STATISTICAL LEARNING - CONCEPTS

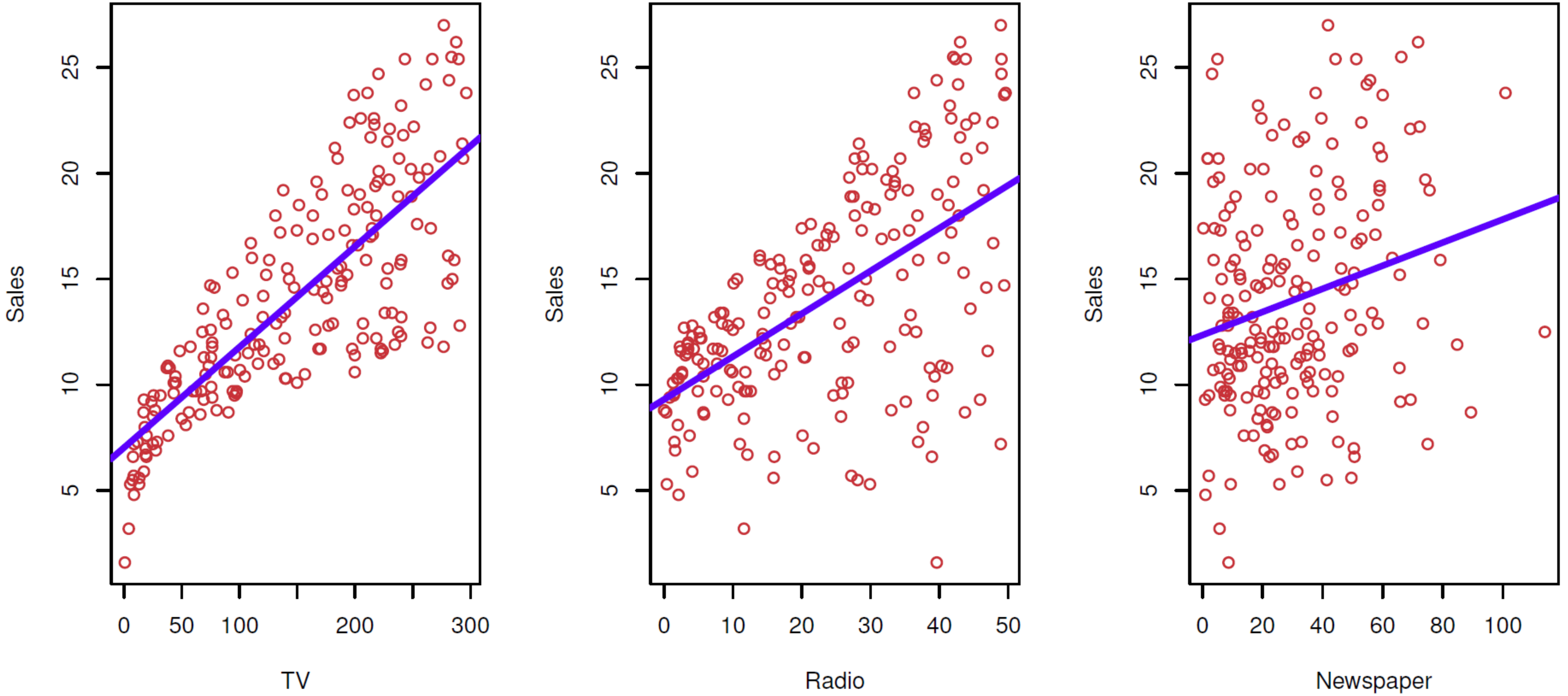
Part I – What is Statistical Learning?

Outline

- Why estimate f?
- How do we estimate f?
- The trade-off between flexibility and interpretability
- Statistical learning problems

Advertising data

- The Advertising data set consists of the sales of a product in 200 different markets, along with budgets for three different media: TV, radio, and newspaper
- Provide advice on how to improve sales of that product



Some of the figures and tables in this presentation are taken from "*An Introduction to Statistical Learning, with Applications in R*" (Springer) with permission from the authors: G. James, D. Witten, T. Hastie, and R. Tibshirani

- Input variables: TV, radio, and newspaper
  - Predictors, independent variables, covariates
  - Features, variables
- Output variable: sales
  - Response, dependent variable

The general framework

- Suppose we have a *quantitative* response Y and p predictors, $X_1, X_2, \dots, X_p$
- We assume that there is some relationship between Y and $X = (X_1, X_2, \dots, X_p)^T$

$$Y = f(X) + \epsilon$$

  - f is some unknown function of $X_1, X_2, \dots, X_p$
  - ϵ is a random *error* term with mean zero, and is independent of X

What is statistical learning?

- In essence, statistical learning refers to a set of approaches for estimating f
  - Why estimate f?
  - How do we estimate f?

Why estimate f?

- Prediction
- Inference
- A combination of the two

Prediction

- Given $X = x$, we can predict Y using

$$\hat{Y} = \hat{f}(x)$$

  - In this setting, $\hat{f}$ is often treated as a *black box*
  - Stock market data. Other examples?

Inference

- In this situation, we wish to understand how Y changes as a function of X
  - Which predictors are associated with the response?
  - What is the relationship between the response and each predictor?
  - Can the relationship be summarized using a linear equation, or is the relationship more complicated?

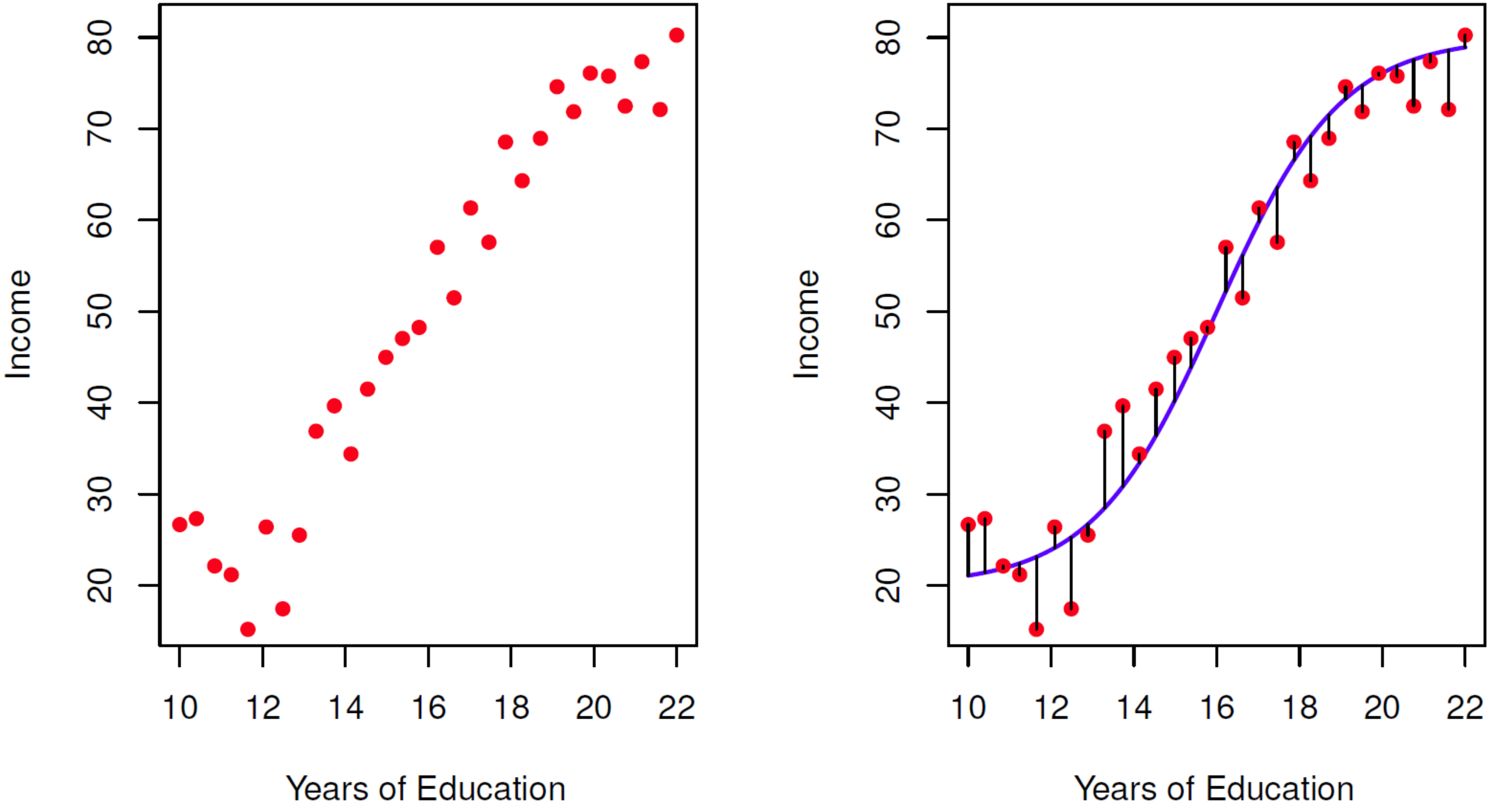
Advertising data

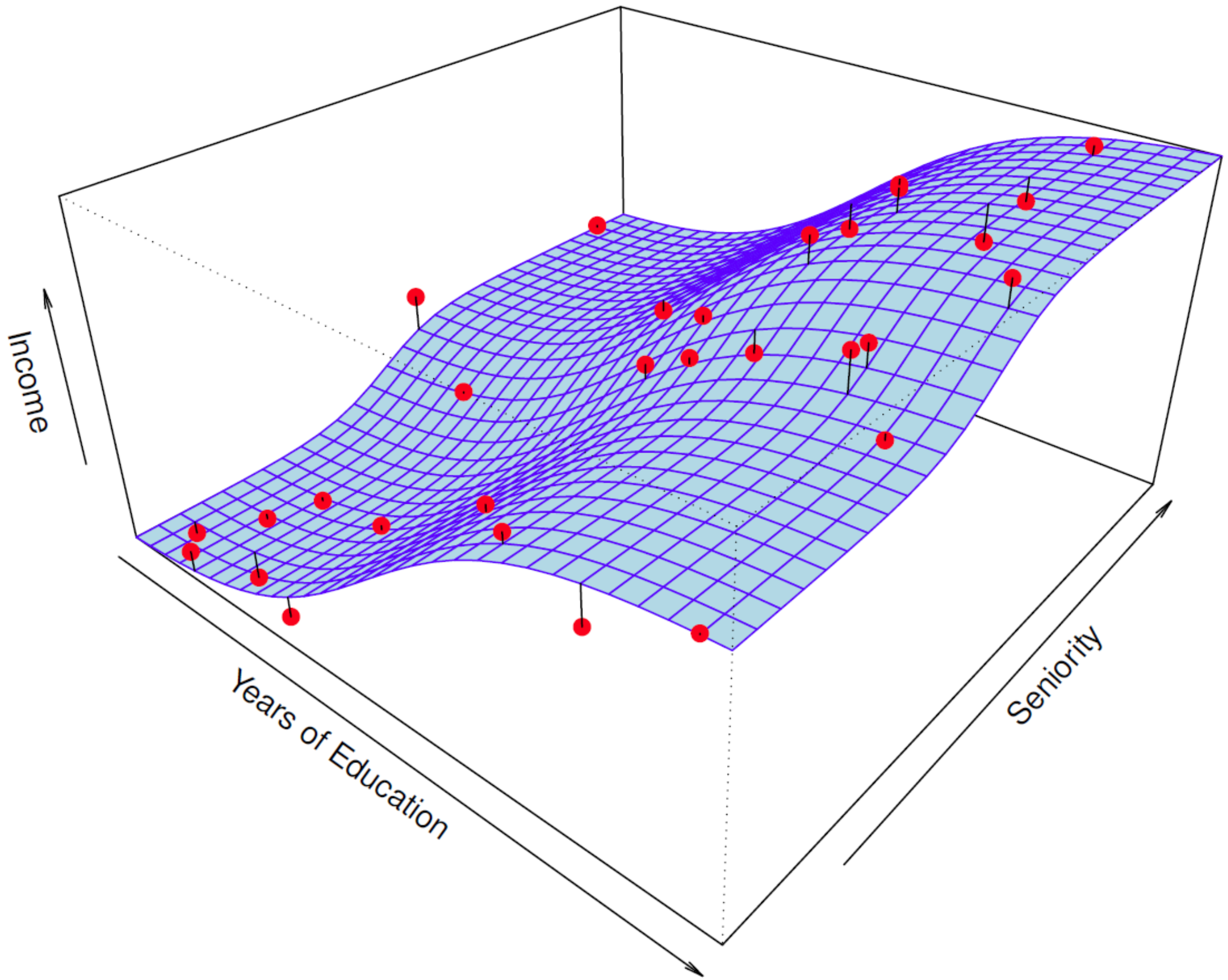
- *Which media contribute to sales?*
- *Which media generate the biggest boost in sales?*
- *How much increase in sales is associated with a given increase in TV advertising?*

How do we estimate f?

Income data

- A simulated data set consists of income (Y) and year of education (X) for 30 individuals (Income)
- The true association (f) between Y and X is known


Income
80
70
60
50
40
30
20
10
12
14
16
18
20
22
Years of Education
Income
80
70
60
50
40
30
20
10
12
14
16
18
20
22
Years of Education


Income
Years of Education
Seniority

- Training data

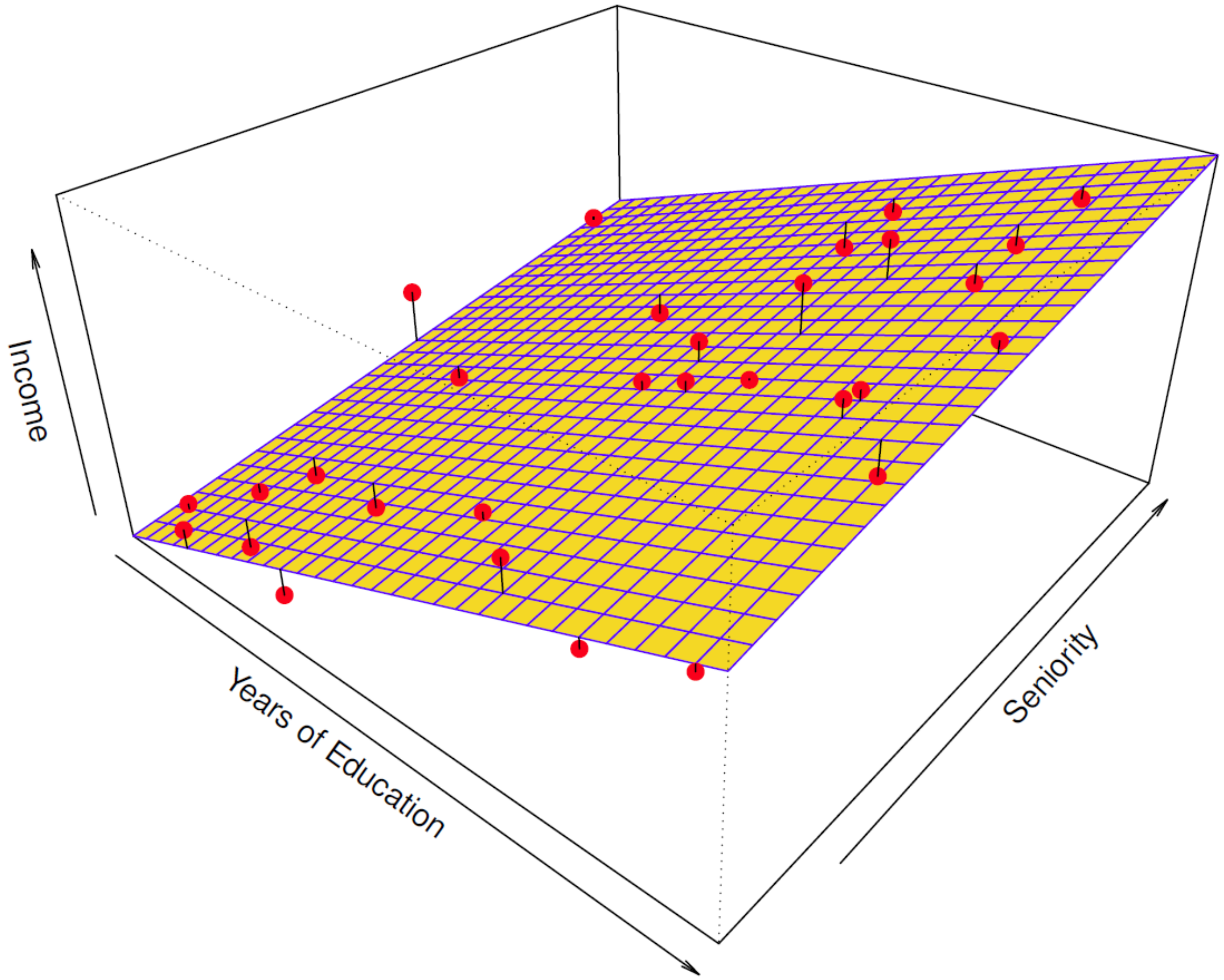
$$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

  - $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$
  - E.g., $n = 30$ and $p = 2$ in the Income data
- Apply a statistical learning method to the training data to estimate the unknown function f

- Linear or non-linear methods
- Parametric or non-parametric methods

Parametric methods

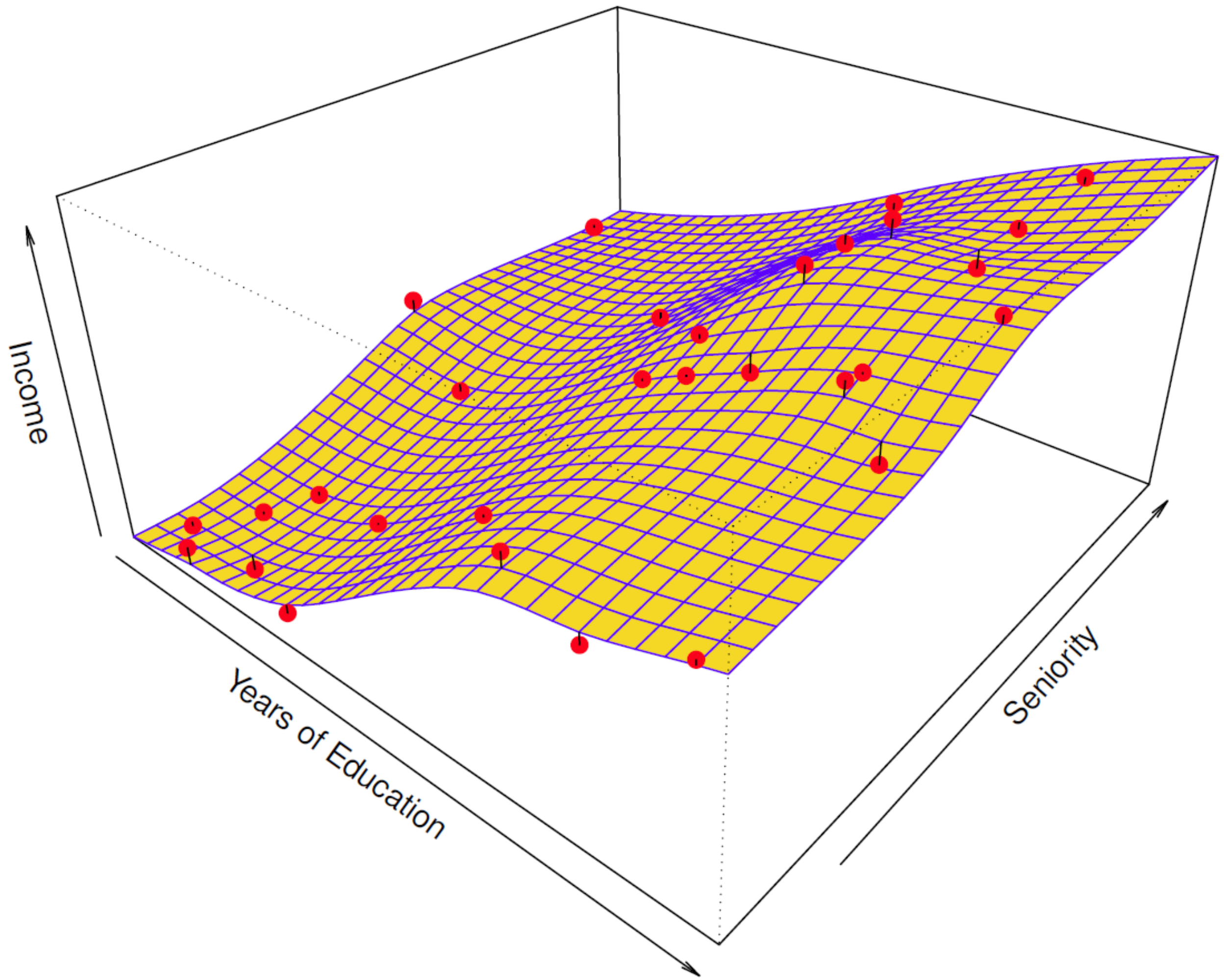
- Make an assumption about the functional form, or shape, of f
  - E.g., $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$
- Use a procedure to fit, or train, the model
  - E.g., least squares



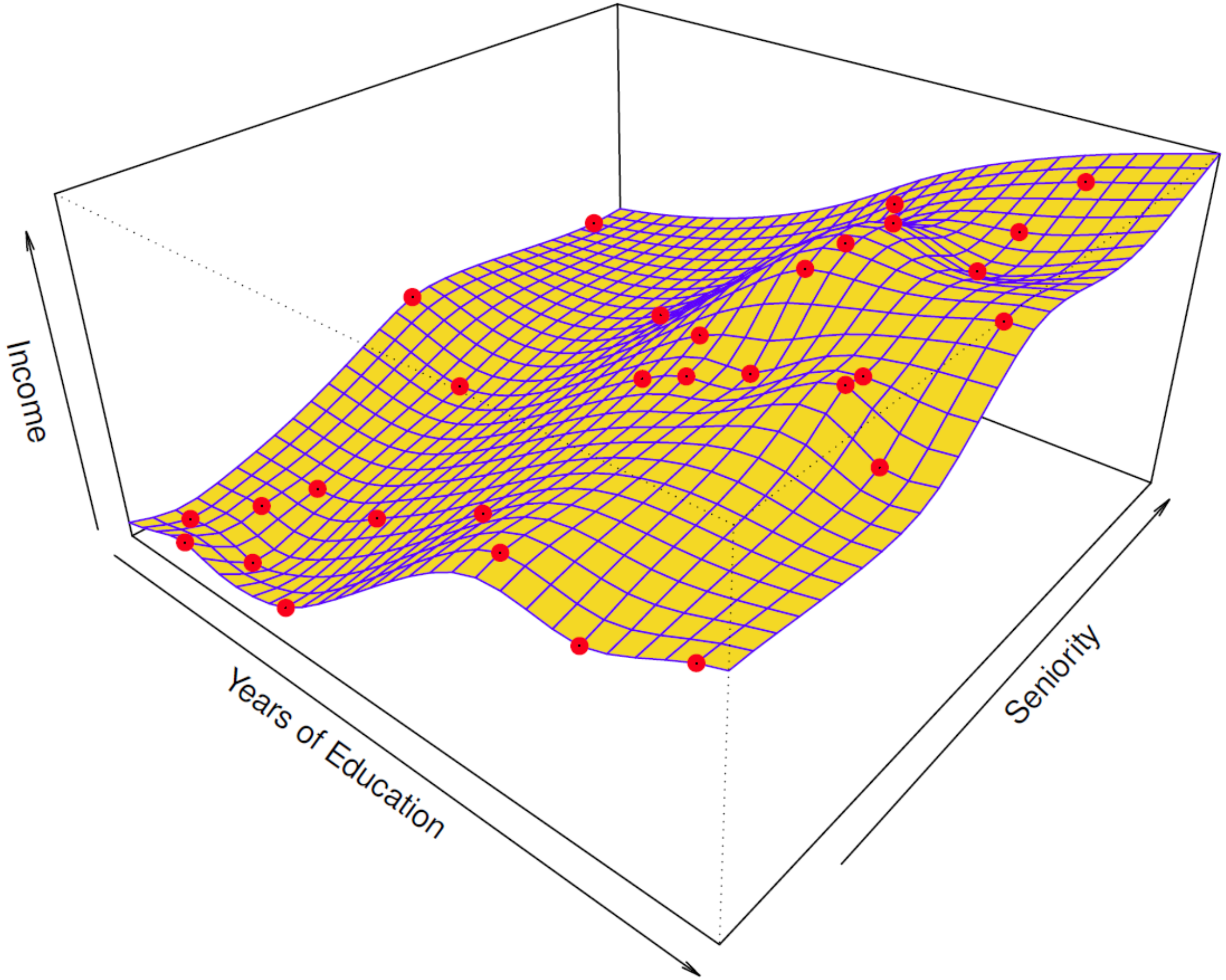
A linear model fit by least squares to the Income data

Nonparametric methods

- Do not make explicit assumptions about the functional form of f



A smooth thin-plate spline fit to the Income data



A rough thin-plate spline fit to the Income data

Advantages and disadvantages

- Parametric methods simplify the problem of estimating f to one of estimating a set of parameters, but the model will usually not match the true unknown form of f
- Nonparametric methods are much more flexible, but have the potential of overfitting the data

Trade-off between flexibility and interpretability

A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods

- Less flexible, or more restrictive, models are much more interpretable
- Different methods for estimating f may be appropriate, depending on whether our ultimate goal is prediction, inference, or a combination of the two

Statistical learning problems

Supervised and unsupervised learning

- Supervised statistical learning
  - We fit a model that relates y_i to x_i, with the aim of either prediction or inference
- Unsupervised statistical learning
  - For every observation $i = 1,2, \dots, n$, we observe a vector of measurements x_i, but no associated response y_i
  - We seek to understand the relationships between the variables or between the observations

Regression and classification

- Regression problems
  - Quantitative or numerical responses
  - Wage data, Advertising data, Income data
- Classifications problems
  - Qualitative or categorical responses
  - Smarket data