- In this setting, one may be interested in answering the following questions:
   i. Which predictors are associated with the response?
   ii. What is the relationship between the response and each predictor?
   iii. Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

Example: In a real estate setting, one may seek to relate values of homes to inputs such as crime rate, zoning, distance from a river, air quality, schools, income level of community, size of houses, and so forth. In this case

- one might be interested in how the individual input variables affect the prices—that is, how much extra will a house be worth if it has a view of the river? This is an inference problem.
- Alternatively, one may simply be interested in predicting the value of a home given its characteristics: is this house under- or over-valued?
  This is a prediction problem.

How Do We Estimate f?

Our goal is to apply a statistical learning method to the training data (the observed set of data points used to train, or teach, our method how to estimate f) in order to estimate the unknown function f.

In other words, we want to find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y).

Most statistical learning methods for this task can be characterized as either *parametric* or *non-parametric*.

Parametric Methods

Parametric methods involve a two-step model-based approach.

1. First, we make an assumption about the functional form, or shape, of f.

For example, one very simple assumption is that f is linear in X:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

2. After a model has been selected, we need a procedure that uses the training data to fit or train the model.

In the case of the linear model, we need to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ such that

$$Y \approx \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

Notes:

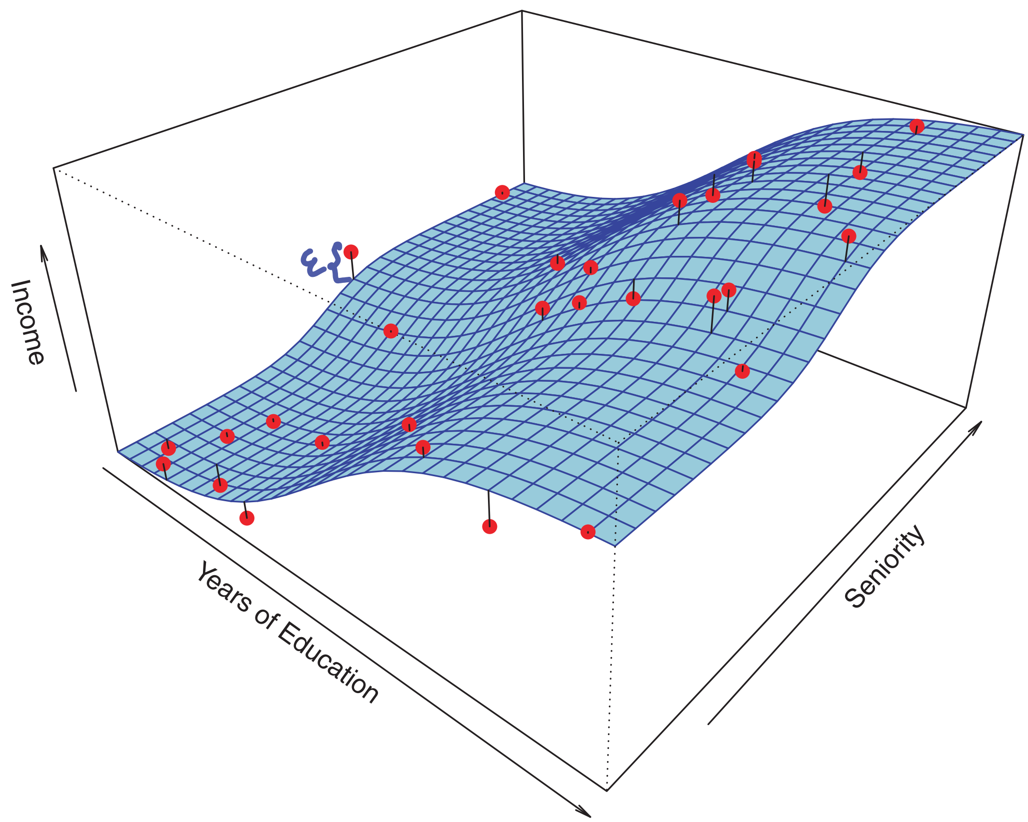
i. Assuming a parametric form for f simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters than it is to fit an entirely arbitrary function f.

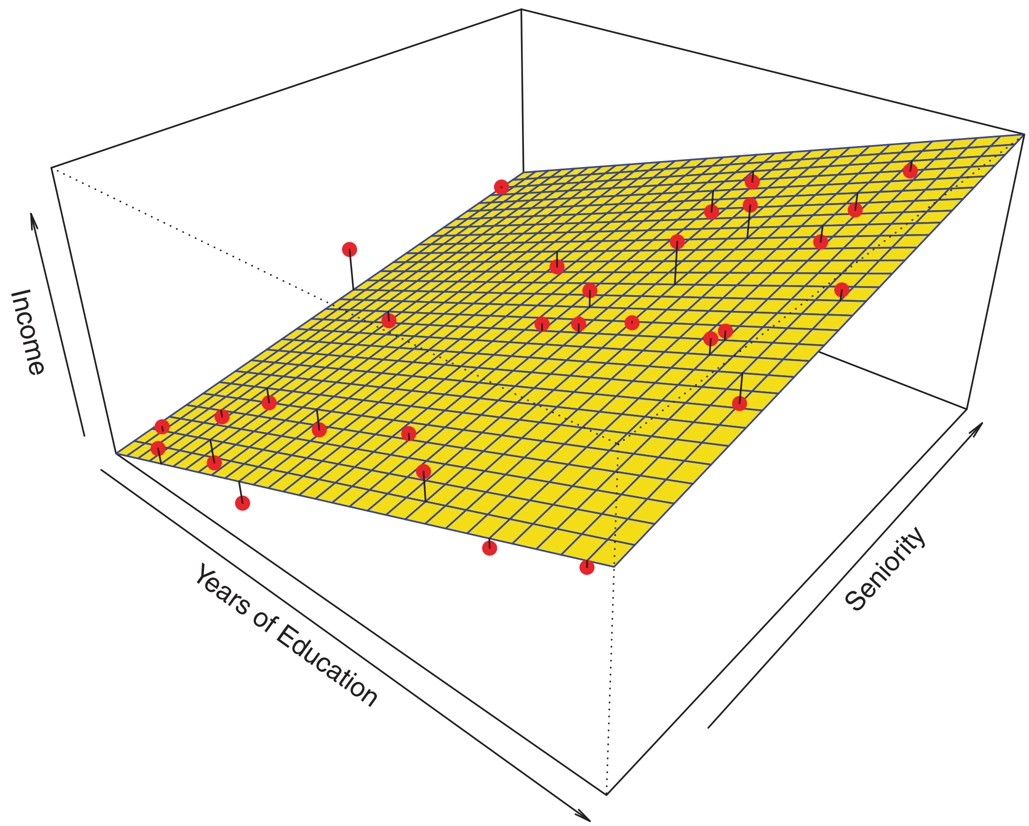
ii. The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of f leading to a poor estimate.

iii. As a solution we can choose flexible models that can fit many different possible functional forms for f. But in general, fitting a more flexible model requires estimating a greater number of parameters.

iv. These more complex models can lead to a phenomenon known as *over-fitting* the data, which essentially means they follow the errors, or noise, too closely.

Example: The plot displays income as a function of years of education and seniority for 30 individuals. The blue surface represents the true underlying relationship between income and years of education and seniority, which is known since the data are simulated. The red dots indicate the observed values of these quantities for 30 individuals.





we have fit a linear model of the form

$$\text{income} \approx \beta_0 + \beta_1 \times \text{education} + \beta_2 \times \text{seniority}$$

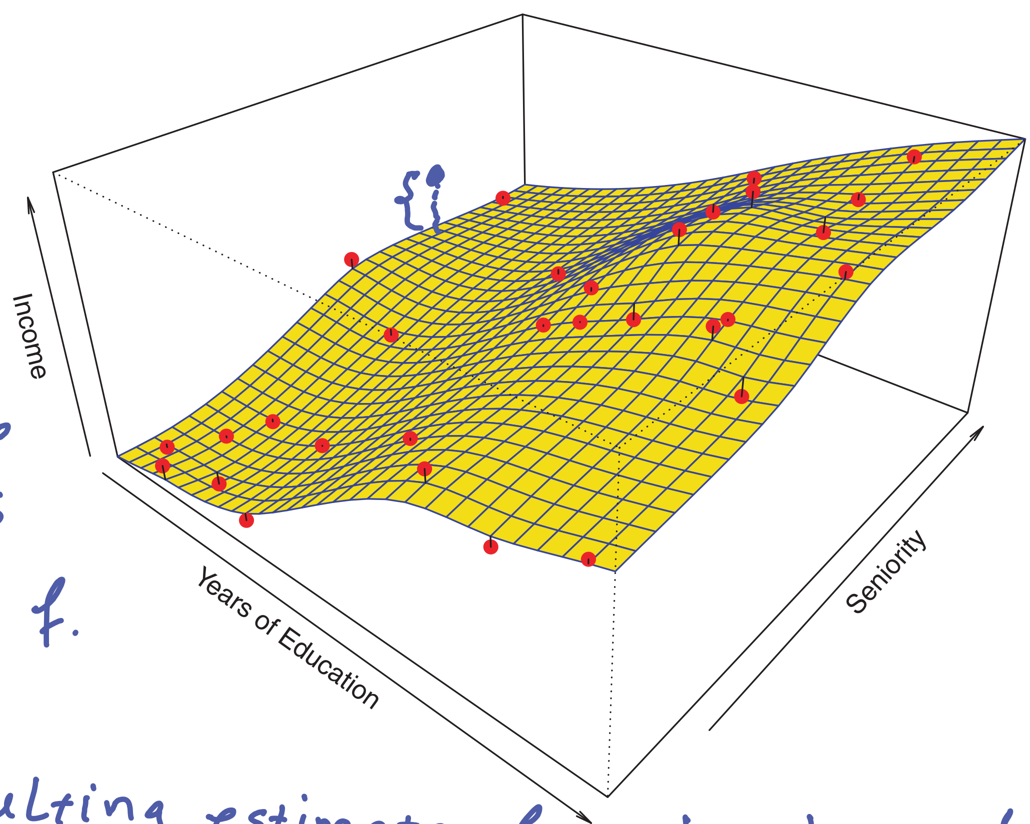
In this case the fitting problem reduces to estimating $\beta_0, \beta_1, \beta_2$.

Non-parametric Methods

Non-parametric methods do not make explicit assumptions about the functional form of f. Instead they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly.

- A major advantage over parametric approaches: by avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f.
- A major disadvantage: since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

Here a thin-plate spline is used to estimate f.



The resulting estimate fits the observed data perfectly. However, the spline fit is far more variable than the true function f.

This is an example of overfitting the data.