Statistical Learning

What is Statistical Learning?

The Supervised Learning Problem

Starting point:

- Outcome measurement Y (also called dependent variable, response, target).
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables).
- In the regression problem, Y is quantitative (e.g price, blood pressure).
- In the classification problem, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
- We have training data $(x_1, y_1), \ldots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.

Objectives

On the basis of the training data we would like to:

- Accurately predict unseen test cases.
- Understand which inputs affect the outcome, and how.
- Assess the quality of our predictions and inferences.
- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working [simpler methods often perform as well as fancier ones!]

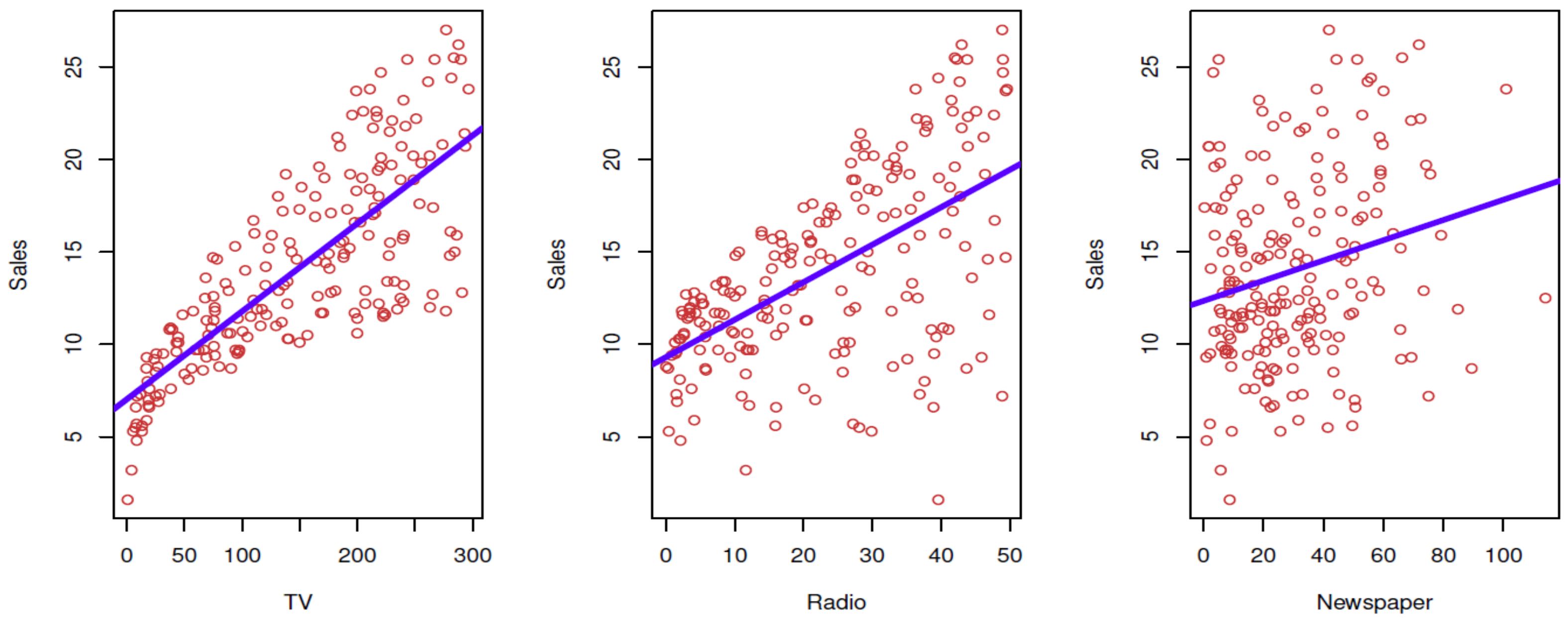
Unsupervised learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective is more fuzzy — find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- difficult to know how well your are doing.
- different from supervised learning, but can be useful as a pre-processing step for supervised learning.

Statistical Learning versus Machine Learning

- Machine learning arose as a subfield of Artificial Intelligence.
- Statistical learning arose as a subfield of Statistics.
- There is much overlap — both fields 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, and precision and uncertainty.
- But the distinction has become more and more blurred, and there is a great deal of “cross-fertilization”.
- Machine learning has the upper hand in Marketing!

What is Statistical Learning?



- Shown above are Sales vs TV, Radio and Newspaper
- Blue linear-regression lines are fitted separately to each.
- Can we predict Sales using TV, Radio and Newspaper ?
- We can construct a model to make the prediction.
- Sales ≈ f(TV, Radio, Newspaper)

Notation

- Sales is a *response* or *target* that we wish to predict.
- We usually use the variable Y to denote the response.
- TV, Radio, and Newspaper are known as features, inputs, or predictors
- We usually use the variables X_i to denote the set of features, inputs, or predictors
- We can refer to the input collectively as the *input vector*

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

- Now we can write our model as

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

where ϵ captures measurement errors and other discrepancies

What is $f(X)$ good for?

- With a good f we can make predictions of Y at new points $X = x$.
- We can understand which components of $X = (X_1, X_2, \ldots, X_p)$ are important in explaining Y.
- We can understand which components of X are irrelevant in explaining Y.

For example - Seniority and Years of Education have a big impact on Income, but Marital Status typically does not.

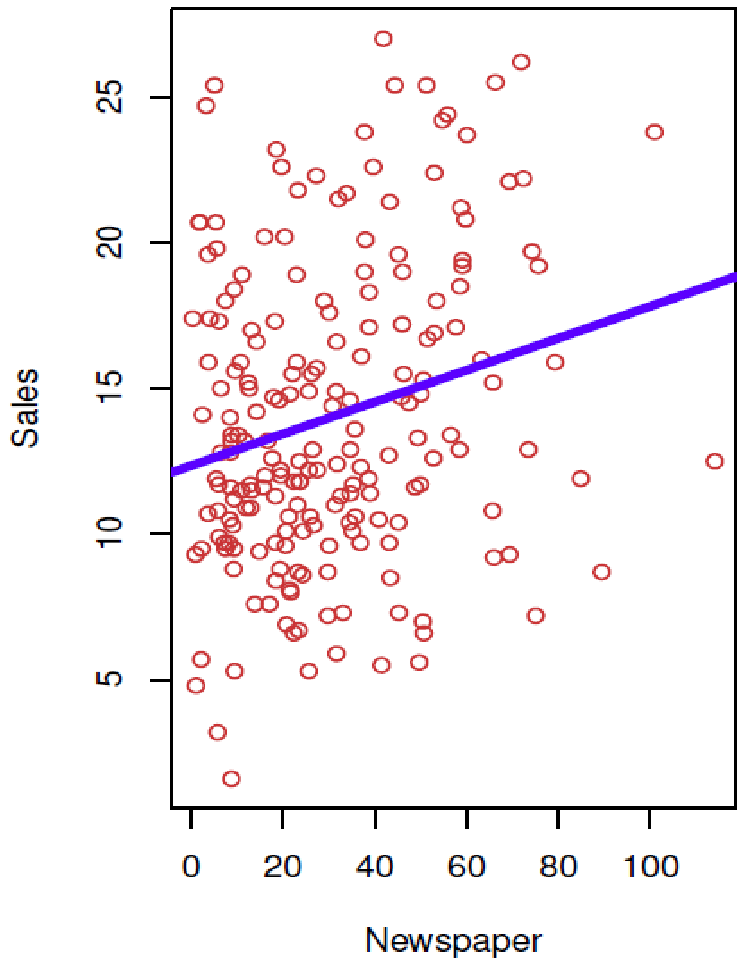
- Depending on the complexity of f, we may be able to understand how each component X_i of X affects Y.

Is there an ideal $f(X)$?

- In particular, what is a good value for $f(X)$ at any selected value of X, say X = 40?
- There can be many Y values at X = 40. A good value is

$$f(40) = E(Y|X = 40)$$

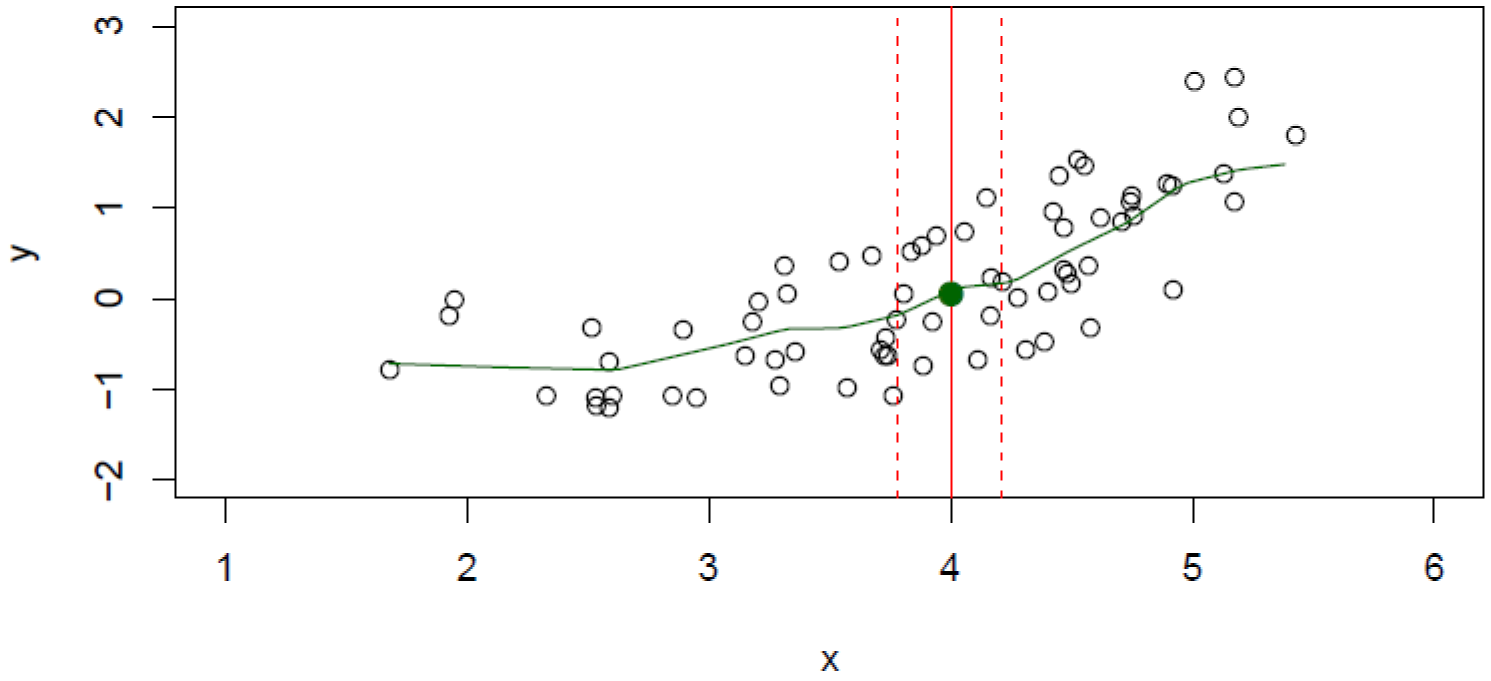
- $E(Y|X = 40)$ is the *expected value* (average) of Y given $X = 40$
- This ideal $f(x)$ = $E(Y|X = x)$ is called the *regression function*



How to estimate f

- Typically we have few if any data points with $X = 40$ exactly.
- So we cannot compute $E(Y|X = x)$!
- Relax the definition and let $\hat{f}(x) = E(Y|X \in N(x))$ where $N(x)$ is some neighborhood of x.
- Nearest neighbor averaging can be pretty good for small p — i.e. $p \leq 40$ and large-ish N.
- We will discuss smoother versions, such as kernel and spline smoothing later in the course.
- Nearest neighbor methods can be *lousy* when p is large. Reason: the *curse of dimensionality*. Nearest neighbors tend to be far away in high dimensions.
  - We need to get a reasonable fraction of the N values of y_i to average to bring the variance down—e.g. 10%.
  - A 10% neighborhood in high dimensions need no longer be local, so we lose the spirit of estimating $E(Y|X = x)$ by local averaging.

How to estimate f



Why Do We Estimate f?

- Statistical Learning, and this course, are all about how to estimate f.
- The term statistical learning refers to using the data to “learn” f.
- Why do we care about estimating f?
- There are 2 reasons for estimating f,
  - Prediction and
  - Inference

1. Prediction

- If we can produce a good estimate for f (and the variance of ϵ is not too large) we can make accurate predictions for the response, Y, based on a new value of X.
- Example: Direct Mailing Prediction
  - Interested in predicting how much money an individual will donate based on observations from 90,000 people on which we have recorded over 400 different characteristics.
  - Don’t care too much about each individual characteristic.
  - Just want to know: For a given individual should I send out a mailing?

1. Prediction

- If we can produce a good estimate for f (and the variance of ϵ is not too large) we can make accurate predictions for the response, Y, based on a new value of X.
- Example: Direct Mailing Prediction
  - Interested in predicting how much money an individual will donate based on observations from 90,000 people on which we have recorded over 400 different characteristics.
  - Don't care too much about each individual characteristic.
  - Just want to know: For a given individual should I send out a mailing?

2. Inference

- Alternatively, we may also be interested in the type of relationship between Y and the X's.
- For example,
  - Which particular predictors actually affect the response?
  - Is the relationship positive or negative?
  - Is the relationship a simple linear one or is it more complicated etc.?

- Example: Housing Inference
  - Wish to predict median house price based on 14 variables.
  - Probably want to understand which factors have the biggest effect on the response and how big the effect is.
  - For example how much impact does a river view have on the house value etc.

How Do We Estimate f?

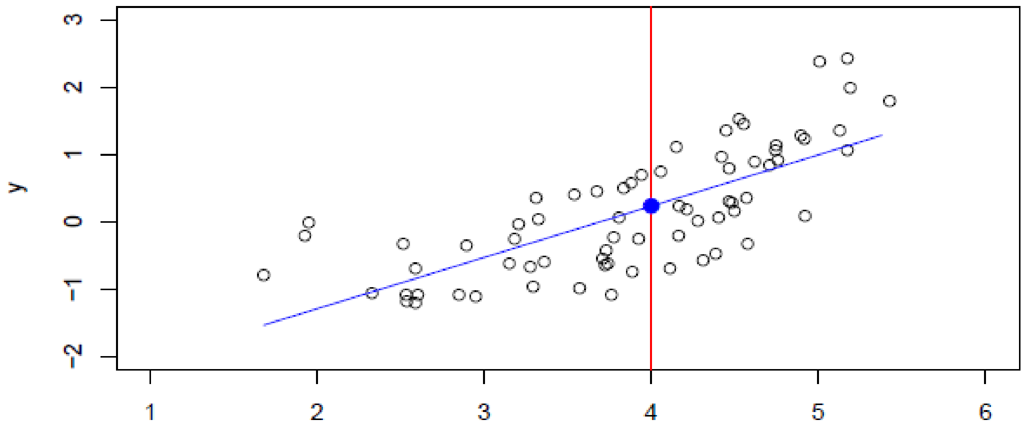
- We will assume we have observed a set of training data
- We must then use the training data and a statistical method to estimate f.
- Statistical Learning Methods:
  - Parametric Methods
  - Non-parametric Methods

Parametric and structured models

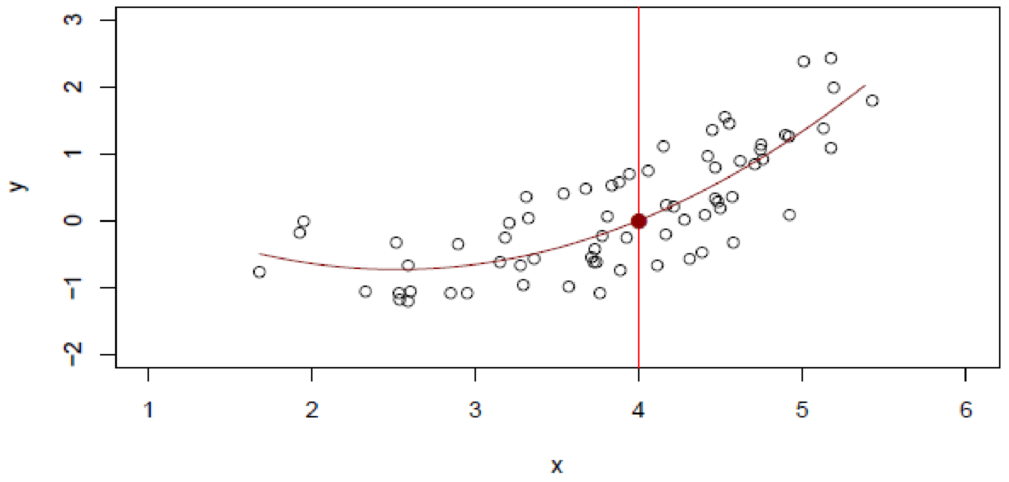
- The *linear* model is an important example of a parametric model:
$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p.$$
- A linear model is specified in terms of p + 1 parameters $\beta_0, \beta_1, \ldots \beta_p$.
- We estimate the parameters by fitting the model to training data.
- Although it is *almost never correct*, a linear model often serves as a good and interpretable approximation to the unknown true function $f(X)$.

Linear VS nonlinear

- A linear model $\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ gives a reasonable fit here



- A quadratic model $\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ fits slightly better.



The regression function $f(X)$

- Is also defined for vector X; e.g.

$$f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

- Is the ideal or optimal predictor of Y with regard to mean-squared prediction error: $f(x) = E(Y|X = x)$ is the function that minimizes

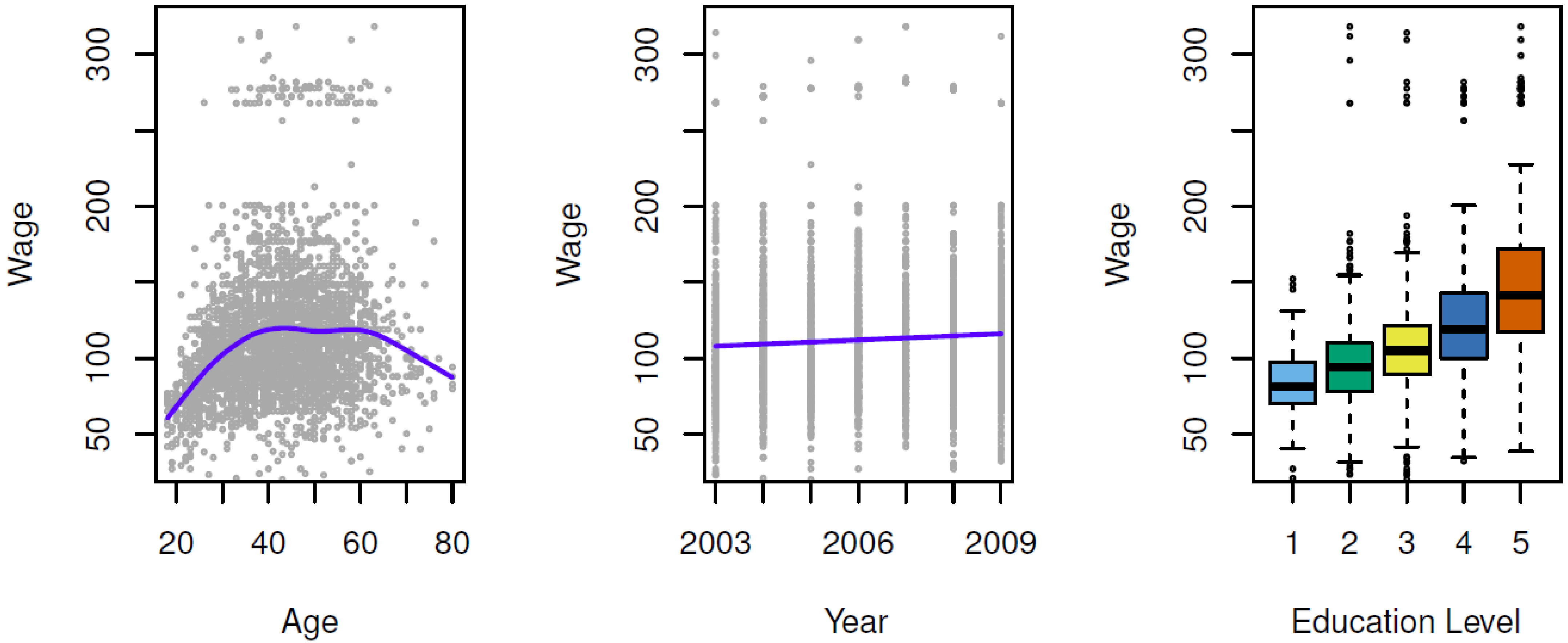
$E[(Y - g(X))^2 | X = x]$ over all functions g at all points $X = x$.

- $\epsilon = Y - f(x)$ is the irreducible error — i.e. even if we knew $f(x)$, we would still make errors in prediction, since at each $X = x$ there is typically a distribution of possible Y values.

- For any estimate $\hat{f}(x)$ of $f(x)$, we have

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

Example: Data for Salary vs. Age and Education

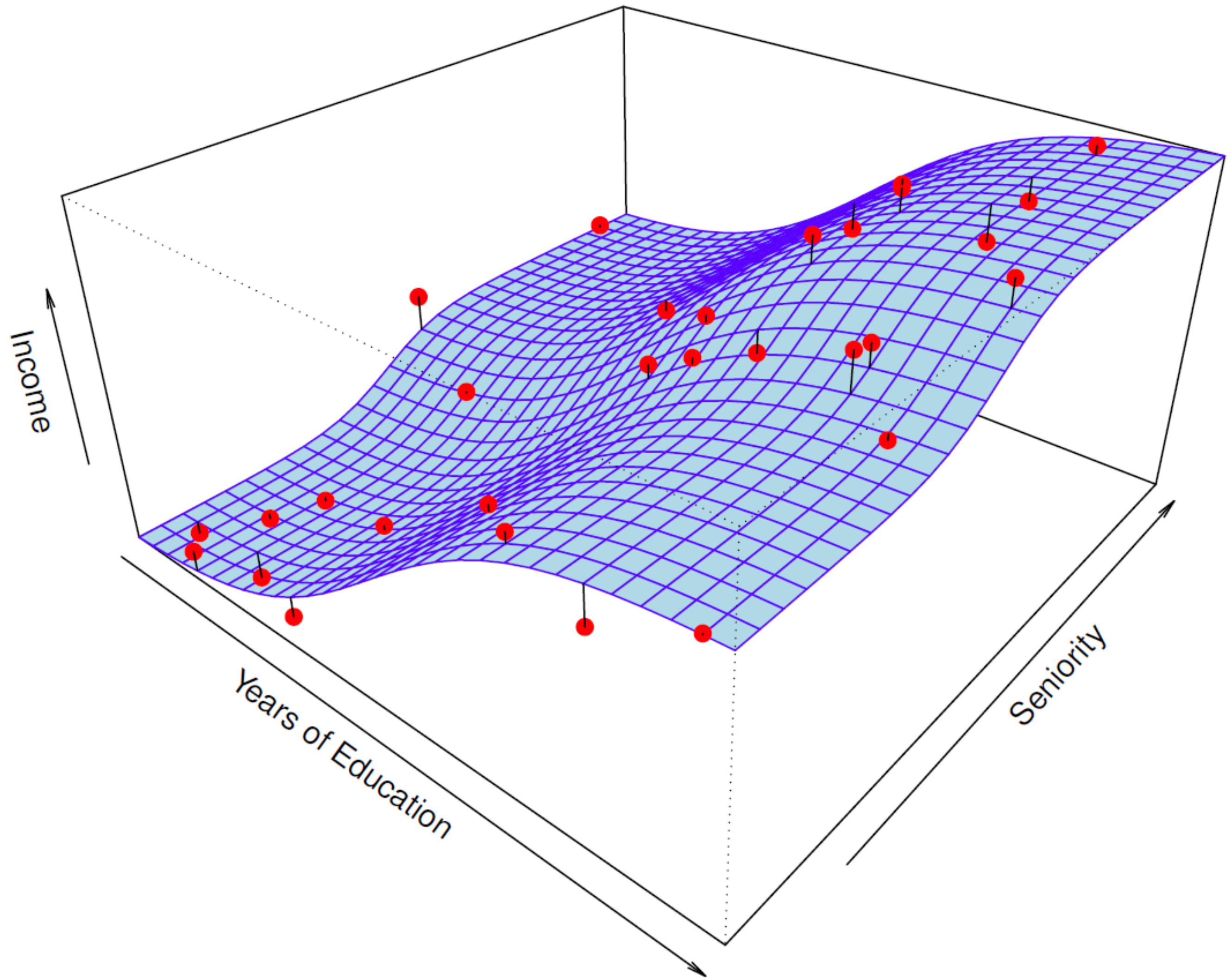


Simulated Example

Red points are simulated values for income from the model

Income = f(education, seniority) + ϵ

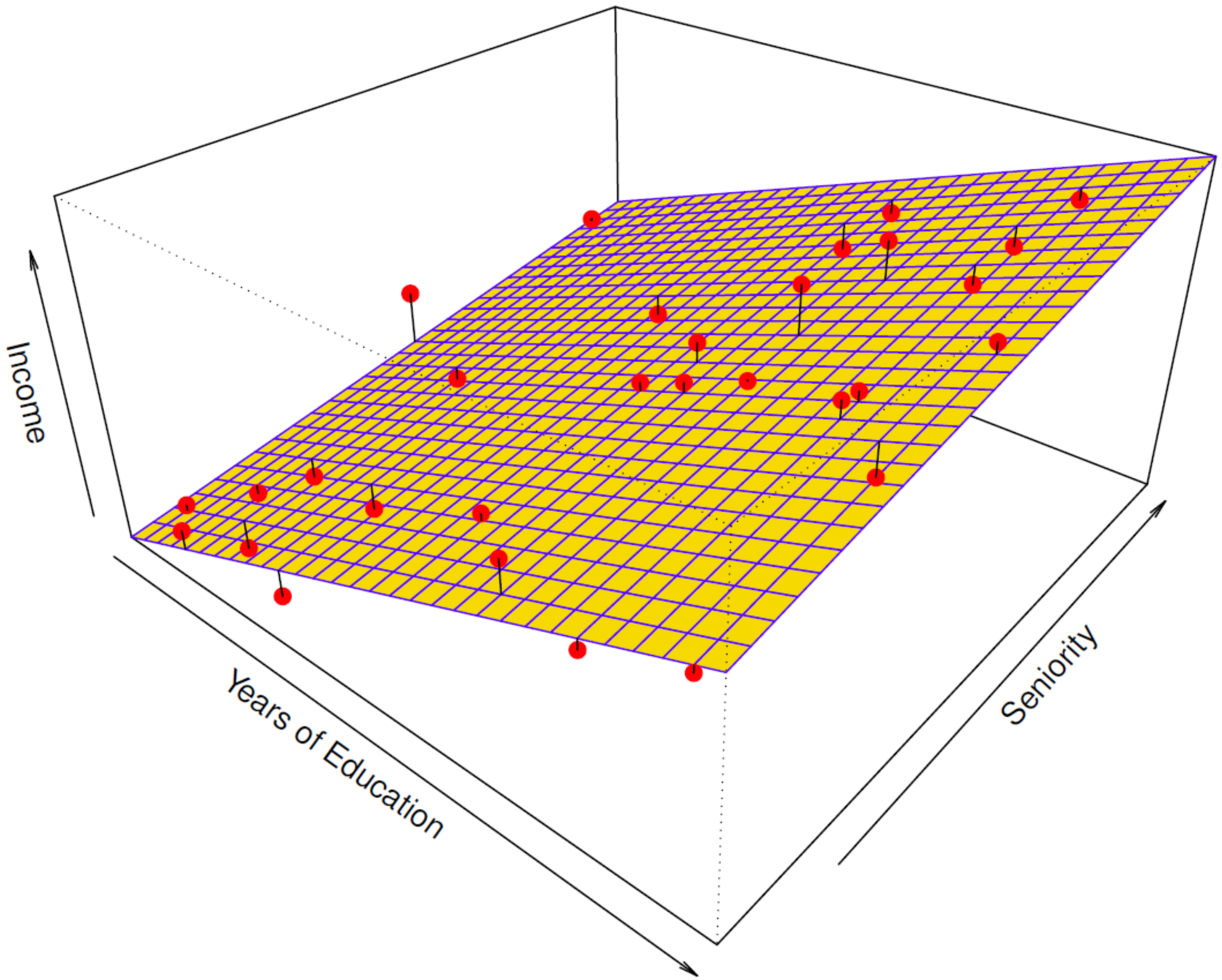
f is the *blue* surface



Example Cont.

Linear regression model t to the simulated data.

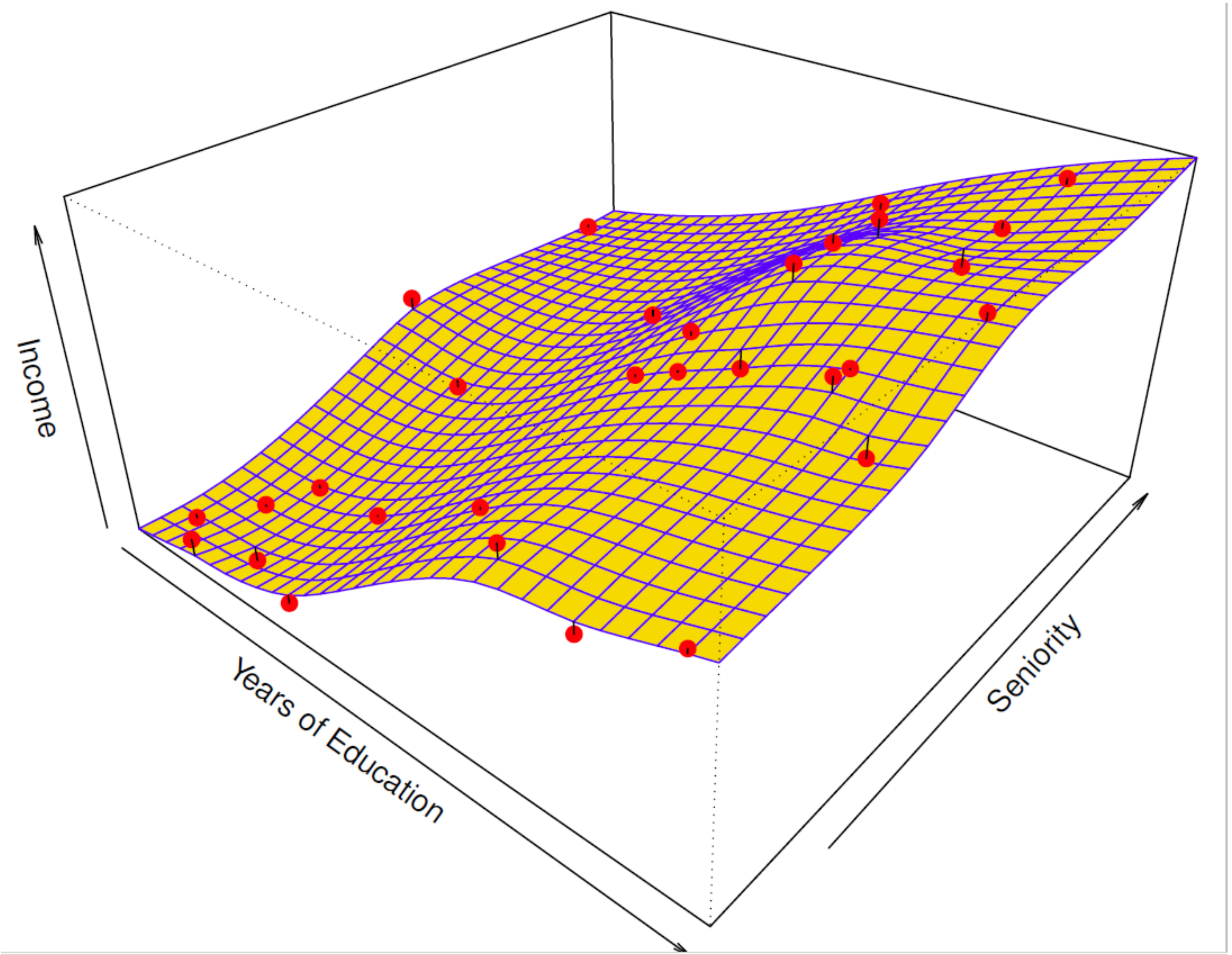
$$\hat{f}_L(education, seniority) = \hat{\beta}_0 + \hat{\beta}_1 \times education + \hat{\beta}_2 \times seniority$$



Example Cont.

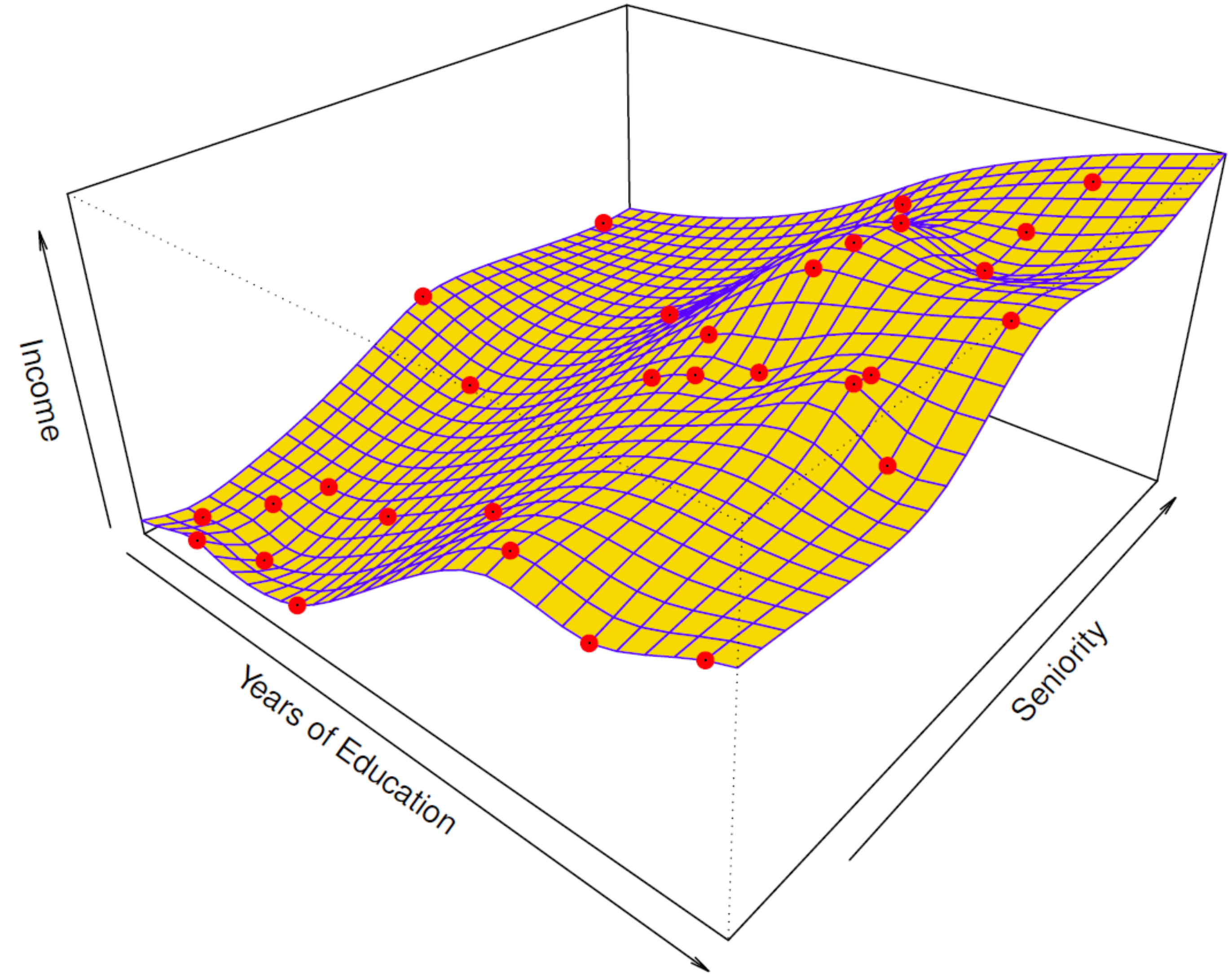
More flexible regression model $\hat{f}_S(\color{red}{education}, \color{red}{seniority})$ fit to the simulated data. Here, a technique called a thin-plate spline is used to fit a flexible surface.

(Chapter 7)



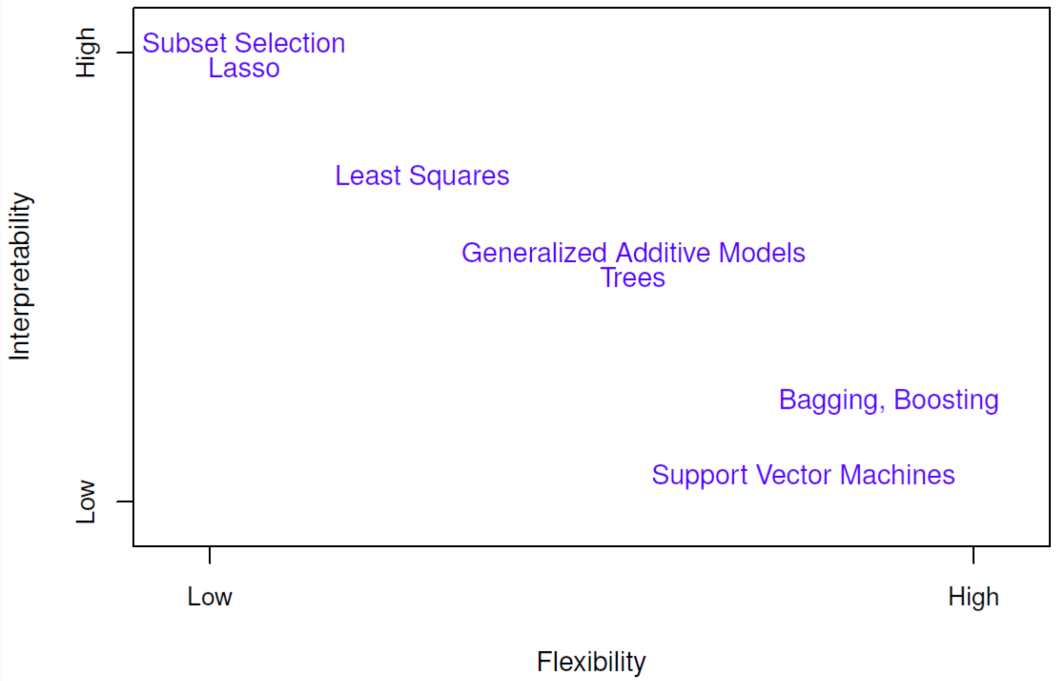
Example cont.

Even more flexible spline regression model $\hat{f}_S(\textit{education}, \textit{seniority})$ fit to the simulated data. Here the fitted model makes no errors on the training data! Also known as overfitting.



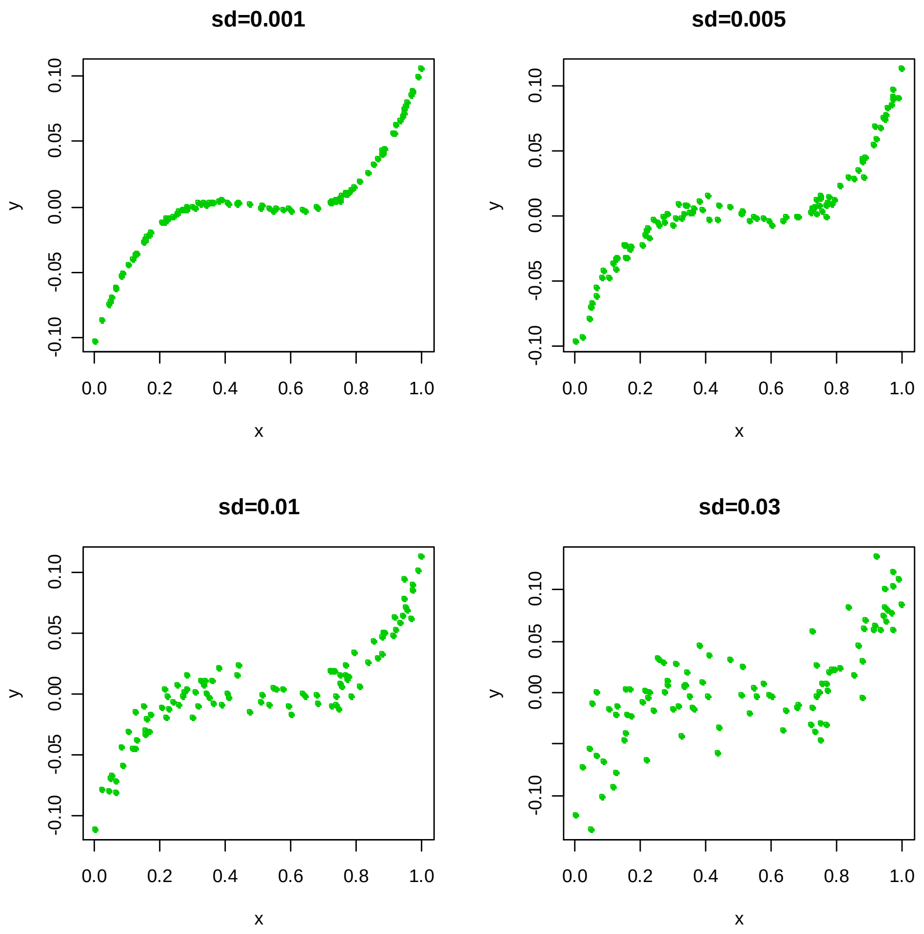
Trade-offs

- Prediction accuracy versus interpretability.
  - Linear models are easy to interpret; thin-plate splines are not.
- Good fit versus over-fit or under-fit.
  - How do we know when the fit is just right?
- Parsimony versus black-box.
  - We often prefer a simpler model involving fewer variables over a black-box predictor involving them all.



Different Standard Deviations

- The difficulty of estimating f will depend on the standard deviation of the ϵ's.
- The smaller the standard deviation of the data, the more accurate the estimate of f



Assessing Model Accuracy

Suppose we fit a model $\hat{f}(x)$ to some training data $Tr = \{x_i, y_i\}_1^N$, and we wish to see how well it performs.

- We could compute the average squared prediction error over Tr:

$$MSE_{Tr} = \text{Ave}_{i \in Tr} [y_i - \hat{f}(x_i)]^2$$

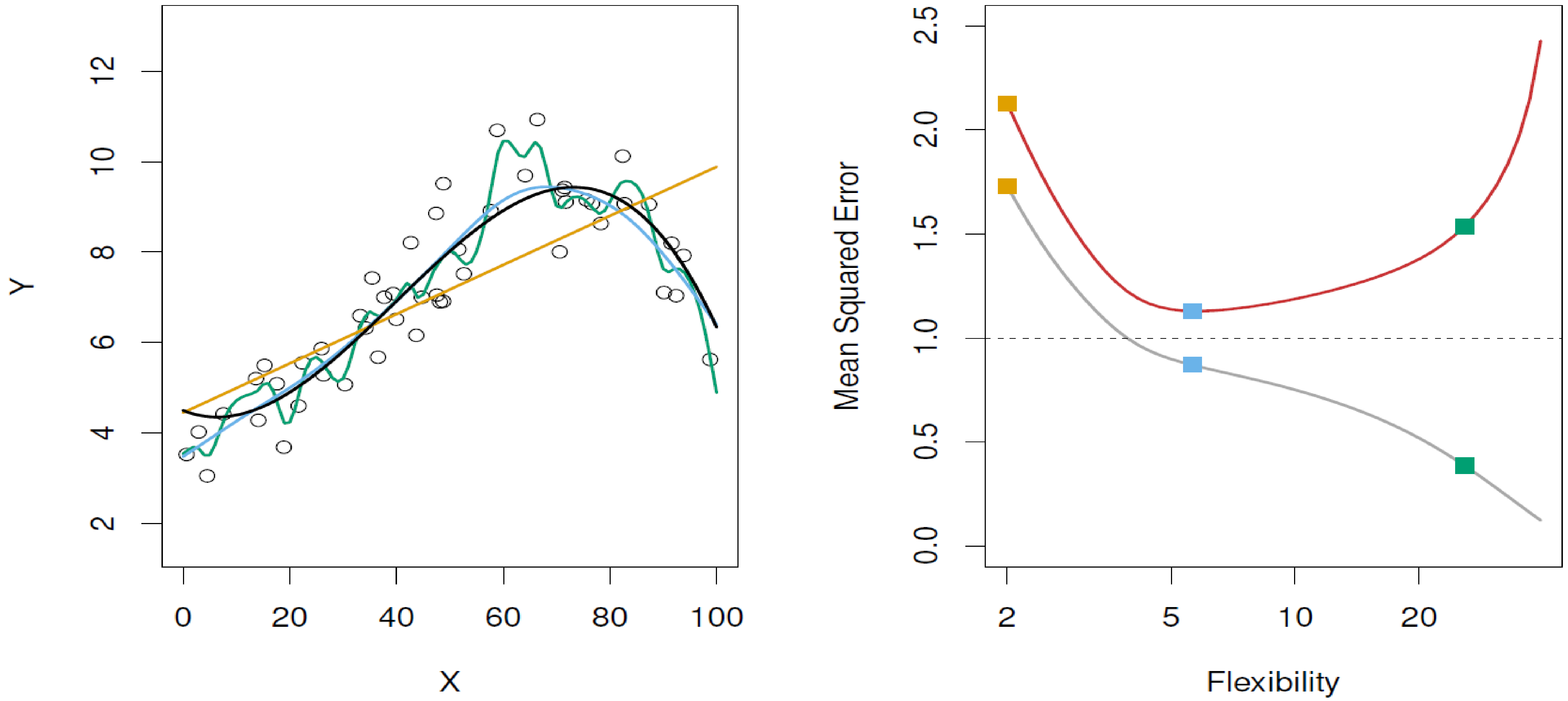
This may be biased toward more over-fit models.

- Instead we should, if possible, compute it using fresh test data $Te = \{x_i, y_i\}_1^M$

$$MSE_{Te} = \text{Ave}_{i \in Te} [y_i - \hat{f}(x_i)]^2$$

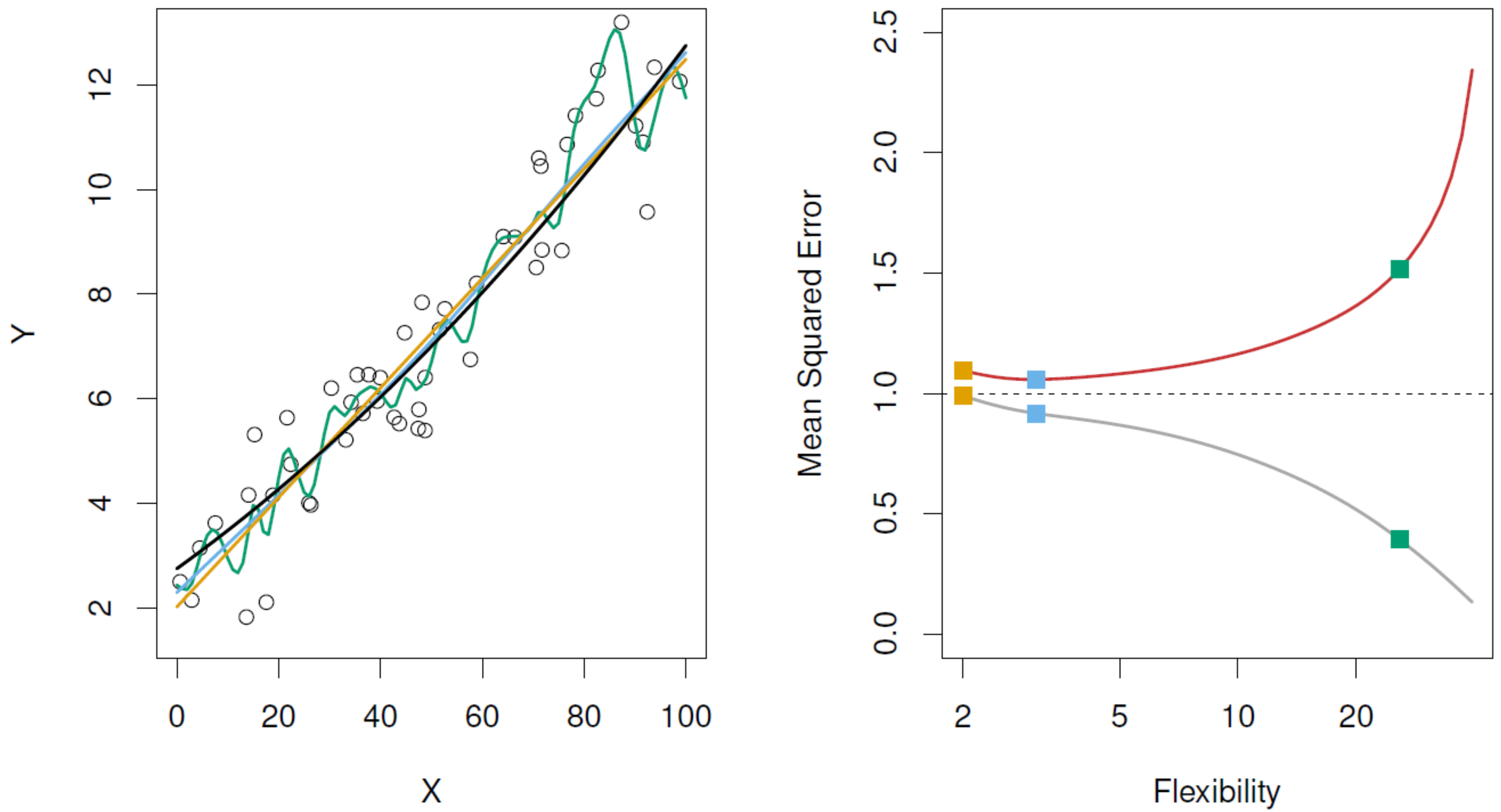
Training

Black curve is truth. Red curve on right is MSE_{Te}, grey curve is MSE_{Tr}. Orange, blue and green curves/squares correspond to fits of different flexibility.



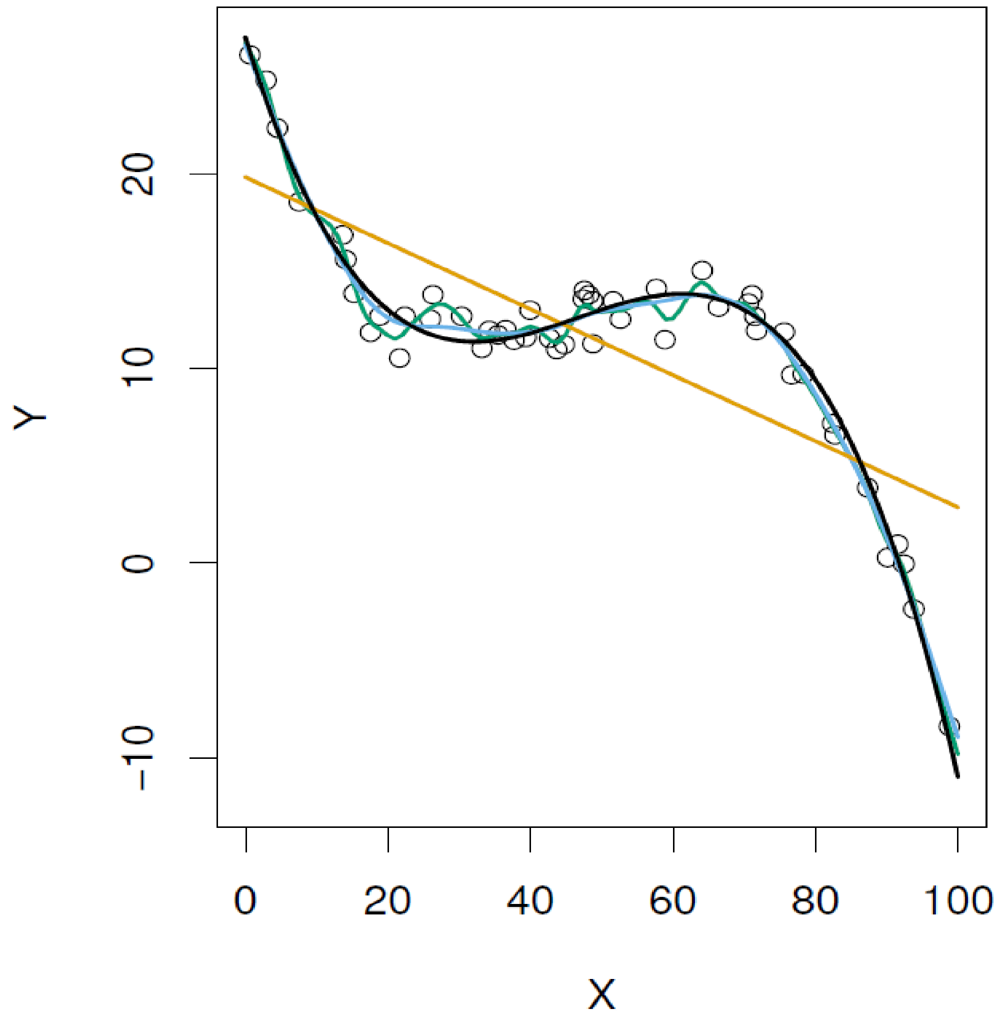
Training

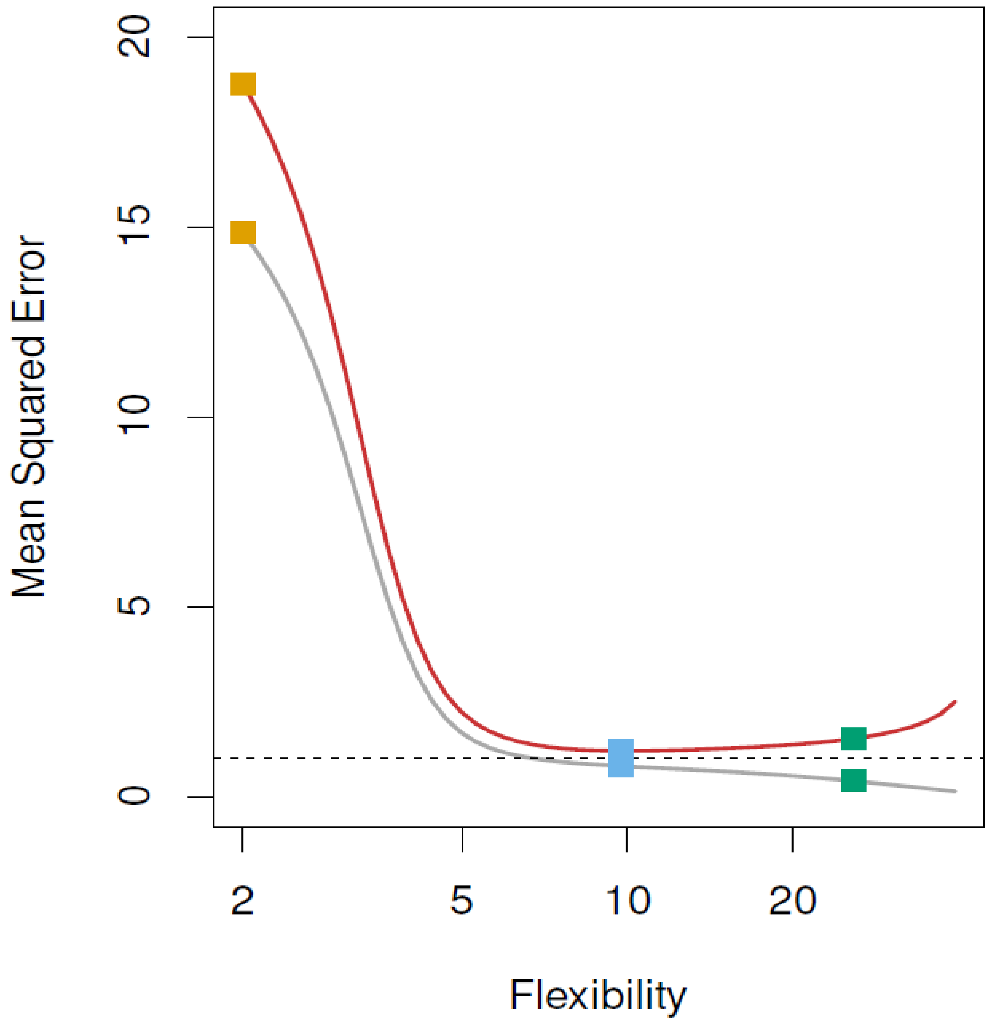
Here the truth is smoother, so the smoother fit and linear model do really well.



Fitting

Here the truth is wiggly and the noise is low, so the more flexible fits do the best.





Bias-Variance Trade-off

- Suppose we have fit a model $\hat{f}(x)$ to some training data Tr, and let (x_0, y_0) be a test observation drawn from the population. If the true model is
$$Y = f(X) + \epsilon \; (f(x) = E(Y|X = x)),$$
- Then
$$E(y_0 - \hat{f}(x_0))^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$
- The expectation averages over the variability of y_0 as well as the variability in Tr. Note that Bias $\hat{f}(x_0)$ = E$[\hat{f}(x_0)]$ - $f(x_0)$.
- Typically as the flexibility of $\hat{f}$ increases, its variance increases, and its bias decreases. So choosing the flexibility based on average test error amounts to a bias-variance trade-off.

Bias-variance trade-off for the three examples