DASC 4113 Machine Learning

Ukash Nakarmi

Lecture 2

Learning Objectives

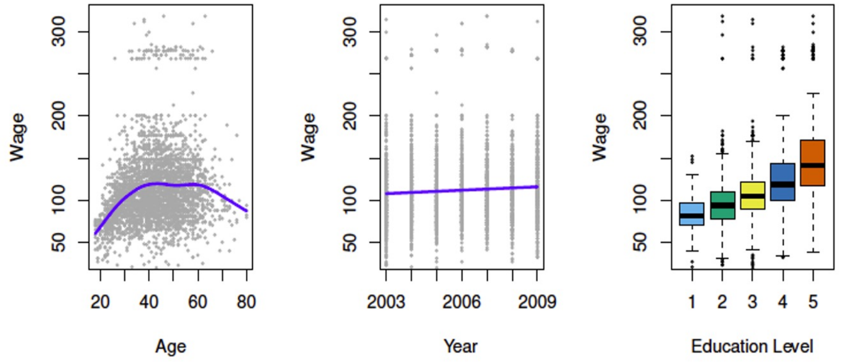
In this class, we will learn about following concepts:

- The function f() and its estimation
- How do we assess the accuracy of the model that estimates the function f()

To understand 1. and 2., we will introduce several tools/terms that we will see through out this course.

The function $f(\cdot)$

Relates the input variables with the output variable.



Some Notations and Key terms

Input variables

- Features
- Predictors
- Independent variables
- Variables
- Often denoted by the symbol X

Output variables

- Target variables
- Response
- Response variable
- Often denoted by the symbol Y

Examples:

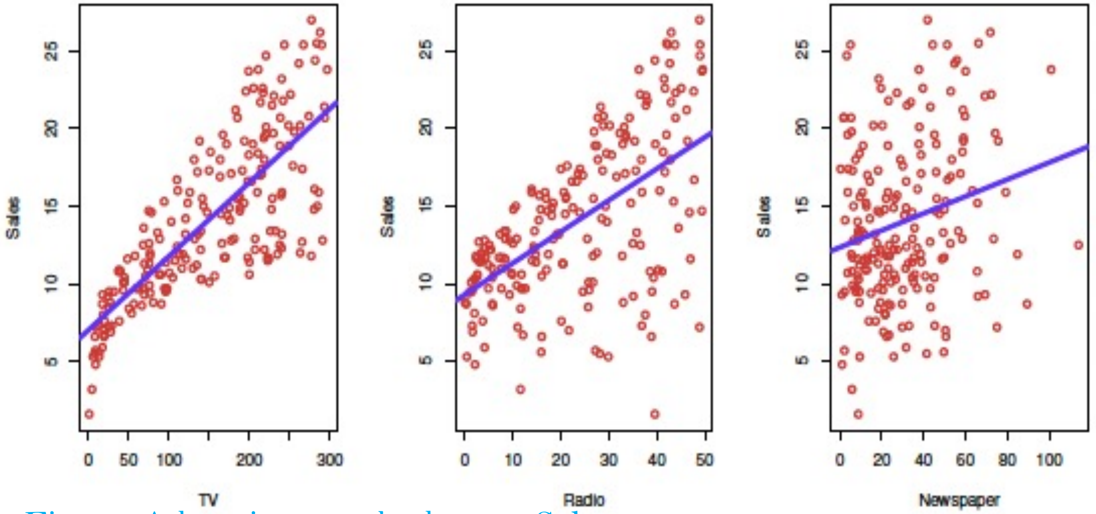


Fig 2.1 Advertisement budget vs Sales

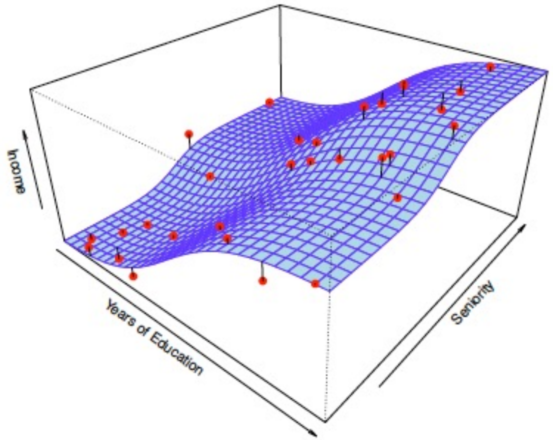


Fig 2.3 Education and Seniority vs Income

Inputs:

- TV budget (X_1)
- Radio Budget(X_2)
- Newspaper Budget (X_3)

Output:

- Sales (Y)

Inputs:

- Years of Education (X_1)
- Seniority(X_2)

Output:

- Income (Y)

Back to the function $f(\cdot)$

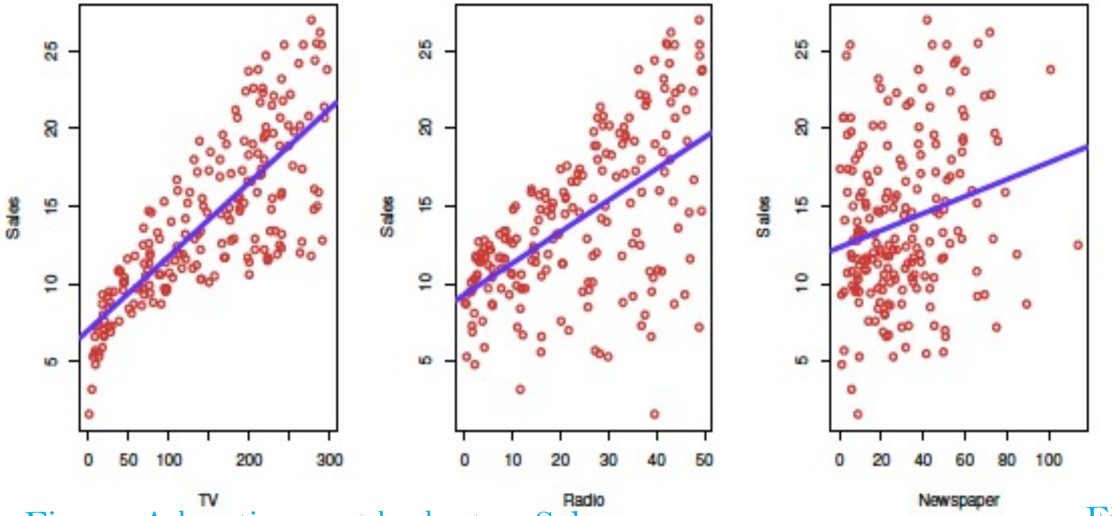


Fig 2.1 Advertisement budget vs Sales

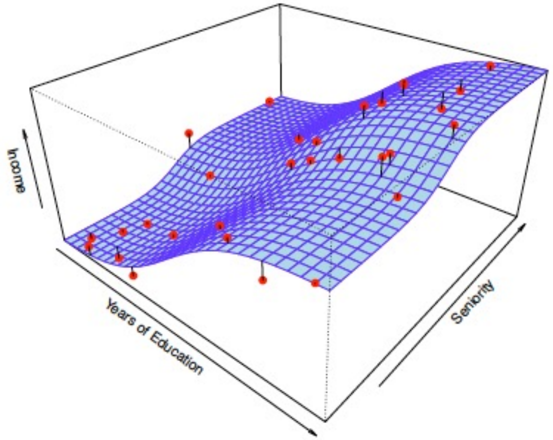


Fig 2.3 Education and Seniority vs Income

- In both figures, we see some relation between I/Ps and O/P.
- If we know this function f(), that relates I/P (X) to O/P (Y),

What could we do?

Back to the function $f(\cdot)$

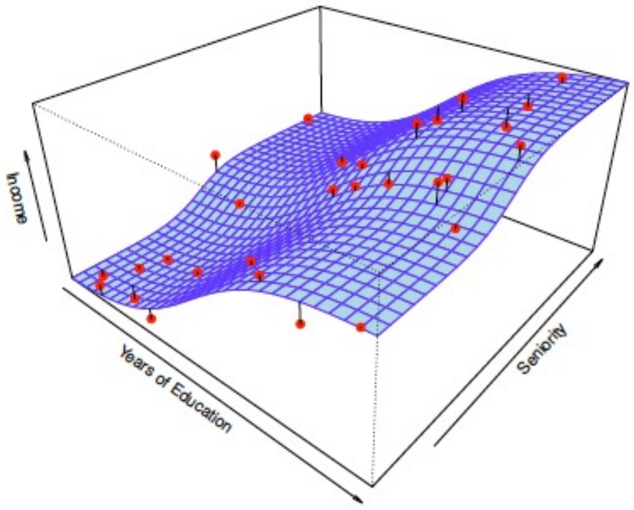


Fig 2.3 Education and Seniority vs Income

What could we do?

- We can make some predictions.

(Given some values for I/Ps (X), we can predict what could be the O/P (Y)

Example: Given a person's age is 35 and seniority level is 8, what could be his/her income?

The function $f(\cdot)$

The relation between I/Ps and O/P can be modeled as:

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

Random error term independent of X.

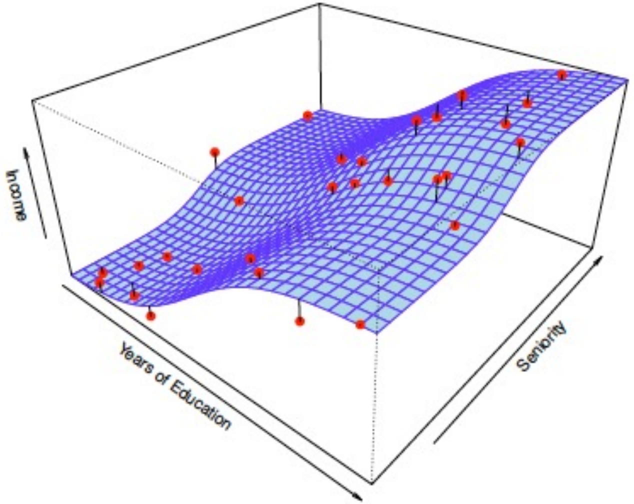


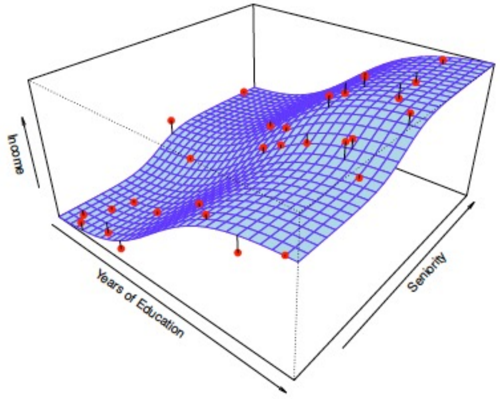
Fig 2.3 Education and Seniority vs Income

Most of the times, when we are doing machine learning/statistical learning based on data, we are simply trying to estimate this function $f(\cdot)$

Estimation of function $f(\cdot)$ and the Error

In general, the function $f(\cdot)$ is unknown.

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



So, given some data, we estimate $f(\cdot)$ as $\hat{f}(X)$ such that:

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

Once we have an estimate $\hat{f}(X)$, we can make predictions $\hat{Y}$

Q. From above 2 equations, what factors would affect the accuracy of our prediction $\hat{Y}$?

Estimation of function $f(\cdot)$ and the Error

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

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

The accuracy of our prediction depends on two quantities:

1. Reducible Error

- Comes from ML/statistical technique that we use.
- Can be improved by choosing a better model.

2. Irreducible Error (ϵ)

- Comes from some unknown factor.
- Cannot be improved by improving model.
- No matter how well we improve estimation $\hat{f}(X)$.

Why ?

Estimation of function $f(\cdot)$ and the Error

2. Irreducible Error (ϵ)

- Comes from some unknown factor.
- Cannot be improved by improving model.
- No matter how well we improve estimation $\hat{f}(X)$.

Why ?

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

- Simply because the way we defined our model.
- We make an estimate of $f(\cdot)$ from X and use it to predict Y.
- But from our system model, $Y = f(X) + \epsilon$, Y not only depends on X but also on ϵ, which is independent of X.

The irreducible Error (ϵ)

Improving model estimation cannot improve ϵ.
Rather, important questions to ask are:

Q. What is this ϵ ?
Q. Why shouldn't we simply formulate our problem as :

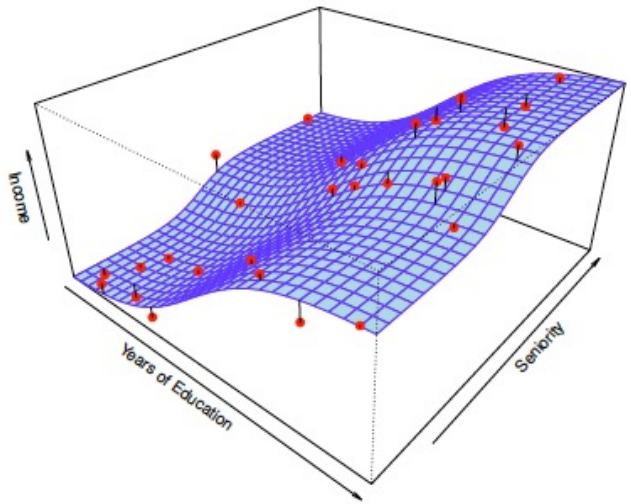
$$Y = f(X)$$

The irreducible Error (ϵ)

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

- $\epsilon \Rightarrow$ Accounts for all unknown causes (variables) that might influence our output.

Example:



- There must be several other factors besides years of education and seniority that could impact the income.
- So, no matter how well we fit the data, there is some irreducible error.
- Moreover, often we only have limited data.

(Sample might not represent the population exactly.)

ML people, Statisticians, Data Scientists being modest ☺

The Error in Prediction

- Given $\hat{f}(\cdot)$ as an estimate of $f(\cdot)$ and $\hat{Y}$ as a prediction of Y,
- We can express the error of prediction as :

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

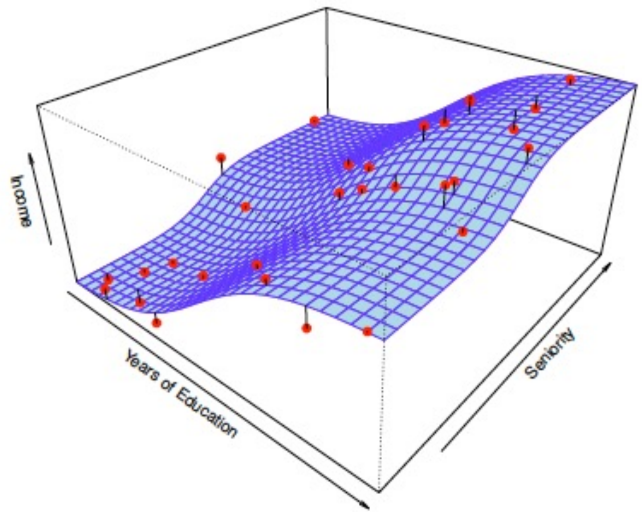
Our Goal is to reduce the reducible error

Is Prediction all we can do from $\hat{f}(\cdot)$?

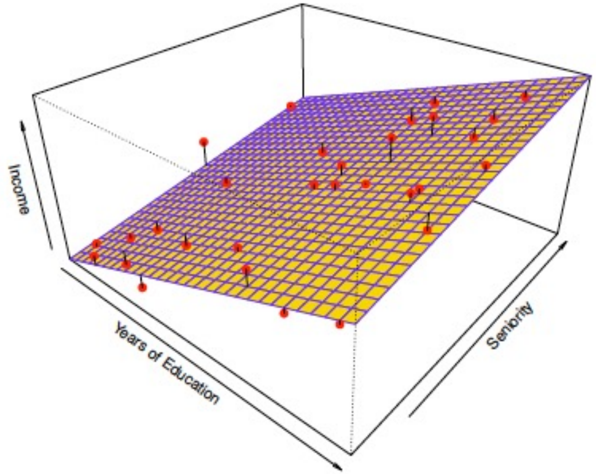
Is Prediction all we can do from $\hat{f}(\cdot)$?

Inference

What is the relation between individual predictors and the response?
Which predictors have role in the response?
Can we model the relation between predictors and O/P as a linear or do we need more complex modeling?



Non-linear modeling. (Red markers are data points observed)



Linear modeling

How do we estimate $f(\cdot)$

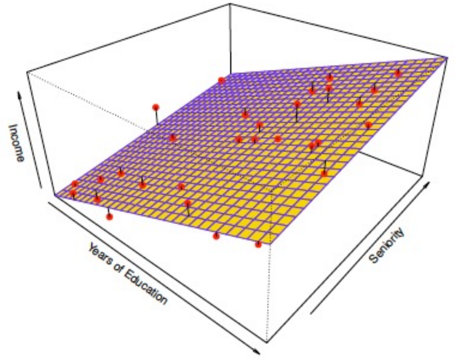
Now we have introduced concept of $f(\cdot)$, let look at two general ways we could use to estimate $f(\cdot)$.

A. Parametric

B. Non-Parametric

How do we estimate $f(\cdot)$

A. Parametric Method



Linear modeling

Two Steps Approach:

1. Make an assumption about the functional form of $f(\cdot)$
   Linear, Nonlinear ……

Example: Linear Assumption

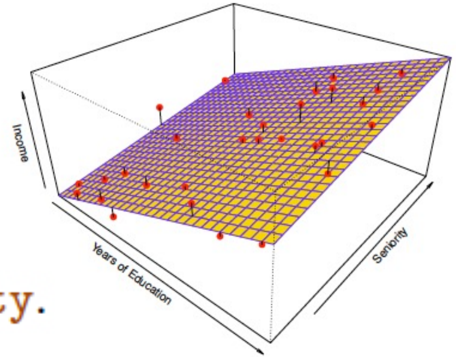
$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

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

How do we estimate $f(\cdot)$

A. Parametric Method

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

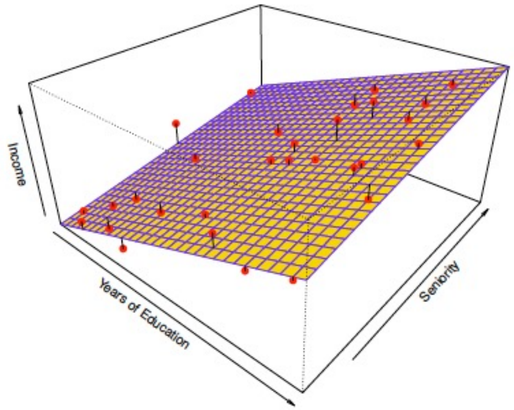


Linear modeling

Two Steps Approach:
2. Estimate parameters βs using given data set.
(Fit/Train in popular terms ☺)

- Some ways to fit are:
- Ordinary least squares
- LASSO
- Many other techniques

How do we estimate $f(\cdot)$



Linear modeling

A. Parametric Method

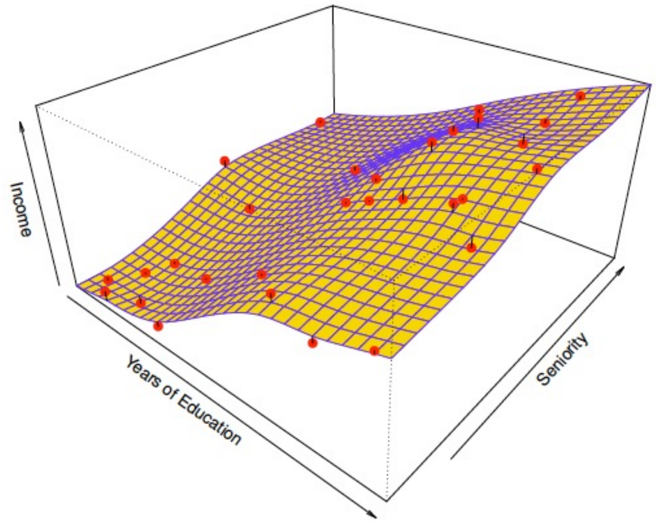
$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{education} + \beta_2 \times \texttt{seniority}.$$

- Simplifies the model
- More **Interpretable**
- Less **Flexible:** Model might not represent true $f(\cdot)$ very well.

How do we estimate $f(\cdot)$

B. Non-Parametric Method

- Does NOT make assumption about the functional form of $f(\cdot)$.
- More **flexible**
- Less **interpretable**.
- More parameters to estimate.

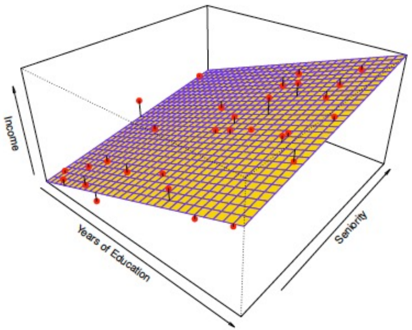


Smooth thin plate modeling of the income data.

Flexibility vs Interpretability

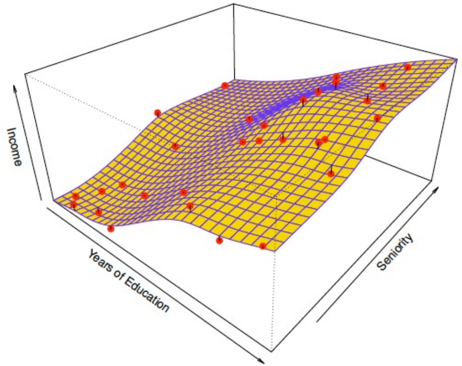
A. Parametric Method

- More **Interpretable**
- Less **Flexible**



B. Non-Parametric Method

- Less **Interpretable**.
- More **Flexible**



Interpretable : We have better idea about the exact form of relation between the predictors and output.

Flexibility: Governs how the model f() would change when there is some change in the training data.

Flexibility vs Interpretability

Lecture 2 | Part 2

Assessing Model Accuracy

- Key Terms:
- Evaluation Metric
- Training/Testing Data
- Interpretability/Flexibility
- Bias-Variance Tradeoff

Assessing Model Accuracy

1. Evaluation Metric : Tells us how good is our model.

Example: Mean Squared Error (MSE)

For Training data:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

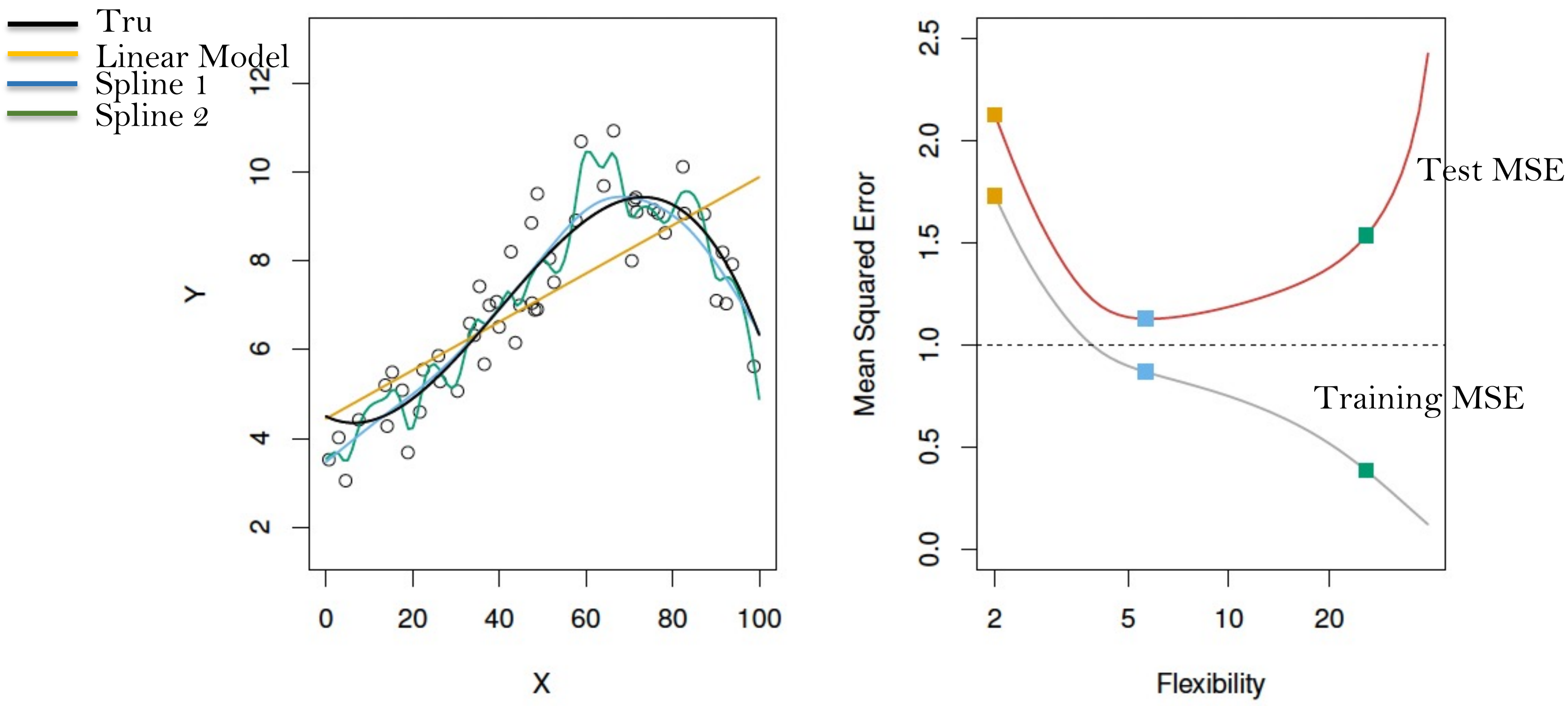
For Test data:

$$\mathrm{Ave}(y_0 - \hat{f}(x_0))^2.$$

Q. Why do we need to assess our model in training and also in the test data?

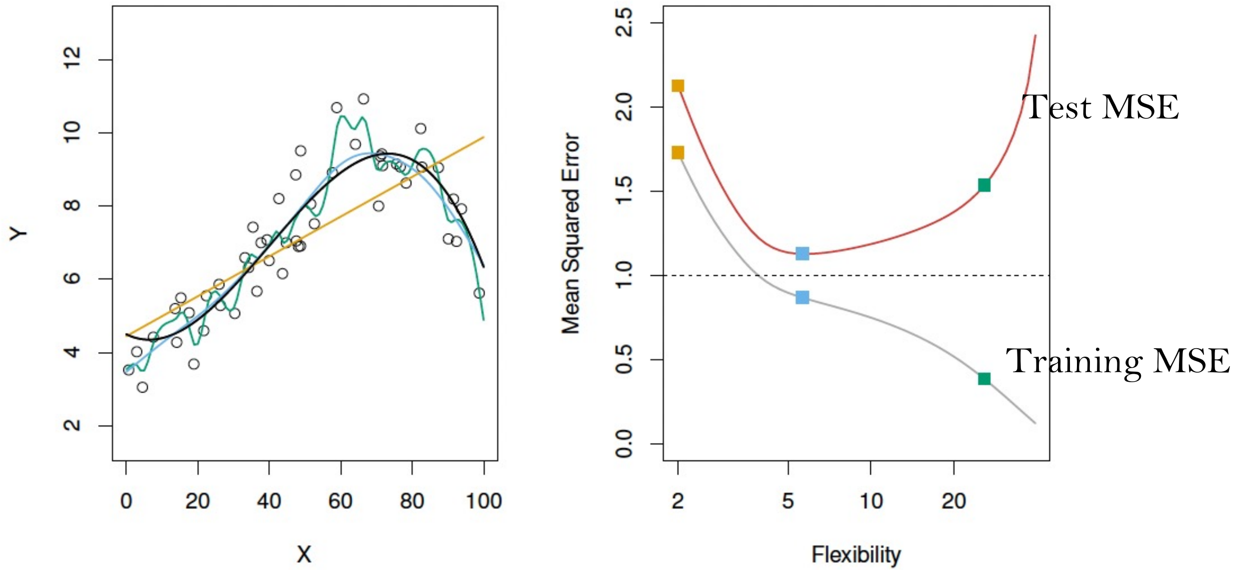
Assessing Model Accuracy

Relation between Training Metric value and the Test Metric value



Q. Which model would you choose? (Note the Spline 2 (Green) model has smallest training MSE)

Assessing Model Accuracy

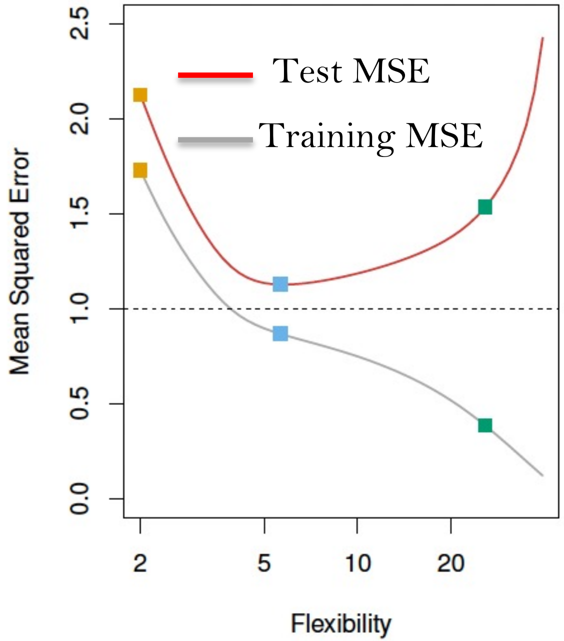


Q. Which model would you choose? (Note the Spline 2 (Green) model has smallest training MSE)

- Model with the minimum TEST MSE.
- In this case

Assessing Model Accuracy

Bias –Variance Tradeoff



From this graph, we see that Training MSE and the Test MSE of a model could tell very different stories.

This graph gives us one more important information:

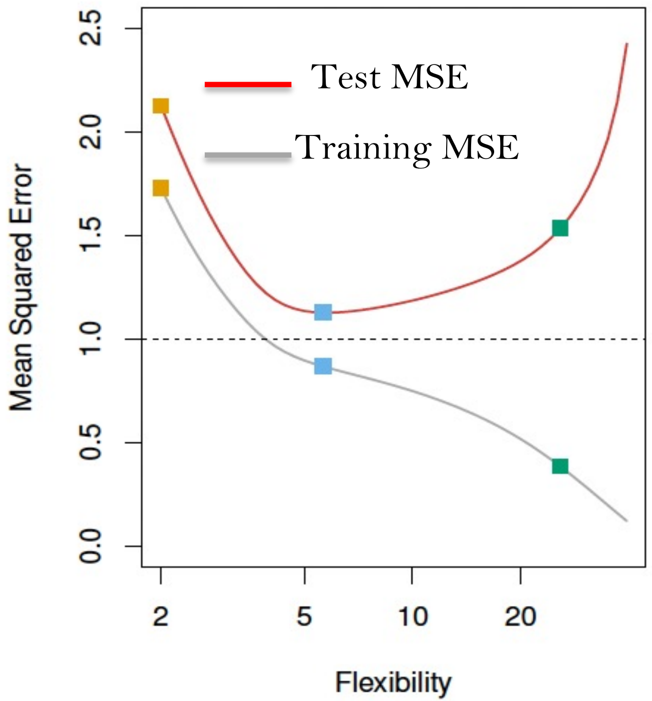
???

- Model with the minimum TEST MSE.
- In this case

Assessing Model Accuracy

Bias –Variance Tradeoff

- As flexibility of model increases, the training error decreases.
- But, if we keep increases the flexibility of a model, the train error will decrease but the test error might increase.



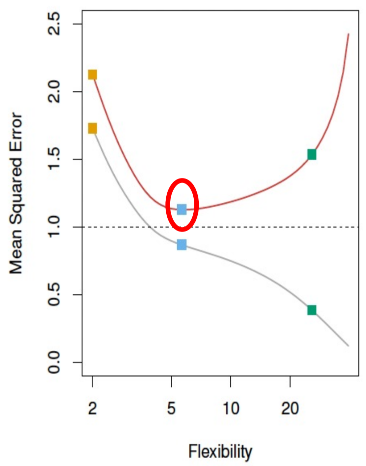
This introduces us to two **competing** properties of Statistical Learning methods:

Bias of $\hat{f}(\cdot)$

Variance of $\hat{f}(\cdot)$

Assessing Model Accuracy

Bias –Variance Tradeoff



Competing properties of Statistical Learning methods:

What is the Bias of $\hat{f}(\cdot)$? What is the Variance of $\hat{f}(\cdot)$?

Why do we call them Competing ?

Recall: We want to choose the model with minimum Test Error.
The Test Error can be expressed as:

Cannot do much about this term

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

Goal: Minimize the LHS.

- Cannot do much about Var(ϵ).
- So, we need to find a model that decreases the first two RHS terms simultaneously.

Assessing Model Accuracy

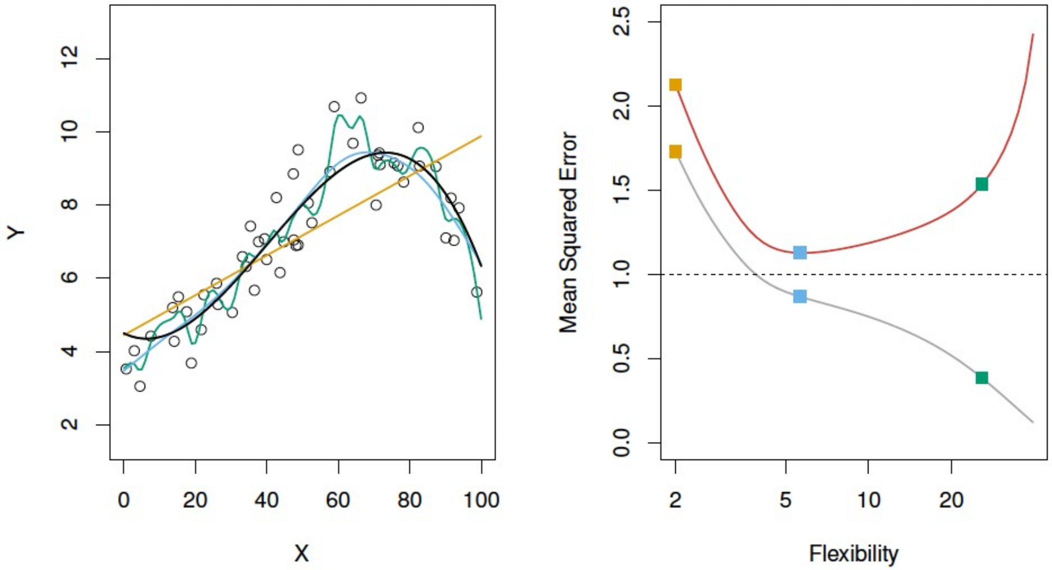
Bias –Variance Tradeoff

Variance:

Amount by which the $\hat{f}(\cdot)$ would change when the model is trained on a different training data

- Generally, higher the flexibility of the model, higher is is the variance.

Q. If we change two points in the figure in the left,



Assessing Model Accuracy

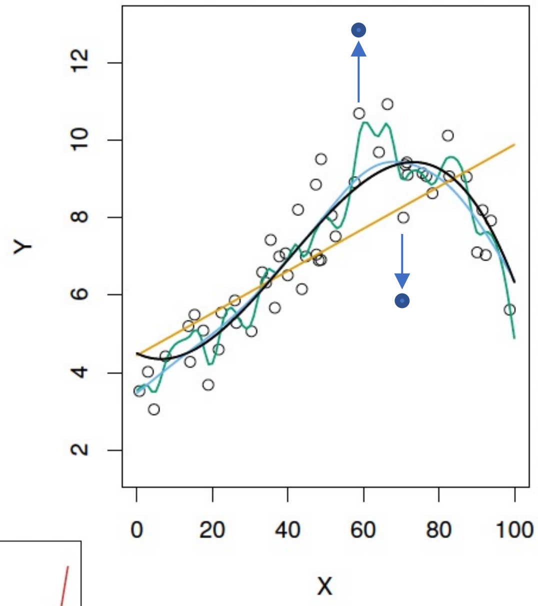
Bias –Variance Tradeoff

- Generally, higher the flexibility of the model, higher is is the variance.

Example:

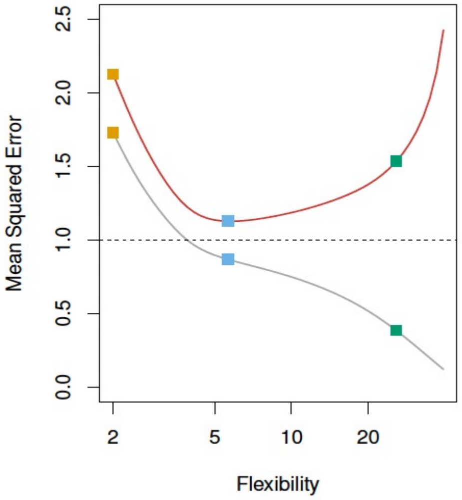
Q. If we change two data points in the figure, which model would change more?

- Linear (Yellow) ? Or
- Spline 2 (Green) ?



Q. Which one is more Flexible ?

- Linear (Yellow) ? Or
- Spline 2 (Green) ?



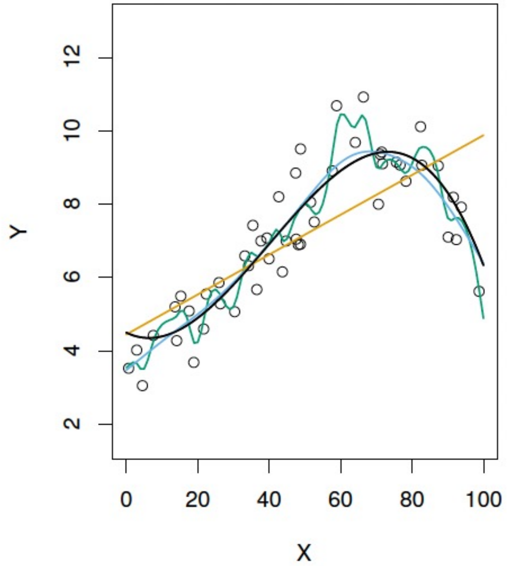
Assessing Model Accuracy

Bias –Variance Tradeoff

Bias:

Refers to the Error introduced by approximating real function f() by a simple function.

- Generally, higher the flexibility of the model, smaller is is the bias.
- No matter how many more training data we add might not be possible to accurately estimate the true f ()



Example: If we assume our model to be linear, no matter how many training samples we add, we can never accurately approximate true $f(\cdot)(Black\ curve\ in\ the\ figure.)$

Assessing Model Accuracy

Bias –Variance Tradeoff

- Generally, higher the flexibility of the model, higher is is the variance.
- Generally, higher the flexibility of the model, smaller is is the bias.

Flexibility $\uparrow$ => Variance $\uparrow$ Bias $\downarrow$

Flexibility $\downarrow$ => Variance $\downarrow$ Bias $\uparrow$

Cannot do much about this term

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

- We want a model that decreases both Var and Bias
- But, when we try to decrease one, the other increases

Competitive characteristics

Q. How shall we choose the model?

Assessing Model Accuracy

Bias –Variance Tradeoff

Q. Suppose we have two models A and B that fit the training data.

Model A

For every 1 unit decrease in Bias, variance increases by 3 units.

Model B

For every 1 unit decrease in Bias, variance increases by 0.001 units.

Choose Model A or Model B ?

Assessing Model Accuracy

Bias –Variance Tradeoff

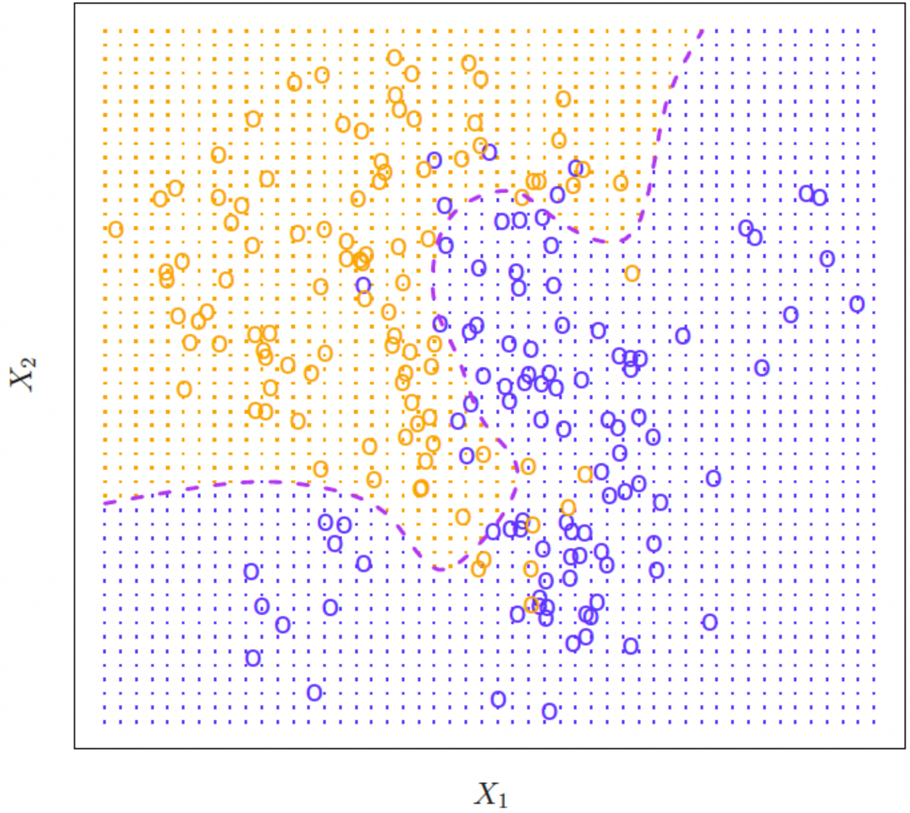
Q. Suppose we have two models A and B that fit the training data.

Model A	Model B
For every 1 unit decrease in Bias, variance increases by 3 units.	For every 1 unit decrease in Bias, variance increases by 0.001 units.

Choose Model A or Model B ?

Generally, we choose a model that has a higher ratio of decrease to increase.

Assessing Model Accuracy



Generally, we could do similar discussion for other type of problems such as Classificiation.

Take Away!

- The function f() and its estimation:
  - How to estimate,
    - Predict
    - Inference
    - The capability of estimate of f() (Reducible vs Irreducible Error)
    - Concept of Flexibility and Interpretability
- How do we assess the accuracy of the model that estimates the function f()
  - Concept of Bias and Variance of a model.
  - Competing nature of Bias and Variance of a model.
  - How to choose a better model?