Introduction to Statistical Learning

(ISLR 2.1)

Yingbo Li

Southern Methodist University

STAT 4399

Outline

The Advertising data

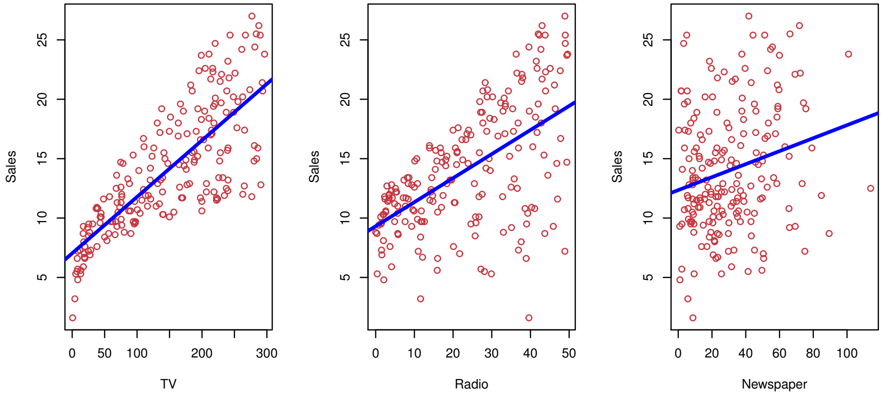
For $n = 200$ different markets

- `Sales`: sales of the product in this market (Y)
- `TV`: advertising budget for TV (X_1)
- `Radio`: advertising budget for radio (X_2)
- `Newspaper`: advertising budget for newspaper (X_3)

```
> Advertising = read.csv(file = 'Advertising.csv');
> head(Advertising)
  X    TV Radio Newspaper Sales
1 1 230.1  37.8      69.2  22.1
2 2  44.5  39.3      45.1  10.4
3 3  17.2  45.9      69.3   9.3
4 4 151.5  41.3      58.5  18.5
5 5 180.8  10.8      58.4  12.9
6 6   8.7  48.9      75.0   7.2
```

We believe that there is a relationship between Y and X

- Y: output variable, response
- $X = (X_1, X_2, X_3)$: input variables, predictors



Model the relationship between Y and X

The regression function

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

- f: unknown function
- ϵ: random error with mean zero, i.e., $E(\epsilon) = 0$.

Model the relationship between Y and X

The regression function

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

- f: unknown function
- ϵ: random error with mean zero, i.e., $E(\epsilon) = 0$.

In the `Advertising` example:

$$f(X_1, X_2, X_3) = E(Y \mid X_1, X_2, X_3)$$

Statistical learning, and this course, are all about how to estimate f. Why?

- Prediction.
- Inference.

Prediction

If we can get a good estimate for f, we can make accurate predictions for the response Y, based on a new value of X.

- For a new market, given three media budgets, what's the sales?
- Just want to predict sales, not to know which media is more important.

Prediction

If we can get a good estimate for f, we can make accurate predictions for the response Y, based on a new value of X.

- For a new market, given three media budgets, what's the sales?
- Just want to predict sales, not to know which media is more important.

Suppose our estimate for f is $\hat{f}$, the output Y for input X is predicted as

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

Mean square error

$$E(Y - \hat{Y})^2 = \underbrace{E[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{V(\epsilon)}_{\text{Irreducible}}$$

Inference

We are often interested in understanding the relationship between that Y and each of $X_1, \ldots, X_p$. For example

1. Which predictors actually affect the response?
2. Is the relationship positive or negative?
3. Is the relationship a simple linear one or is it more complicated?

Inference

We are often interested in understanding the relationship between that Y and each of $X_1, \ldots, X_p$. For example

1. Which predictors actually affect the response?
2. Is the relationship positive or negative?
3. Is the relationship a simple linear one or is it more complicated?

- How much impact does TV budges have on the sales.
- Which media generate the biggest boost in sales?

How to estimate f

Use the training data and a statistical method to estimate f.

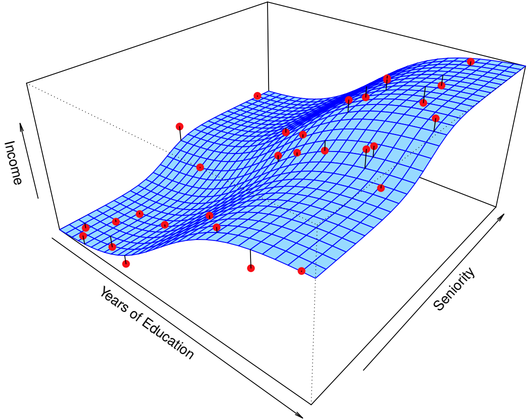
- We have observed a set of *training data*

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

  where each $x_i = (x_{i,1}, x_{i,2}, \dots, x_{i,p})'$, and y_i is a scalar.
- Statistical learning methods:
  - parametric
  - non-parametric

Income vs Education, Seniority

```
> Income = read.csv(file = 'Income2.csv')[, -1];
> head(Income)
  Education Seniority   Income
1  21.58621  113.1034 99.91717
2  18.27586  119.3103 92.57913
3  12.06897  100.6897 34.67873
4  17.03448  187.5862 78.70281
5  19.93103   20.0000 68.00992
6  18.27586   26.2069 71.50449
> dim(Income)
[1] 30  3
```



Parametric methods

Reduces the problem of estimating f down to one of estimating a (finite) set of parameters. A two-step model based approach:

1. Come up with a model (some functional form assumption about f). The most common example is a *linear model*.

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

- We only need to estimate $p+1$ parameters $\beta_0, \beta_1, \ldots, \beta_p$.
- Although it is almost never correct, a linear model often serves as a good and interpretable approximation to the unknown true $f(X)$.

Parametric methods

Reduces the problem of estimating f down to one of estimating a (finite) set of parameters. A two-step model based approach:

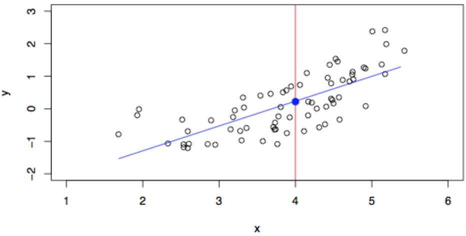
1. Come up with a model (some functional form assumption about f). The most common example is a *linear model*.

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

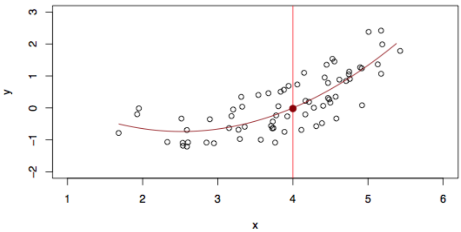
   - We only need to estimate $p+1$ parameters $\beta_0, \beta_1, \ldots, \beta_p$.
   - Although it is almost never correct, a linear model often serves as a good and interpretable approximation to the unknown true $f(X)$.

2. Use the training data to fit the model. Estimate the unknown parameters such as $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$.
   - The most common approach is ordinary least squares (OLS).
   - We will see later that there are other superior approaches.

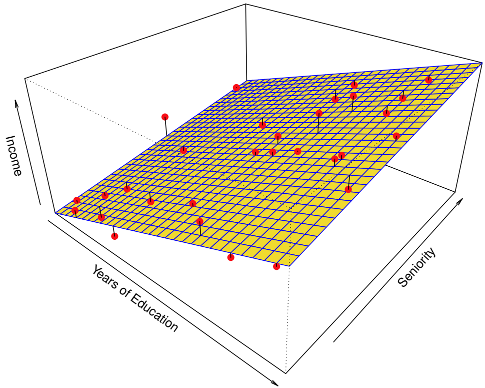
A linear model $\hat{f}_L(X) = \beta_0 + \beta_1 X$ gives a reasonable fit here.



A quadratic model $\hat{f}_Q(X) = \beta_0 + \beta_1 X + \beta_2 X^2$ fits slightly better.



A linear regression fit to the Income data

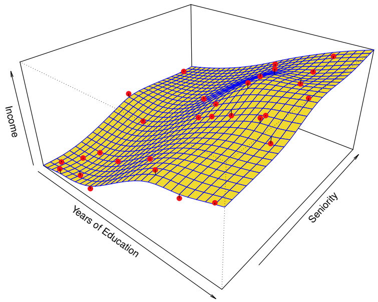


$$\text{Income} = \beta_0 + \beta_1 \times \text{Education} + \beta_2 \times \text{Seniority}$$

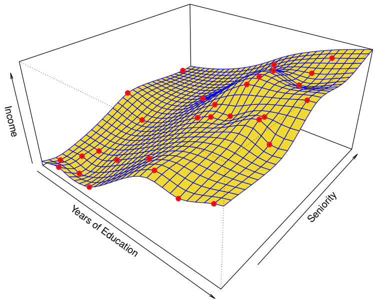
Non-parametric methods

They do not make explicit assumptions about the functional form of f.

- Advantages: accurately fit a wider range of possible shapes of f.
- Disadvantages: require large n to obtain an accurate estimate.



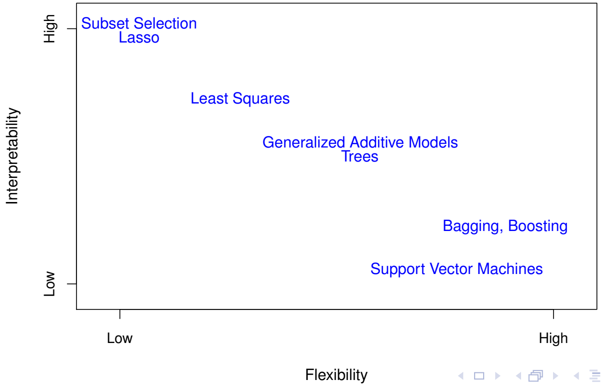
A smooth thin-plate spline fit: flexible



A rough thin-plate spline fit: overfitting

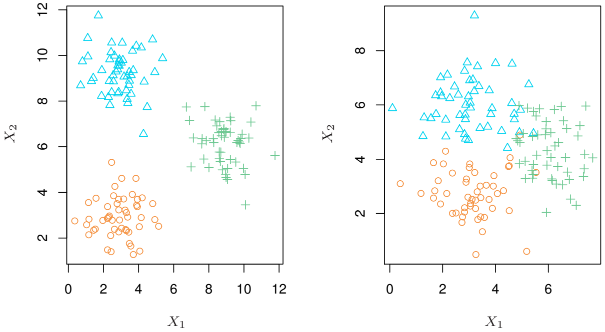
Some trade-offs

- Prediction accuracy vs model interpretability
  Linear models are easy to interpret; thin-plate splines are not.
- Good fit vs over-fit
  A model that overfits the training data may not predict well.



Supervised vs unsupervised learning

- Supervised learning: both X and Y are available
- Unsupervised learning: only X is available; there is no Y.
  - Example: market segmentation where we try to divide potential customers into groups based on their characteristics.
  - A common approach is *clustering*.



Regression vs classification

- Regression: Y is continuous (quantitative).
  - Predicting the value of the Dow in 6 months.
  - Predicting the value of a given house based on various inputs.
- Classification: Y is categorical (qualitative).
  - Will the Dow be up (U) or down (D) in 6 months?
  - Is this email a SPAM or not?