Statistical Learning for Data Science

Lecture 04

唐晓颖

电子与电气工程系

南方科技大学

February 27, 2023

Statistical Learning VS. Machine Learning

- Machine learning arose as a subfield of *Artificial Intelligence*
- Statistical learning arose as a subfield of *Statistics*
- There is much overlap – both focus on supervised and unsupervised problems:
  - ❑ Machine learning has a greater emphasis on *large scale* applications and *prediction accuracy*.
  - ❑ Statistical learning emphasizes *models* and their interpretability, *precision* and *uncertainty*.
- The distinction has become more and more blurred

Statistical Learning VS. Machine Learning

Statistical Learning:

- Use statistical methods to analyze data and make predictions or decisions
- Focus on building statistical models to describe the relationships between variables in a dataset
- Used in fields such as economics, psychology, and social sciences, where understanding the underlying statistical relationships is important

Machine Learning:

- A more modern approach that uses algorithms to automatically learn patterns and relationships in data
- Designed to automatically improve the algorithm performance
- Used to make predictions or decisions in a wide range of applications, such as image recognition, natural language processing, and self-driving cars

Statistical Learning VS. Machine Learning

https://blogs.perficient.com/2018/01/29/machine-learning-vs-statistical-learning/

https://www.jiqizhixin.com/articles/2019-05-06-13

哪种方法更好?

其实这是个很蠢的问题。从关系角度看，没有统计学，机器学习是不存在的。然而，在当前人类所经历的这个信息爆炸的时代中，面对海量数据的涌入，机器学习倒是颇为有用。

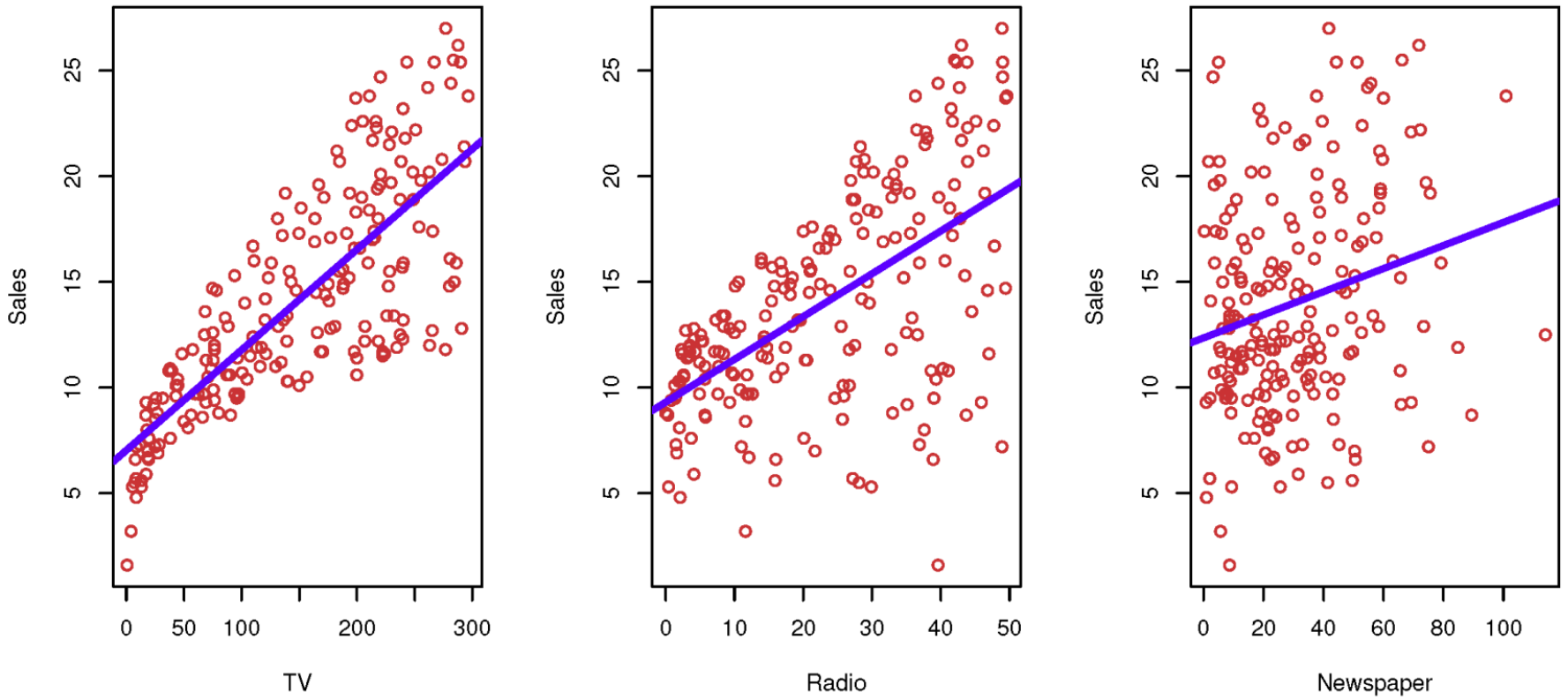
比较机器学习和统计模型确实有些困难。这主要取决于你的目的是什么。如果你想构建一种可以精确预测房价的算法，或是使用数据确定某人是否可能感染某种疾病的话，机器学习可能是更好的选择。如果你想证明变量间的关系或用数据进行推断，那么统计模型则会成为更好的选择。

Here are some of the differences:

1. Both methods are data dependent. However, Statistical Learning relies on rule-based programming; it is formalized in the form of relationship between variables, where Machine Learning learns from data without explicitly programmed instructions.
2. Statistical Learning is based on a smaller dataset with a few attributes, compared to Machine Learning where it can learn from billions of observations and attributes.
3. Statistical Learning operates on assumptions, such as normality, no multicollinearity, homoscedasticity, etc. when Machine Learning is not as assumptions dependent and in most of the cases ignores them.
4. Statistical Learning is mostly about inferences, most of the idea is generated from the sample, population, and hypothesis, in comparison to Machine Learning which emphasizes predictions, supervised learning, unsupervised learning, and semi-supervised learning.
5. Statistical Learning is math intensive which is based on the coefficient estimator and requires a good understanding of your data. On the other hand, Machine Learning identifies patterns from your dataset through the iterations which require a way less of human effort.

Statistical Learning Example

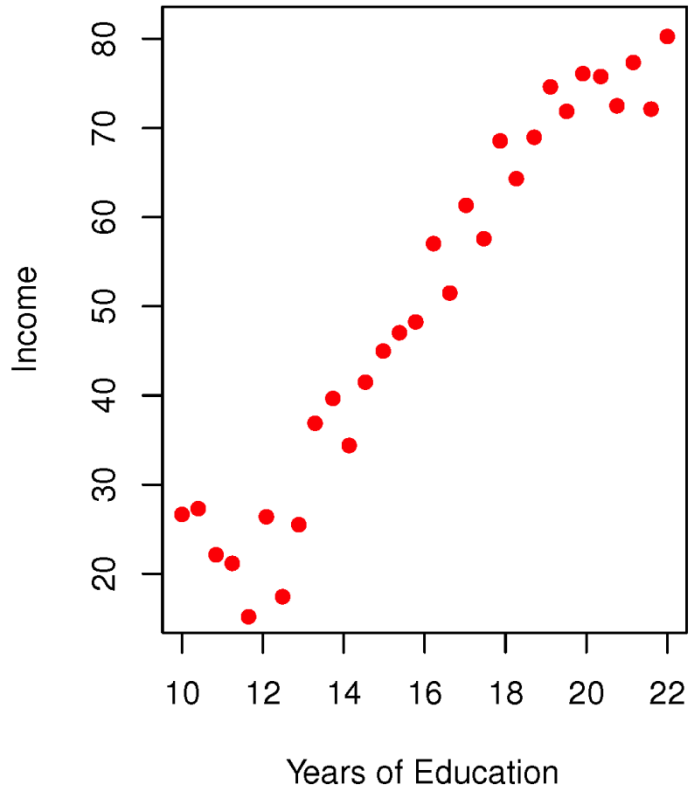
- Provide advice on how to improve sales of a particular product



Can we predict *Sales* using *TV*, *Radio*, and *Newspaper*?

Statistical Learning Example

- Understand the relationship between income and education



Can we predict *Income* using *Years of Education*?

Notation

Y : *Sales or Income* (output/response)

X_1: *TV or Years of Education* (input/predictor)

X_2: *Radio or …* (input/predictor)

X_3: *Newspaper or …* (input/predictor)

.

.

.

We can refer to the *input vector* collectively as:

$$X = \left(X_1, X_2, ..., X_p \right)$$

p : the total number of inputs or predictors

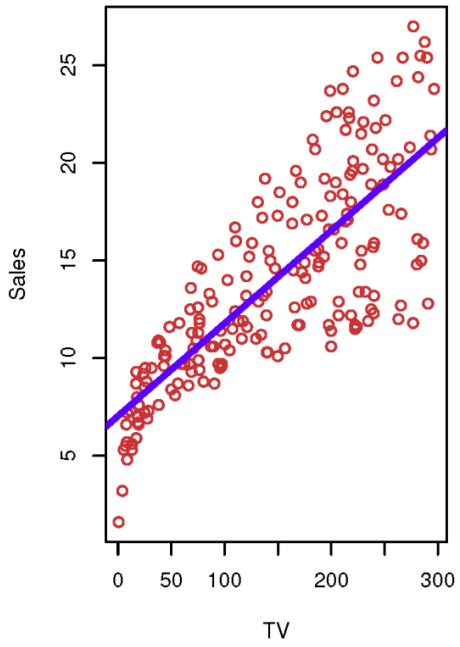
Model

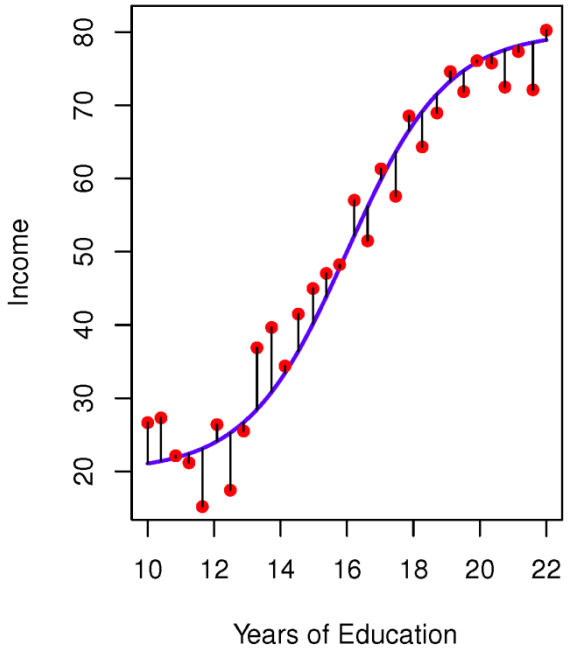
$$Y = f(X) + \varepsilon$$

f : some fixed but unknown function

ε : a random *error term* (independent of the input, has mean zero)

f represents the systematic information that X provides about Y





In essence, statistical learning refers to a set of approaches for estimating f

Why Estimate f

Two important goals in statistical learning: prediction and inference

Prediction:

- Task of making predictions about future observations based on a set of input-output pairs
- Goal: to create a model that accurately predicts the output for new input values that were not present in the training data
- Models are typically evaluated based on their ability to accurately predict the output for new data

Why Estimate f

Two important goals in statistical learning: prediction and inference

Inference:

- Task of understanding the relationship between the input and output variables in a statistical model
- Goal: to identify which variables are important in predicting the output, and to understand the strength and direction of the relationships between variables
- Models are typically evaluated based on their ability to provide insights into the underlying mechanisms that generate the data

Why Estimate f

Two important goals in statistical learning: prediction and inference

Main difference:

- Prediction is focused on accurately predicting the output for new data
- Inference is focused on understanding the relationships between the variables in the model
- Prediction models prioritize accuracy over interpretability, while inference models prioritize interpretability over accuracy

Why Estimate f

1. Prediction

   - With a good one we can make predictions of Y at new points $X = x$

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

     $\hat{f}$: our estimation of f

     $\hat{Y}$: the resulting prediction for Y

     Example: predict a patient's risk for a severe adverse reaction to a particular drug based on his/her blood sample characteristics.

Why Estimate f

1. Prediction
   - The accuracy of $\hat{Y}$ is very important
     - *Reducible error*

       Error induced by the estimation of f

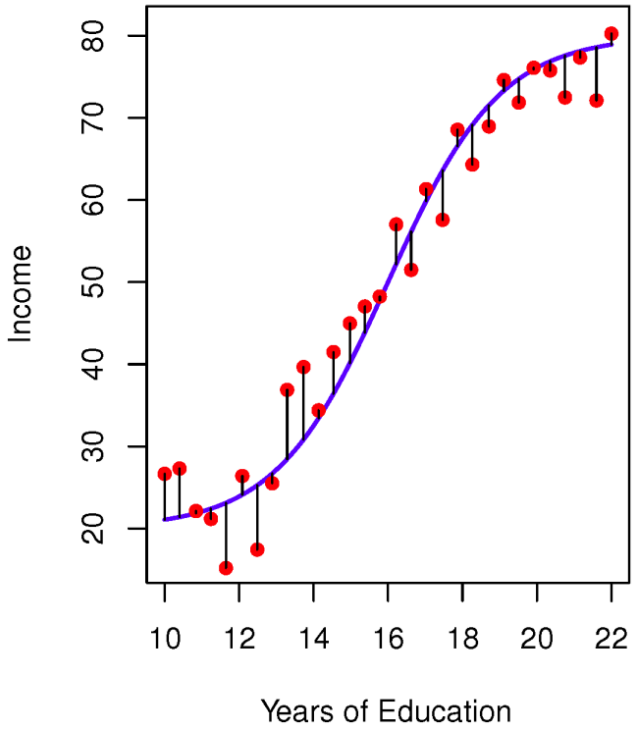
     - *Irreducible error*

       Error induced by ε

Model

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

ε : a random *error term* (independent of the input, has mean zero)



The vertical lines represent the error terms ε

The random error is caused by some uncontrollable reasons, such as the environment, the temperature, when real data is generated. It can be positive and negative. It is not measurable.

Why Estimate f

1. Prediction

   The mean squared error (MSE)

$$E(Y-\hat{Y})^2 = E\left[f(X)+\varepsilon-\hat{f}(X)\right]^2$$

$$=\underbrace{\left[f(X)-\hat{f}(X)\right]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\varepsilon)}_{\text{Irreducible}}$$

$$\begin{aligned}
\text{Var}(X) &= \text{E}\left[(X-\text{E}[X])^2\right] \\
&= \text{E}\left[X^2 - 2X\,\text{E}[X] + \text{E}[X]^2\right] \\
&= \text{E}\left[X^2\right] - 2\,\text{E}[X]\,\text{E}[X] + \text{E}[X]^2 \\
&= \text{E}\left[X^2\right] - \text{E}[X]^2
\end{aligned}$$

Why Estimate f

1. Prediction

   - Our goal is to provide techniques for estimating f with the aim of minimizing the reducible error.

   - The irreducible error will always provide an upper bound on the accuracy of our prediction. This bound is almost always unknown in practice.

   - The prediction accuracy is the most important. We may not care the exact format/structure of $\hat{f}$.

Why Estimate f

2. Inference

Understand how Y changes as a function of $X_1, X_2, ..., X_p$

Now $\hat{f}$ cannot be treated as a black box, we need to know its exact form

- *Which predictors are associated with the response?*

  Identifying the few important predictors among a large set of possible variables can be extremely useful

- *The relationship between the response and each predictor*

  Some predictors may have a positive relationship with the output while some other may have a negative relationship. The relationship between the response and a given predictor may also depend on the values of the other predictors.

Why Estimate f

Prediction example

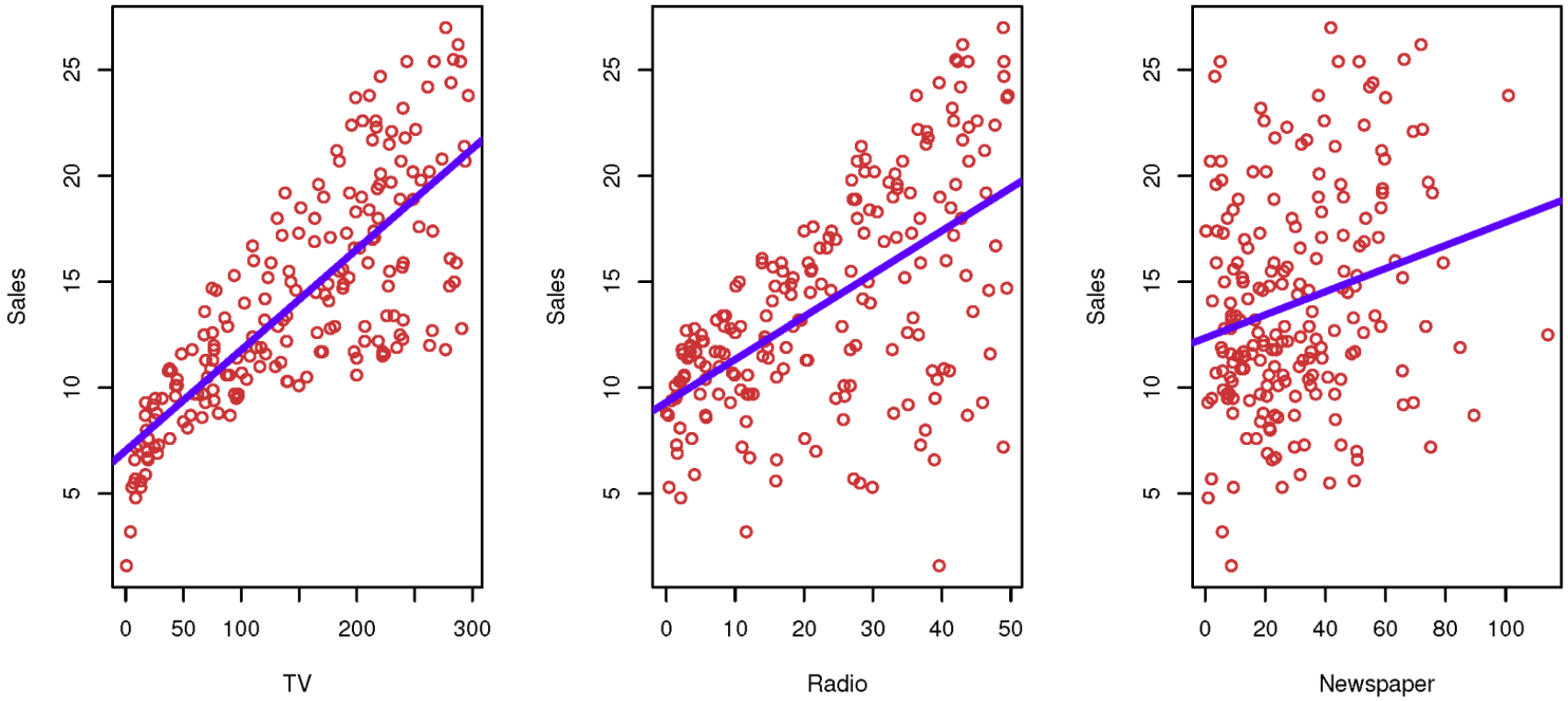
Identify individuals who will respond positively to a mailing, based on observations of demographic variables measured on each individual.

-- *Not interested in obtaining a deep understanding of the relationships between each individual predictor and the response.*

-- *Simply wants an accurate model to predict the response using the predictors.*

Why Estimate f

Inference example



-- Which media contribute to sales?

-- Which media generate the biggest boost in sales?

Why Estimate f

Prediction + Inference example

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.

-- *Interested in how the individual input variables affect the prices ("how much extra will a house be worth if it has a view of the river?")* Inference

-- *Interested in predicting the value of a home given its characteristics ("is this house under- or over- valued?")* Prediction

Why Estimate f

Depending on our ultimate goal (prediction, inference, or a combination of the two), different methods of estimating f may be appropriate.

Example: linear models allow for relatively simple and interpretable inferences but may not yield very accurate predictions. In contrast, some of the highly non-linear approaches can potentially provide quite accurate predictions, but this comes at the expense of a less interpretable model for which inference is more challenging.

How Do We Estimate f

Three key elements for statistical learning: Method = model + optimization + algorithm

Model

- A mapping from input to output
- Two forms of models:
  - Probability model (conditional probability distribution $P(Y|X)$)
  - Non-probability model (decision function $Y = f(X)$)
  - Set of conditional probabilities: $\mathcal{F} = \{P \mid P(Y \mid X)\}$
  - Set of decision functions: $\mathcal{F} = \{f \mid Y = f(X)\}$

How Do We Estimate f

Three key elements for statistical learning: Method = model + optimization + algorithm

Optimization

- Consider what criteria to learn or choose the best model.
- Loss function: evaluation of a prediction

$$L(Y, f(X)) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases}$$

0-1 loss function

$$L(Y, f(X)) = (Y - f(X))^2$$

Quadratic loss function

$$L(Y, f(X)) = |Y - f(X)|$$

Absolute loss function

How Do We Estimate f

Three key elements for statistical learning: Method = model + optimization + algorithm

Algorithm

- Algorithms for solving optimization problems
- If the optimization problem has an explicit analytical formula, the algorithm is relatively simple
- However, the analytical formula usually does not exist, so the numerical calculation method is needed

How Do We Estimate f

Two broad categories of statistical learning methods: parametric and non-parametric methods

Parametric methods:

- Make assumptions about the underlying distribution of the data
- Use a fixed set of parameters to describe the distribution
- The parametric model has a fixed form and the parameters are learned from the data
- Examples: linear regression, logistic regression, and the Naive Bayes classifier

Non-parametric methods:

- Use flexible models that can adapt to the data
- Do not make assumptions about the underlying distribution
- Examples: decision trees, random forests, and support vector machines

How Do We Estimate f

- Parametric Methods (model-based approach)

  1. Make an assumption about the functional form, or shape, of f

     Example: linear model

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

     Instead of having to estimate an entirely arbitrary p-dimensional function $f(X)$, one only needs to estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_p$

How Do We Estimate f

- Parametric Methods (model-based approach)

  2. Use the training data $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ to fit or train the model

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_p x_{1p} \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_p x_{2p} \\ \quad . \\ \quad . \\ \quad . \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_p x_{np} \end{cases}$$

The most common approach is least squares

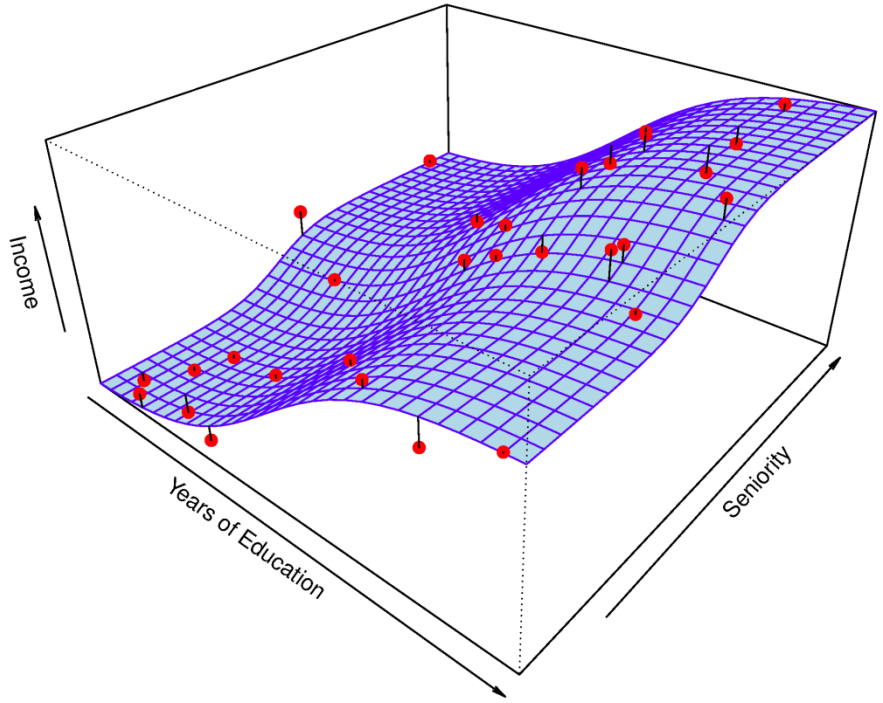
How Do We Estimate f

- Parametric Methods (model-based approach)

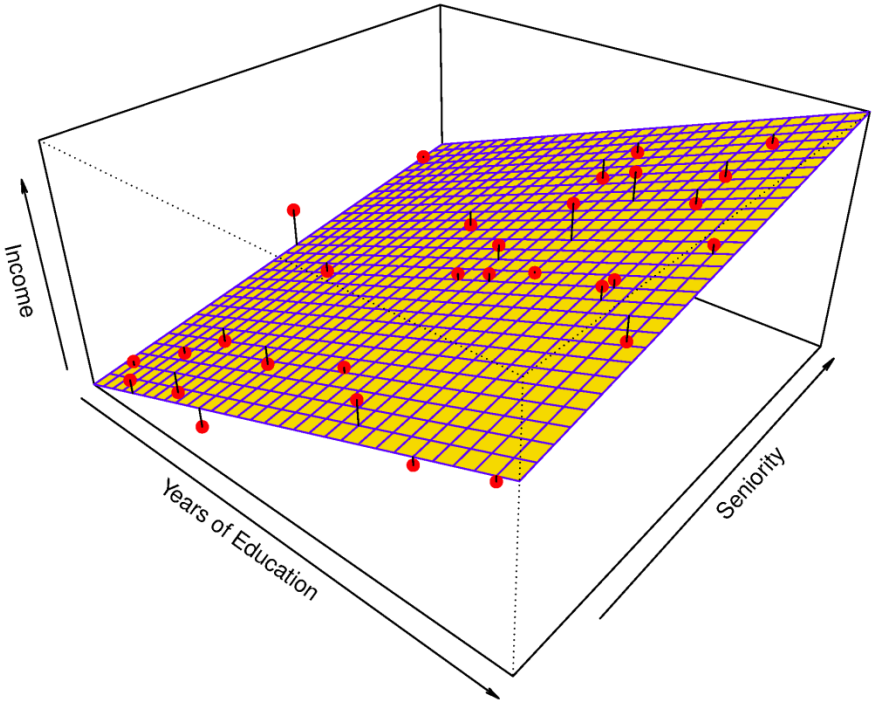
  Pro: Simplifies the estimation problem

  Con: Model choosing is very important

True function



Estimated function using linear model



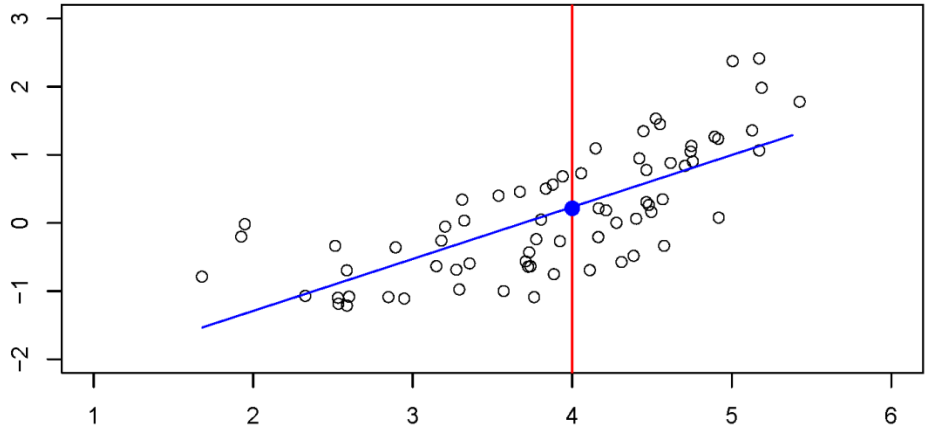
How Do We Estimate f

- Parametric Methods (model-based approach)

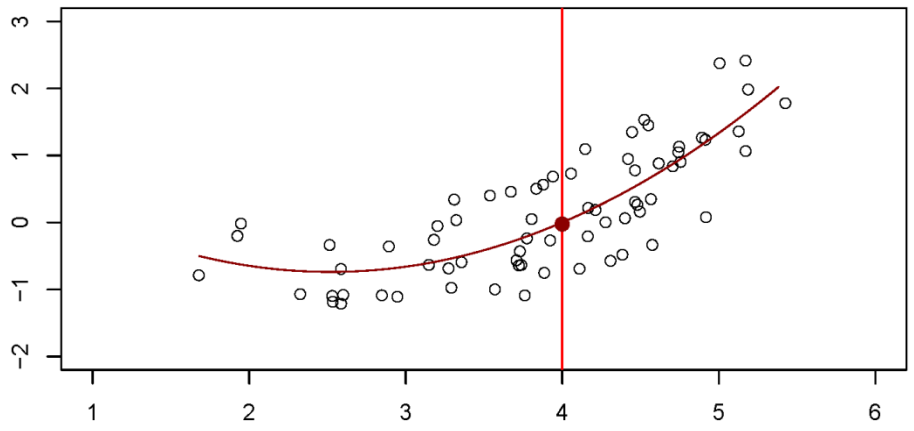
  Pro: Simplifies the estimation problem

  Con: Model choosing is very important

Linear model fitting



Quadratic model fitting



Although it is almost never correct, a linear model often serves as a good and interpretable approximation to the unknown true function