Prediction vs Inference.

W. Wang[1]

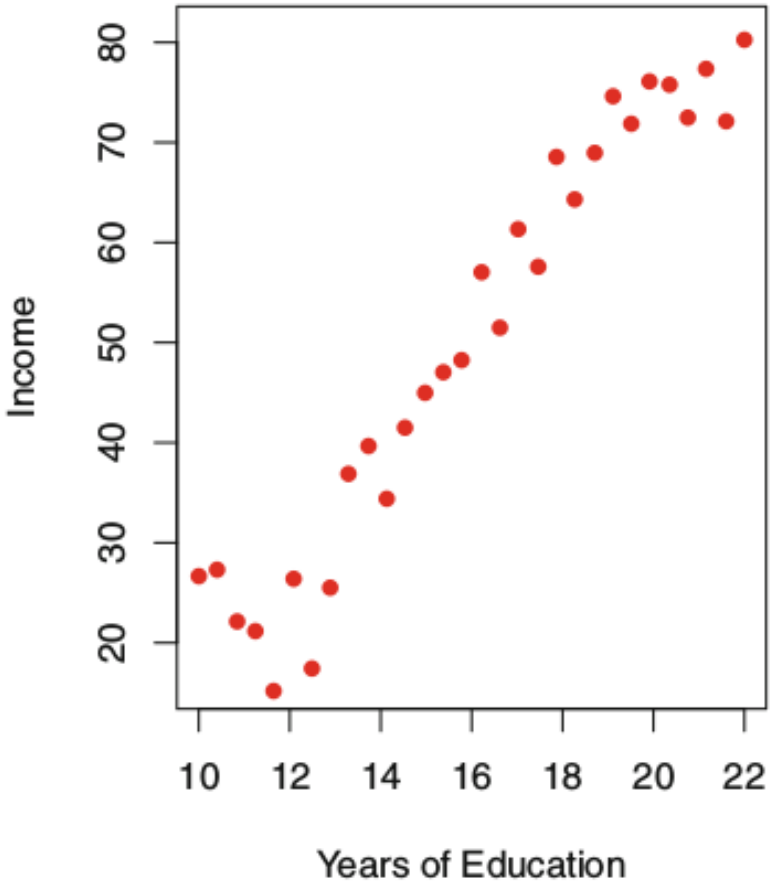
[1]Department of Mathematics
University of Houston

MATH 4323

Supervised Learning: *Income* data.

Example. Consider *Income* data set (*Income*1.*csv*) on a

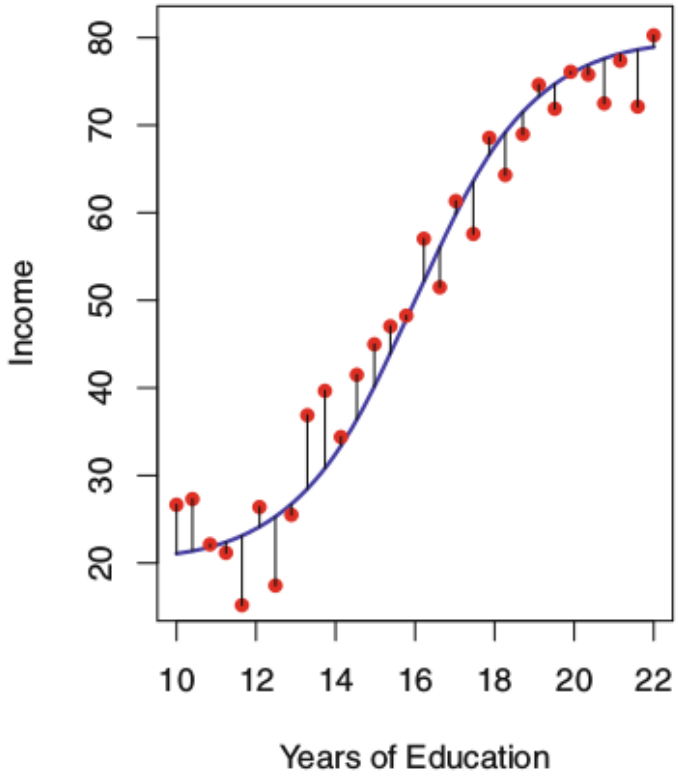
- person's *income* ($Y = income$, response variable),
- as a function of education years ($X = \{\text{education years}\}$, predictor variable).



The goal of supervised statistical learning is to identify that function $f()$ such that $Y \approx f(X)$.

Supervised Learning: *Income* data.

Example (cont'd). Disclaimer: *Income* is actually a *simulated* data set (didn't come from real observed data).



It means that the true form of the underlying function $f()$, such that $Y \approx f(X)$, is available to us. It is shown via blue curve.

As you may see, the data does not lie exactly on the blue curve.

Why? What are those vertical lines between actual data (red points) and the true curve (blue line)?

Supervised Learning: Formula.

In supervised learning we assume the following general formula:

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

where

- $f(X)$ is information about Y provided by X,
- while ϵ describes information about Y not captured by X. It is
  - a random quantity, centered around 0 ($E[\epsilon] \equiv 0$),
  - oftentimes referred to as model error.

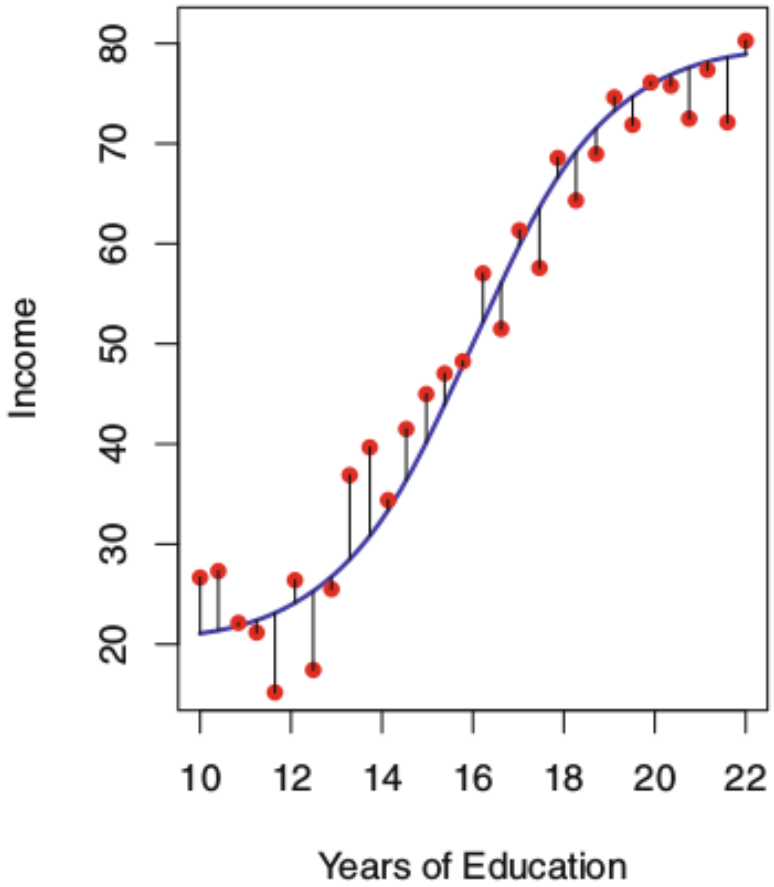
In the model formula above, the observation index $i, i = 1, \ldots, n$ is implicit:

$$Y^{(i)} = f(X^{(i)}) + \epsilon^{(i)}, \ i = 1, \ldots, n$$

Supervised Learning: *Income* data.

Example (cont'd). For our *Income* data set,

- Y - income,
- $X \equiv X_1 =$ years of education.



The vertical lines represent the error terms $\epsilon^{(i)}, i = 1, \ldots, n$.

Some of the 30 observations lie above the blue curve, some - below; on average, errors have approximately mean zero ($E[\epsilon] \approx 0$).

Income example: Two predictors.

In general, the function f may involve more than one input variable.

Example (cont'd). In addition to years of education, let's include *seniority* as a predictor for person's income:

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

where we now have:

- Y - income (response, as before),
- $X = (X_1, X_2)$, X_1 - years of education, X_2 - seniority (two predictors)

Here f is a two-dimensional surface that must be estimated based on the observed data.

Income example: Two predictors.

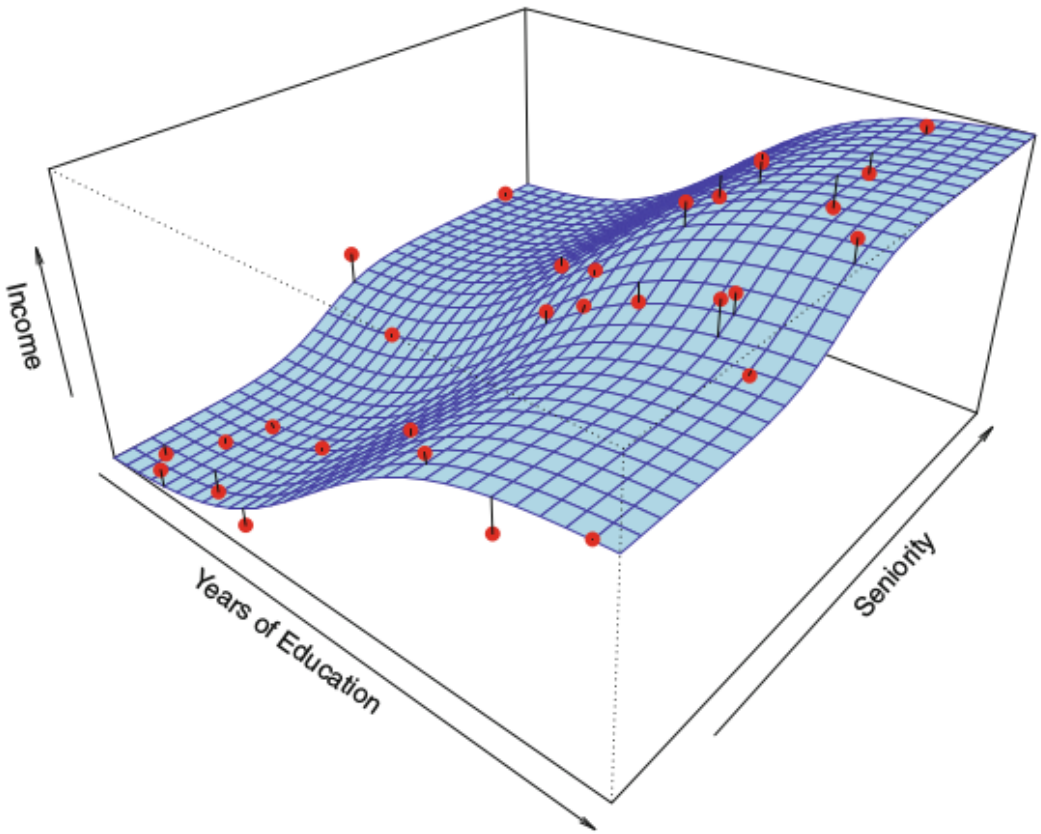


FIGURE 2.3. *The plot displays* `income` *as a function of* `years of education` *and* `seniority` *in the* `Income` *data set. The blue surface represents the true underlying relationship between* `income` *and* `years of education` *and* `seniority`*, which is known since the data are simulated. The red dots indicate the observed values of these quantities for* 30 *individuals.*

Supervised Learning: General Formula.

More generally, suppose that we observe a

- quantitative response Y, and
- p different predictors, $X_1, X_2, \ldots, X_p$.

Then the supervised learning formula becomes:

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

where

- f is some fixed, but unknown, function of $X = (X_1, \ldots, X_p)$,
- ϵ is a random error term, which is
  - independent of X,
  - has mean zero.

In this formulation,

- f represents systematic information that X provides about Y, while
- ϵ captures random aspect that wasn't explained by X with respect to Y (more on it later in this lecture).

Estimating f: Two Main Reasons.

Unlike the simulated example on income data we showed, the true form of function f s.t. $Y \approx f(X)$ is generally unknown.

The main focus of supervised learning is using the observed data in order to come up with an estimate $\hat{f}$ for that function.

There are two main reasons that we may wish to estimate f:

- **prediction**,
- **inference**.

Estimating f: Two Main Reasons.

Both are fully dictated by the eventual goal of your statistical learning:

- **prediction** - (In situations where the inputs X are readily available, but the output Y cannot be easily obtained.) Come up with estimate $\hat{f}$ that produces the best possible **predictions** $\hat{Y} = \hat{f}(X)$ of unobserved cases.
- **inference** - (In situations where our goal is not necessarily to make predictions for Y; but how Y changes as a function of $X_1,\ldots,X_p$) attempt to **infer** the relationship between each predictor X_i and the response Y.

Prediction.

Below are some examples of data problem formulations where prediction is prioritized.

Examples.

- We give a particular drug to patients. Suppose that $X_1, \ldots, X_p$ are characteristics of a patient's blood sample, while Y is whether patient experiences adverse reaction to that drug. We wish to **predict** whether patient with particular characteristics will have an adverse reaction to the drug.
- Suppose $X_1, \ldots, X_p$ are stock price dynamics in the previous p days, while Y - stock's direction (*Up*/*Down*) today. We wish to **predict** whether stock with particular history goes up or down on that day.
- Suppose $X_1, \ldots, X_p$ are characteristics of a real estate, while Y - price of that real estate. We wish to **predict** the price of that real estate, to figure out if it is over- or undervalued.

Prediction.

In all those situations, we use model

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

and estimate $f(X)$ with $\hat{f}(X)$, which can then be used for **prediction**:

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

since the error term ϵ averages to zero.

Here, $\hat{f}$ is often treated as a black box:

- we aren't concerned with the exact form of $\hat{f}$,
- given that it yields accurate predictions for Y.

Most prolific black box models in statistical learning:

- Support Vector Machines (covered in this class),
- Artificial Neural Networks (not covered),
- Random Forests (not covered).

Inference.

We often aim to understand the way that response Y is affected by change in predictors $X_1, \ldots, X_p$, which leads to **inference**.

Some of the general questions of interest when conducting inference:

- Which predictors are associated with the response? It's often the case that only a small subset of predictors actually affects Y.
- What is the relationship between the response and each predictor? Predictor X_1 may have a positive relationship with Y (as $X_1 \Uparrow$, $Y \Uparrow$), X_2 - negative (as $X_2 \Uparrow$, $Y \Downarrow$), etc.
- Can the relationship between response and predictors be well summarized via linear equation, or is it more complicated?

Inference.

As in **prediction**, in **inference** we wish to estimate f from

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

BUT unlike **prediction**, our main goal is to

- **interpret** the relationship between response Y and predictors $X_1, \ldots, X_p$,
- rather than just making best possible predictions for Y.

In **inference**, the estimate $\hat{f}$ can't be treated as a black box, we need to know its exact form.

Most popular inferential models in statistical learning:

- Linear/Logistic Regression
- LASSO regression
- Decision Trees

Inference: *Advertising* example.

Example. *Advertising* data set contains information on sales of a particular product given TV, radio & newspaper advertisement budgets. Hence,

- *Sales* - response variable Y,
- TV, radio and newspaper advertisement budgets constitute predictors $X = (X_1, X_2, X_3)$

While conducting inference, one may ask:

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

A lot of these questions might be partially answered via multiple linear regression, where

$$Y = f(X) \equiv \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Inference: Product Purchase example.

Example. Presume that for shopping items you would like to model

- whether it will be purchased or not (categorical response Y),
- depending on such item properties as price, store location, discount levels, competition price, etc (predictors).

Here, one is most interested in how each of the individual variables affects the probability of purchase.

For instance,

- what effect will changing the price of a product have on sales?
- stores at which locations yield the highest sales?
- is the price in competitor stores an important factor?

Example for both Prediction & Inference.

Example. In a real estate setting, one may seek to relate

- values of homes (response Y),
- to predictors such as crime rate, zoning, distance from a river, air quality, schools, income level of community, size of houses, etc

Here, one might be interested in both

- **inference**: how the individual input variables affect prices, e.g.
  - how much extra will a house be worth if it has a view of the river?
  - how does the neighborhood crime rate affect the price?

- **prediction**: one may simply be interested in predicting the value of a home given its characteristics: is this house under- or over-valued?

Different Methods for Different Goals.

Depending on whether our ultimate goal is

- prediction,
- inference, or
- a combination of the two,

different methods for estimating f may be appropriate. For example,

- **Linear models** lead to simple and interpretable inference, but may not yield as accurate predictions, is not as flexible.
- In contrast, some **non-linear approaches** (e.g. **neural networks**, **support vector machines**) can potentially provide quite accurate predictions for Y ($\Longrightarrow$ are very flexible) but this comes at the expense of a less interpretable model for which inference is more challenging.

Interpretability vs Flexibility.

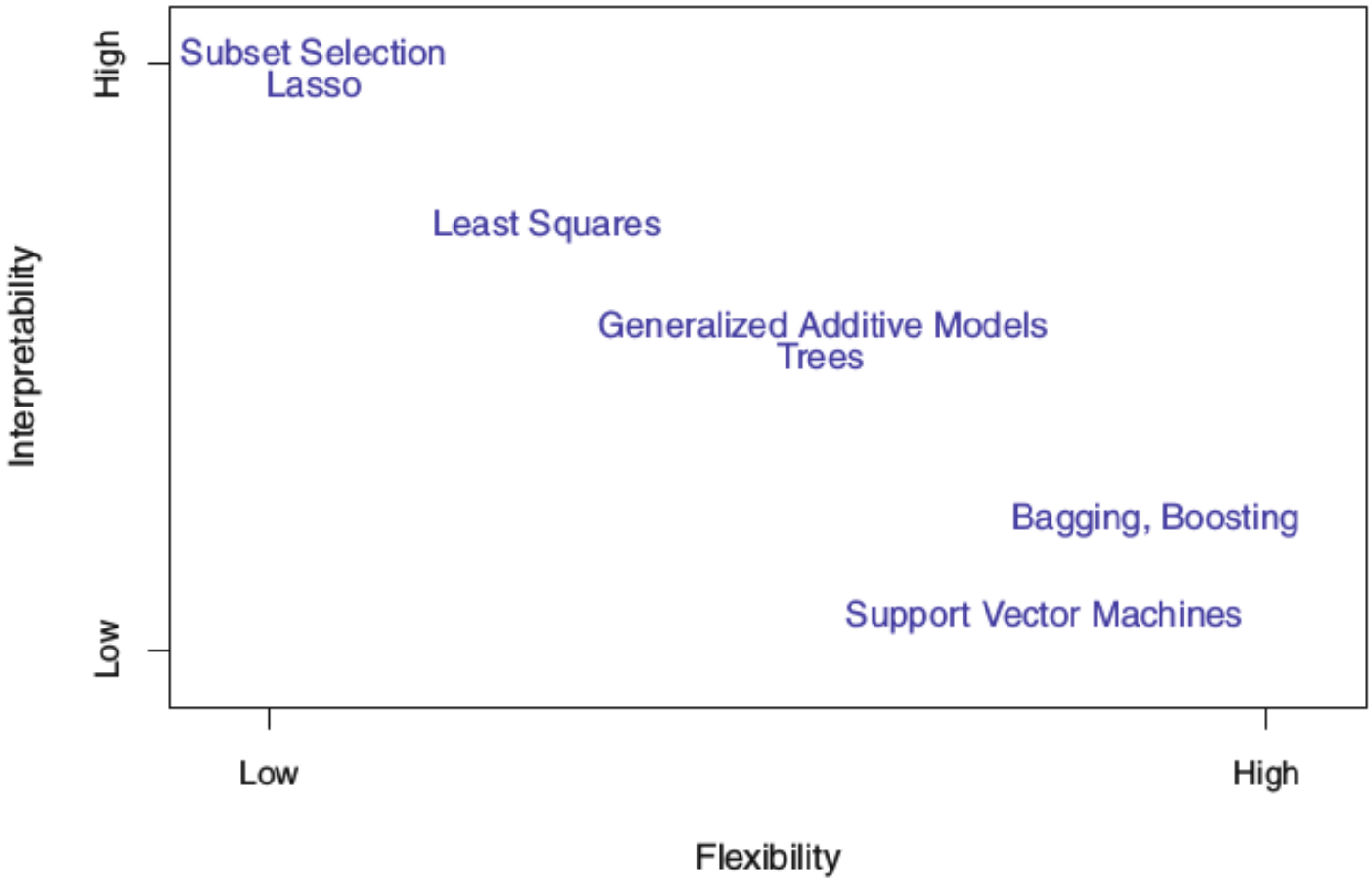


FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

Reducible & Irreducible Errors.

Having talked about various methods to estimate function f in model $Y = f(X) + \epsilon$ (depending on the ultimate goal), let's mention that one commits two types of error when estimating Y with $\hat{Y} = \hat{f}(X)$:

- reducible error:

$$f(X) - \hat{f}(X)$$

  Can be reduced via using a better statistical learning technique to come up with $\hat{f}$ estimate of true function f.

- irreducible error:

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

  No matter how well $\hat{f}(X)$ estimates the true function f, one can't avoid the random ϵ error due to the model. E.g. one's $\hat{f}$ could be a perfect estimate of $f \implies \hat{f}(X) \equiv f(X)$, but then

$$Y - \hat{Y} = (f(X) + \epsilon) - \hat{f}(X) = (f(X) + \epsilon) - f(X) = \epsilon$$

Why have ϵ in the model?

So why do we have that irreducible error ϵ in our model $Y = f(X) + \epsilon$? The quantity ϵ may contain

- unmeasured variables (ones not among our $X_1, \ldots, X_p$) that are useful in predicting Y; we don't measure them $\implies$ $f(X)$ doesn't use them for prediction.
- unmeasurable **natural** variation. For example, the risk of an adverse reaction might vary for a given patient on a given day, depending on
  - manufacturing variation in the drug itself, or
  - the patient's general feeling of well-being on that day.

  which are characteristics that are tough to keep track of & account for in systematic part of the model.

Reducible & Irreducible Errors: Mathematical Formulation (for curious students).

Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = \hat{f}(X)$. Assume for a moment that both $\hat{f}$ and X are fixed. Then, it is easy to show that

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}, \end{aligned} \tag{2.3}$$

where $E(Y - \hat{Y})^2$ represents the average, or *expected value*, of the squared difference between the predicted and actual value of Y, and $\text{Var}(\epsilon)$ represents the *variance* associated with the error term ϵ.

The focus of this book is on techniques for estimating f with the aim of minimizing the reducible error. It is important to keep in mind that the irreducible error will always provide an upper bound on the accuracy of our prediction for Y. This bound is almost always unknown in practice.