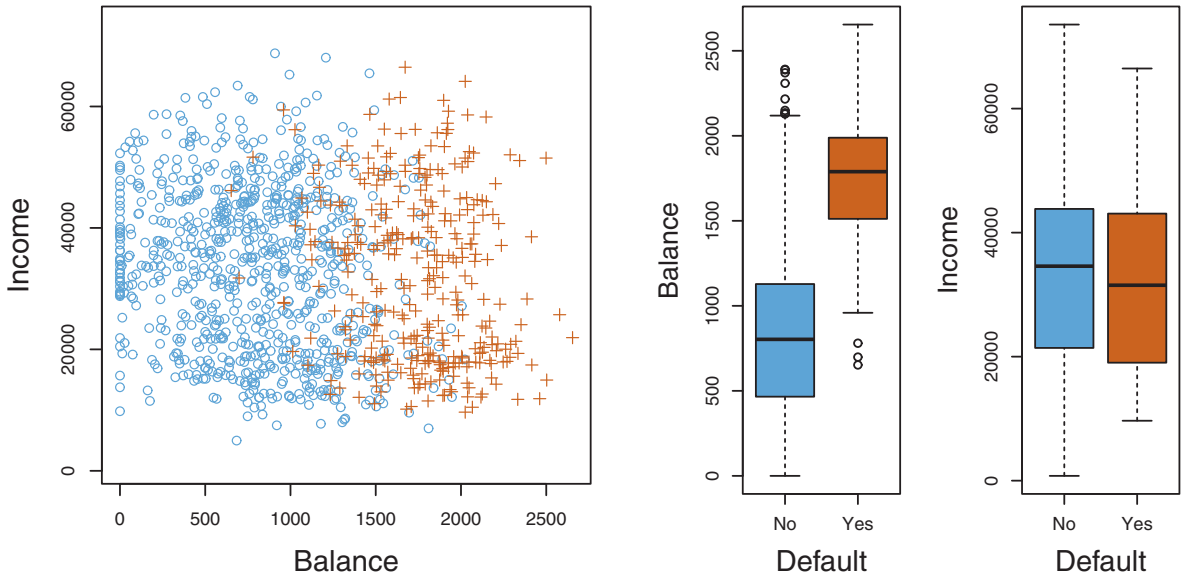
Example: We want to examine a number of factors that relate to wages for a group of males from the Atlantic region of the United States. In particular, we wish to understand the association between an employee's age and education, as well as the calendar year, on his wage.



Given an employee's age, we can use the curve to predict his wage.

This is a supervised learning problem. It involves predicting a continuous or quantitative output value. This is often referred to as a regression problem.

Example: We are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit card balance.



orange: individuals who defaulted
blue: individuals who didn't default
Given the values of balance and income, we would like to predict default.
This is a supervised learning problem. It involves predicting a non-numerical value, that is a qualitative or categorical output. This is known as a classification problem.

Example: We consider 6,830 gene expression measurements for each of 64 cancer cell lines. We are interested in determining whether there are groups, or clusters, among the cell lines based on their gene expression measurements.

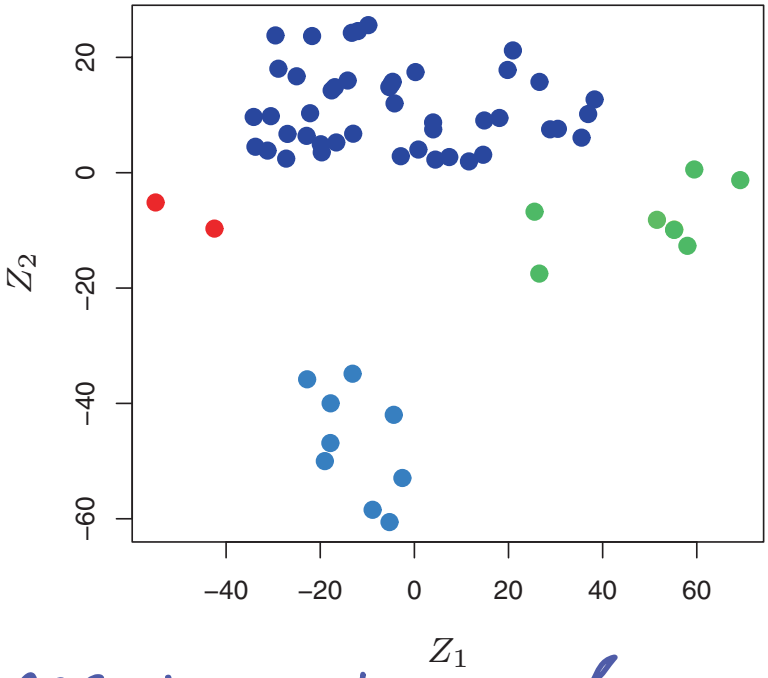


Figure suggest at least four groups of cell lines, represented using separate colors.

This is an unsupervised learning problem. we only observe input variables, with no corresponding output. It does not involve predicting a particular output variable. This is known as a clustering problem.

Measuring the Quality of Fit in the Regression Setting

In order to evaluate the performance of a statistical learning method on a given data set, we need some way to measure how well its predictions actually match the observed data.

- In the regression setting, the most commonly-used measure is the mean squared error (MSE), given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

where $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for the ith observation.

- The MSE will be small if the predicted responses are very close to the true responses, and will be large if for some of the observations, the predicted and true responses differ substantially.

Notes:

i. The MSE in the formula above is computed using the training data that was used to fit the model, and so should more accurately be referred to as the training MSE.

ii. We want to choose the method that gives the lowest test MSE, as opposed to the lowest training MSE. That is, if we had a large number of test observations, we could compute

$$Ave\left(\hat{f}(x_0) - y_0\right)^2$$

the average squared prediction error for these test observations (x_0, y_0).