STA 314: Statistical Methods for Machine Learning I

Lecture 2 - Measurement of the fit and linear models

Xin Bing

Department of Statistical Sciences
University of Toronto

Review

- Supervised learning is about to estimate (learn) f under the generating mechanism
$$Y = f(X) + \epsilon$$

- Obtaining the estimate $\hat{f}$ of f is divided into two categories:
  - parametric methods
  - non-parametric methods

- A more complex estimate $\hat{f}$ is able to capture more complicated relationship f (hence is more flexible) but less interpretable.

Question

- Is there a more systematic way of choosing the best $\hat{f}$ among a set of $\hat{f}$'s?
- What is a good metric for evaluating any given $\hat{f}$?

Metric of $\hat{f}$ for regression problems

We start with the regression problems where Y is quantatitive.

Recall the setup:

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

and we are given $\mathcal{D}^{train}$ consisting of n i.i.d. samples of (X, Y).

Given any $\hat{f}$, ideally, we want to evaluate $\hat{f}$ by the expected **mean squared error** (MSE)

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

where

- (X, Y) is a new random pair that is independent of $\mathcal{D}^{train}$.
- the expectation is taken w.r.t. the random pair (X, Y) as well as the randomness in $\hat{f}$.

Metric of $\hat{f}$

We cannot compute the expected MSE as we do not know the distribution of (X, Y) or that of $\mathcal{D}^{train}$.

One natural option is to use $\mathcal{D}^{train}$ to approximate the expectation by

$$MSE(\hat{f}) := \frac{1}{n}\sum_{i=1}^{N}\left(y_i - \hat{f}(x_i)\right)^2.$$

This is called the **training MSE** as it uses $\mathcal{D}^{train}$.

- However, it is **NOT** a valid metric of the fit for $\hat{f}$ because $\hat{f}$ is typically obtained by using $\mathcal{D}^{train}$ as well.
- In fact, a ubiquitous way of constructing $\hat{f}$ is to minimize $MSE(g)$ over all possible g in cretain class.
- Overfitted $\hat{f}$ usually have smaller (or even zero) $MSE(\hat{f})$.

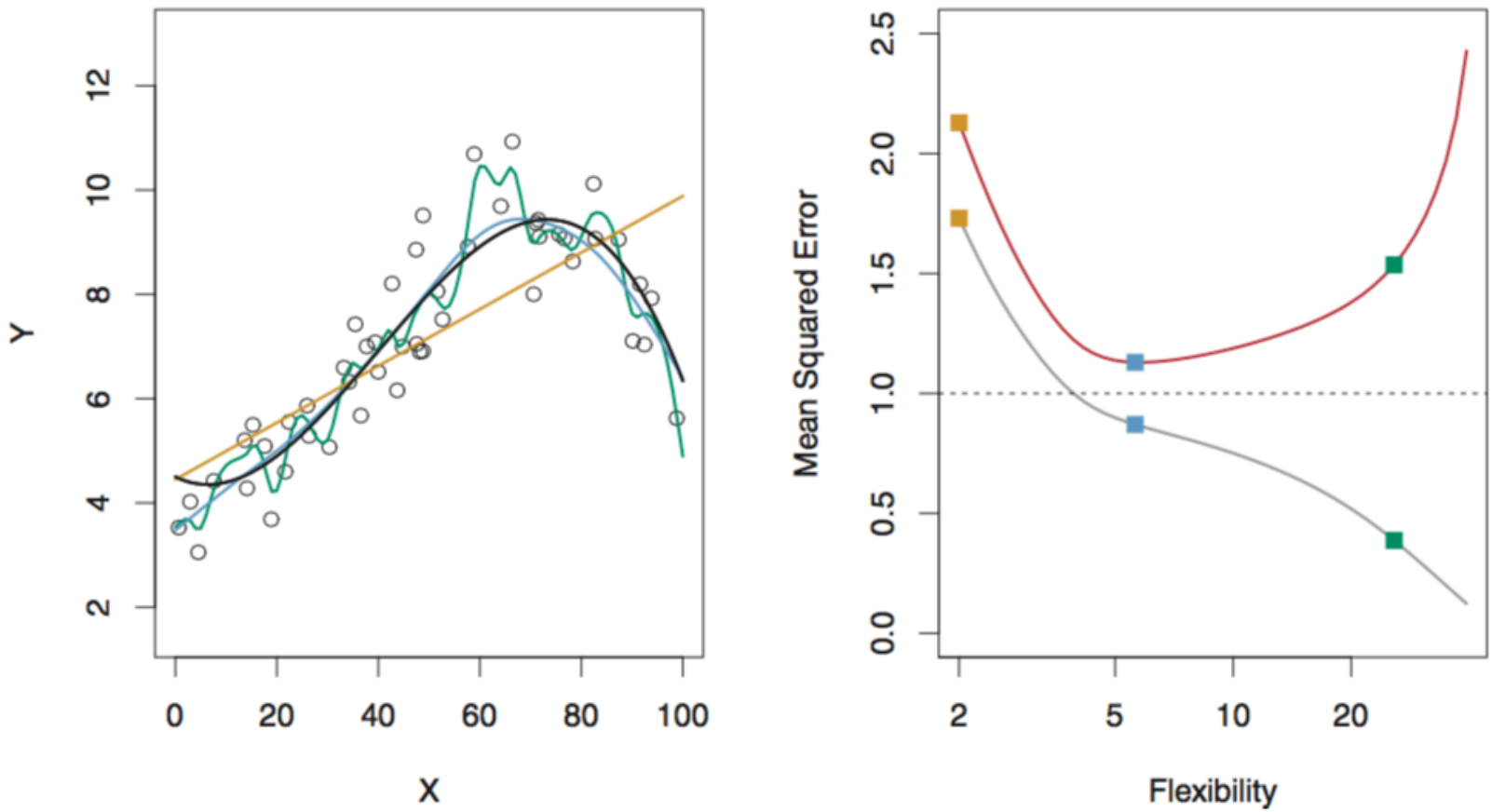
Metric of $\hat{f}$

- **Test data** refers to the data which is not used to train the statistical model (i.e., not used to compute $\hat{f}$).
- **Test MSE**. Suppose we have the test data $\mathcal{D}_{test}$ containing $\{(x_{T1}, y_{T1}), ..., (x_{Tm}, y_{Tm})\}$

$$MSE_T(\hat{f}) = \frac{1}{m}\sum_{i=1}^{m}(y_{Ti} - \hat{f}(x_{Ti}))^2.$$

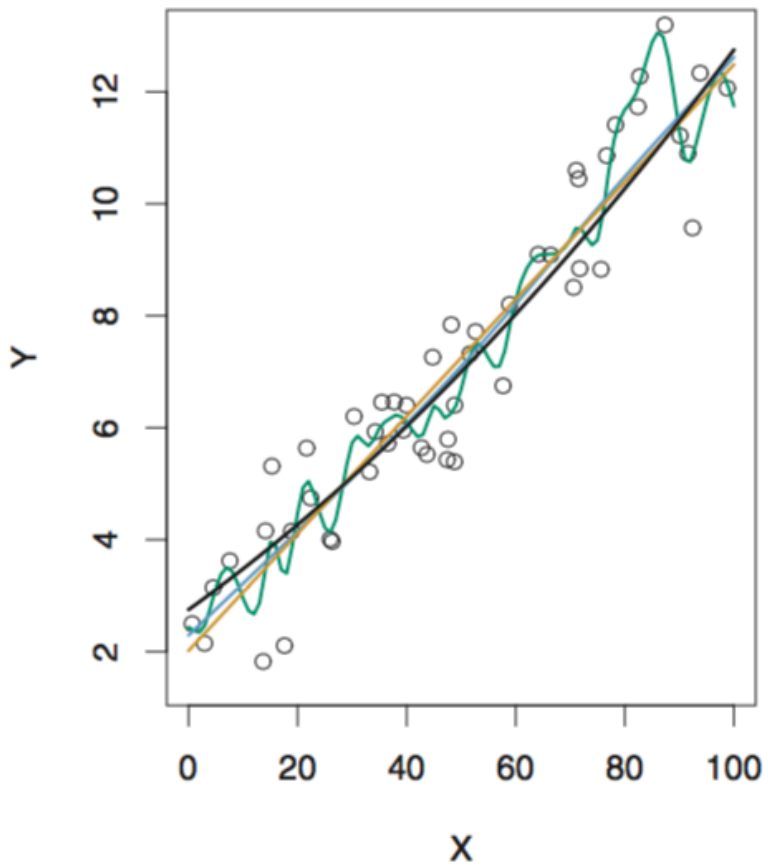
- Instead of using the training MSE, we should look at the test MSE. We'd like to select the model which yields the smallest test MSE.
- How to calculate $MSE_T(\hat{f})$?
  - If test data is available, we can directly compute $MSE_T(\hat{f})$.
  - Otherwise, we use a resampling technique called *cross-validation* (later in Lecture 3).

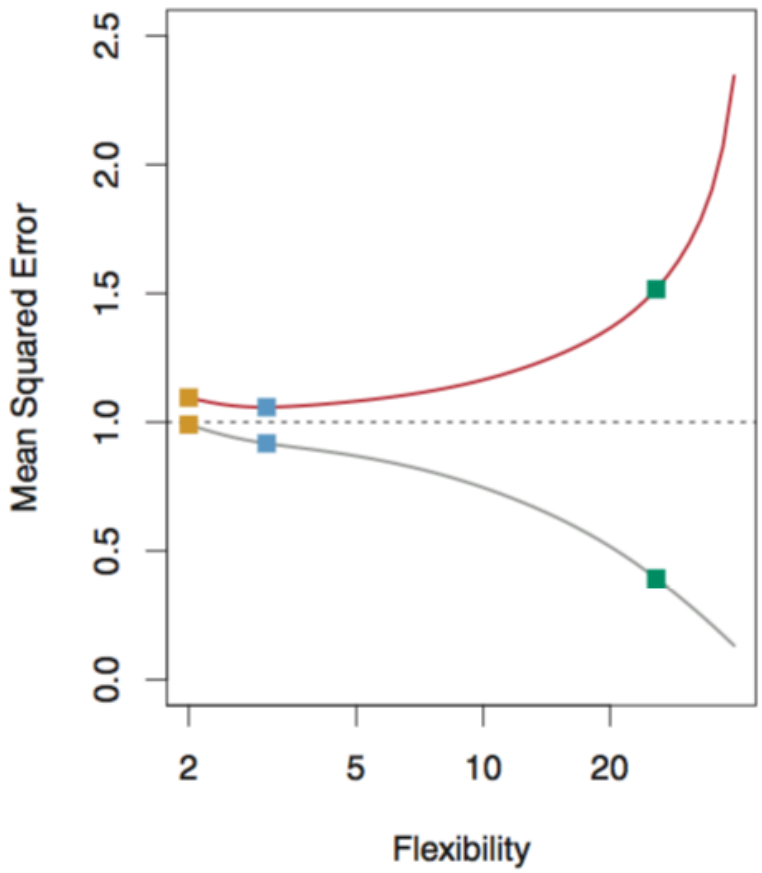
Training MSE vs Test MSE



- Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two nonparametric fits (blue and green curves).
- Right: Training MSE (grey curve), test MSE (red curve), and minimum test MSE over all possible methods (dashed line).
  Squares represent the training and test MSEs for the three fits shown in the LHS panel.

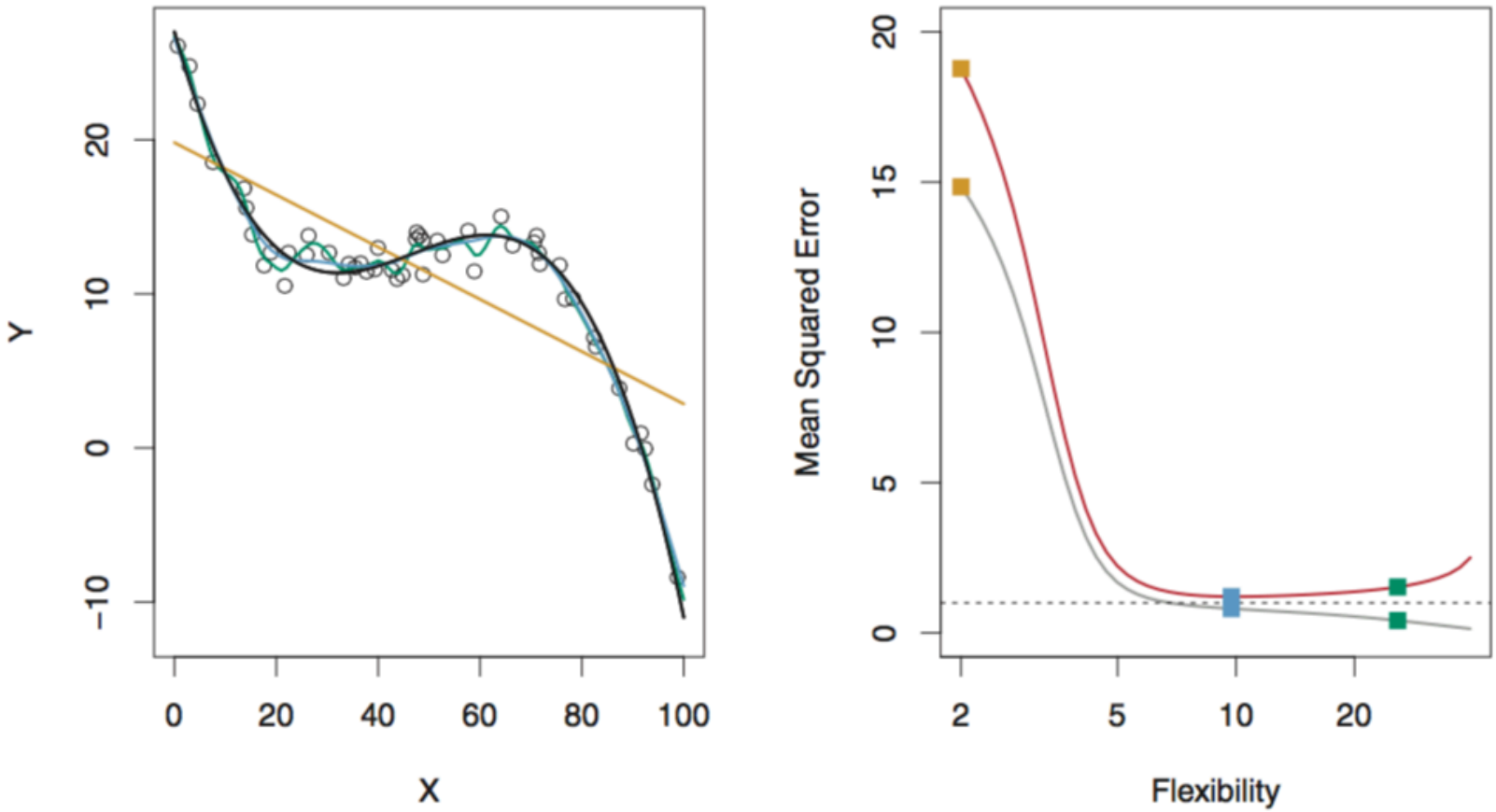
Training MSE vs Test MSE: a linear f





When f is close to linear, the linear predictor provides a very good fit to the data.

Training MSE vs Test MSE: a highly non-linear f



When f is highly non-linear, the linear predictor provides a very poor fit to the data.

- You might have noticed the tradeoff between the test MSE and the flexibility (complexity) of the fitted model $\hat{f}$.

- Question: Is there a universal rule about this trade-off?

Bias-Variance decomposition

Suppose we have an estimate $\hat{f}$ from $\mathcal{D}^{train}$. Let (X, Y) be a new random pair (independent from $\mathcal{D}^{train}$).

Recall that $Y = f(X) + \epsilon$, with $\mathbb{E}[\epsilon|X] = 0$. Then the conditional **expected MSE** at any $X = x$ is

$$\begin{aligned}
&\mathbb{E}\left[\left(Y - \hat{f}(X)\right)^2 | X = x\right] \\
&= \underbrace{\text{Var}(\hat{f}(x))}_{Variance} + \left(\underbrace{\mathbb{E}[\hat{f}(x)] - f(x)}_{Bias}\right)^2 + \underbrace{\text{Var}(\epsilon)}_{Irreducible\ \ error} \\
&\geq \text{Var}(\epsilon)
\end{aligned}$$

- The expectation is over the variability of $Y|X = x$ as well as $\mathcal{D}^{train}$.
- The expected MSE $\geq$ the Irreducible error.
- An ideal $\hat{f}$ should minimize the expected MSE.

What is the Bias-Variance Trade-off?

Variance: how much $\hat{f}$ would change if we estimated it using a different training data set.

Bias: refers to the error that is introduced by parametrizing f.

E.g., the real relationship between response and predictors is nonlinear

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3,$$

but we fit a linear model

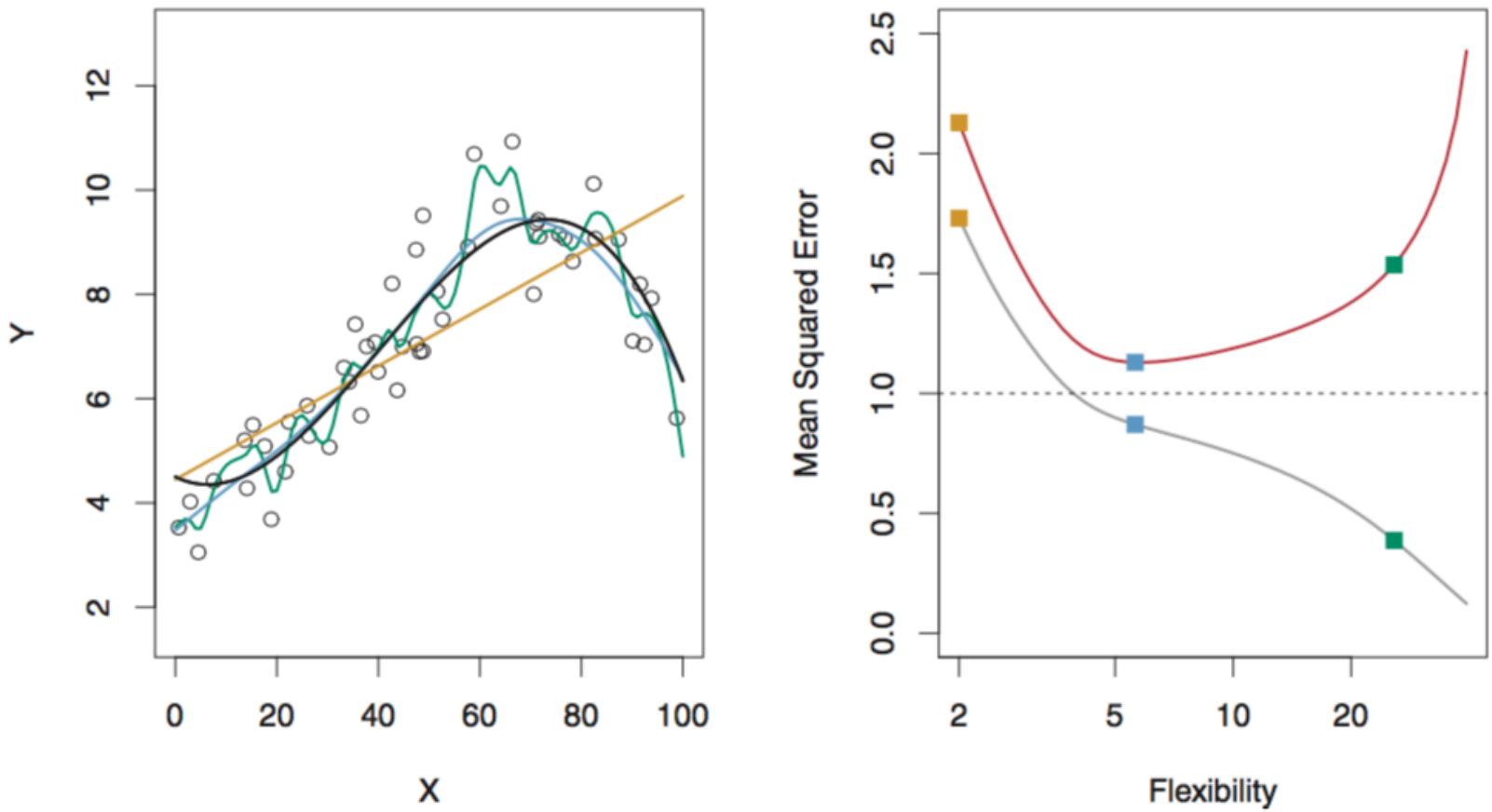
$$\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X.$$

This causes a bias in $\mathbb{E}[\hat{f}(x)] - f(x)$ at $X = x$.

What is the Bias-Variance Trade-off?

- As the complexity (a.k.a. flexibility) of $\hat{f}$ increases (e.g., linear method $\rightarrow$ non-parametric methods), the variance of $\hat{f}$ typically increases whereas its bias decreases.
- On the other hand, if $\hat{f}$ is less complex (e.g., linear model), the variance of $\hat{f}$ is usually small, and its bias is large.
- So choosing the complexity of $\hat{f}$ based on the expected MSE has a bias-variance trade-off.
- When two $\hat{f}_1$ and $\hat{f}_2$ have similar expected MSEs, we usually prefer the more parsimonious (less complex) one.

Example

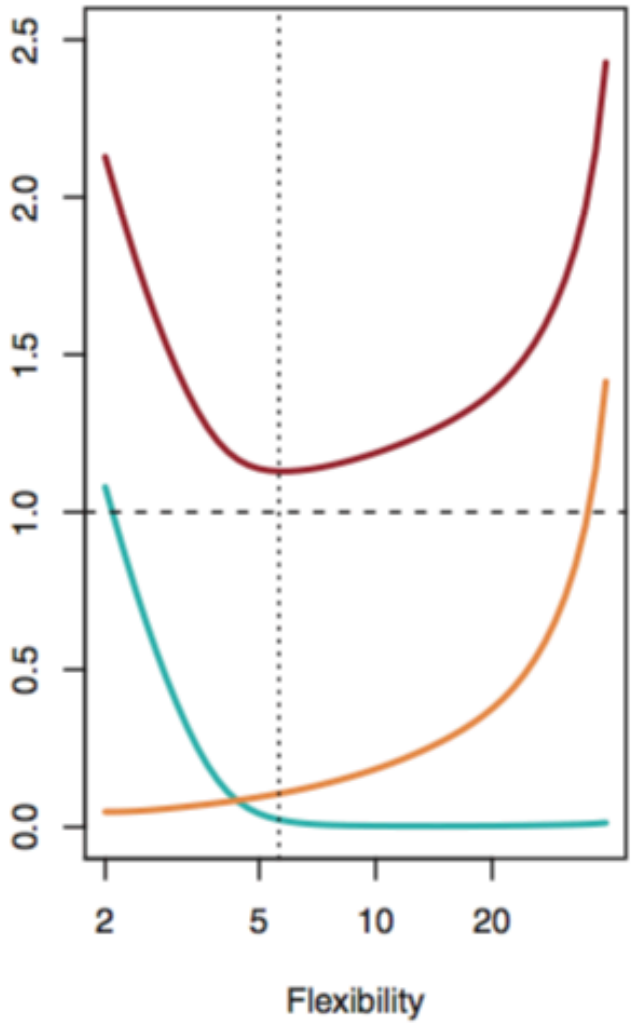


- Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two nonparametric fits (blue and green curves).
- Right: Training MSE (grey curve), test MSE (red curve), and minimum test MSE over all possible methods (dashed line).
  Squares represent the training and test MSEs for the three fits shown in the LHS panel.

Example

- red curve: the test MSE.
- blue curve: $(\mathbb{E}[\hat{f}(x)] - f(x))^2$
- orange curve: $\mathrm{Var}(\hat{f}(x))$

- dashed horizontal line: $\mathrm{Var}(\epsilon)$
- dotted vertical line: the best flexibility corresponding to the smallest test MSE.



- There are alternative metrics for measuring $\hat{f}$, such as the Sum of Absolute Difference (SAM):

$$\mathbb{E}[|Y - \hat{f}(X)|], \qquad \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{f}(x_i)\right|.$$

- Both MSE and SAM are only appropriate for quantatitive Y!

- What about categorical or ordinal Y?
  - Spam email detection: $Y = 0$ for non-spam, $Y = 1$ for spam
  - Hand-written digit recoginition: $Y \in \{0, 1, ..., 9\}$

Metric of $\hat{f}$ for classification

When Y is categorical or ordinal, the **expected error rate** is defined as

$$\mathbb{E}\Big[1\{Y \neq \hat{f}(X)\}\Big].^{1}$$

Analogously, the **training error rate** is

$$\frac{1}{n}\sum_{i=1}^{n} 1\left\{y_i \neq \hat{f}(x_i)\right\}$$

and the **test error rate** is

$$\frac{1}{m}\sum_{i=1}^{m} 1\left\{y_{Ti} \neq \hat{f}(x_{Ti})\right\}.$$

Of course, there also exists other metrics that can be used when Y is categorical or ordinal.

[1] $1\{\}$ is the indicator function. $1\{\text{A}\} = 1$ if A is true and $1\{\text{A}\} = 0$ otherwise.

Summary on the metrics of the fit

- In regression problems, we have the expected MSE, the training MSE and the test MSE.
- In classification problems, we have the expected error rate, the training error rate and the test error rate.
- The best model yields the smallest expected MSE (error rate).
- Bias and variance trade-off exists in both scenarios.
  - A more complex / flexible $\hat{f}$ has smaller bias but larger variance
- Among models that have similar expected MSE (error rate), we always prefer the more parsimonious one.

- We have learned the bias-variance-tradeoff phenomenon for different $\hat{f}$'s.

- In practice, how should we compute the expected MSE to select the best $\hat{f}$?
  (We will come back to this later in Lecture 3).

Questions?

Linear regression

Let $Y \in \mathbb{R}$ be the outcome and $X \in \mathbb{R}^p$ be the (random) vector of p features.

The linear model assumes

$$\begin{aligned} Y &= f(x) + \epsilon \\ &= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \qquad \text{(linearity)} \end{aligned}$$

where:

- $\beta_0, \beta_1, \cdots, \beta_p$ are unknown constants.
  - β_0 is called the **intercept**
  - β_j, for $1 \le j \le p$, are the **coefficients** or **parameters** of the p features
- ϵ is the error term satisfying $\mathbb{E}[\epsilon|X] = 0$.

Linear predictor under the linear regression model

Given some estimates $\hat{\beta}_j$ of β_j for $0 \leq j \leq p$, we predict the response at $X = x$ by the linear predictor

$$\hat{y} := \hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p.$$

Question: how to choose $\hat{\beta}_0, \ldots, \hat{\beta}_p$?

Ordinary Least Squares approach (OLS)

Recall that we want to find a function g by

$$\min_g \mathbb{E}\left[(Y - g(X))^2\right].$$

Under linear model, it suffices to find $\alpha_0, \ldots, \alpha_p$ by

$$\min_{\alpha_0,\ldots,\alpha_p} \mathbb{E}\left[\left(Y - \alpha_0 - \alpha_1 X_1 - \cdots - \alpha_p X_p\right)^2\right]$$

In the model fitting step, we use $\mathcal{D}^{train}$ to approximate the above expectation (w.r.t. X and Y).

Specifically, given the training data $\mathcal{D}^{train}$: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, we choose $\hat{\beta}_0, \cdots, \hat{\beta}_p$ by minimizing the training MSE.

$$(\hat{\beta}_0, \cdots, \hat{\beta}_p) = \underset{\alpha_0,\cdots,\alpha_p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \alpha_0 - \alpha_1 x_{i1} - \cdots - \alpha_p x_{ip}\right)^2.$$

Ordinary Least Squares approach (OLS)

Using the matrix notation,

- $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \cdots, \hat{\beta}_p]^\top \in \mathbb{R}^{p+1}$, $\boldsymbol{\beta} = [\beta_0, \cdots, \beta_p]^\top \in \mathbb{R}^{p+1}$, $\boldsymbol{\alpha} = [\alpha_0, \cdots, \alpha_p]^\top \in \mathbb{R}^{p+1}$,
- $\mathbf{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$, $\mathbf{X} = [1, x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times (p+1)}$,

the OLS estimator of $\boldsymbol{\beta}$ is defined as

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\alpha} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i^\top \boldsymbol{\alpha} \right)^2 \\ &= \underset{\boldsymbol{\alpha} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \ \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|_2^2. \end{aligned}$$

Comments

- The idea of estimating f by minimizing the training MSE can be applied to (almost) all supervised problems. Specifically,

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \tag{1}$$

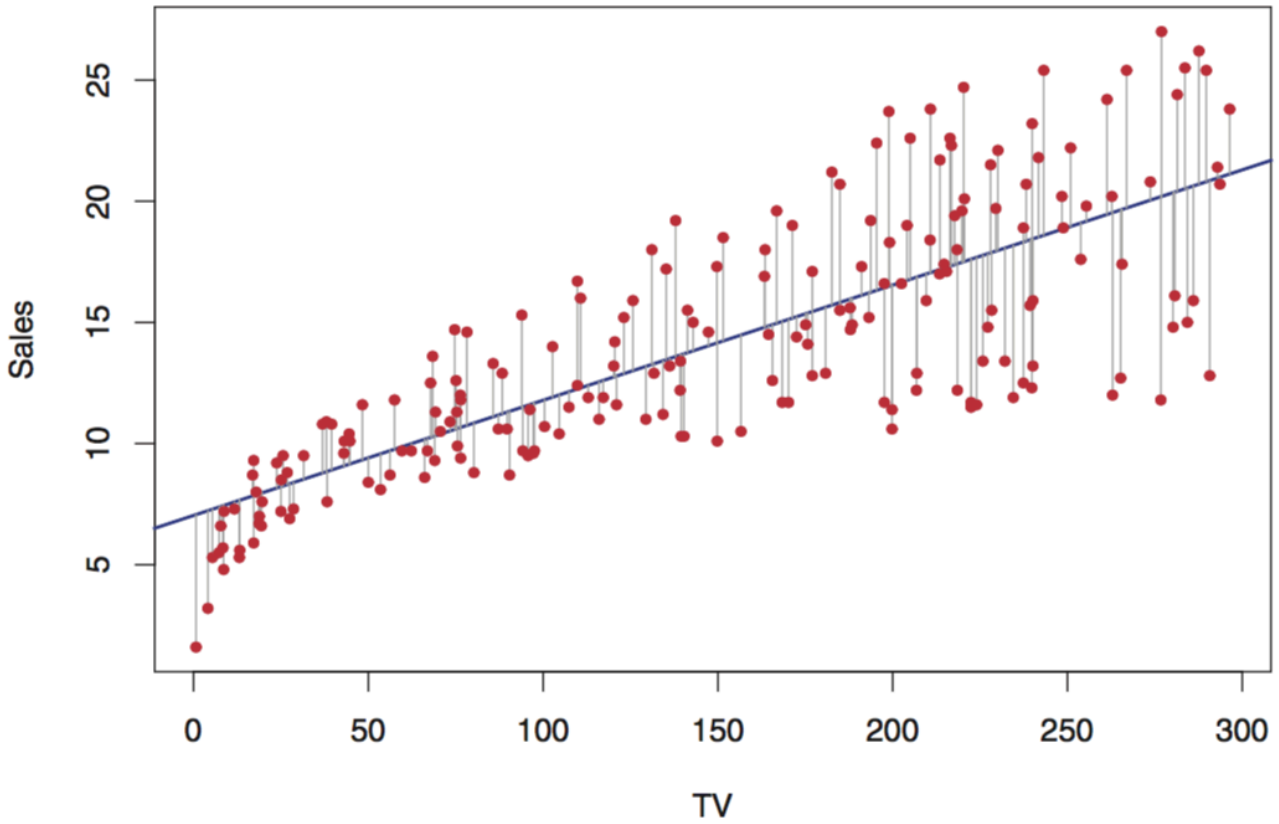
  where $L(\cdot, \cdot)$ is a loss function and $\mathcal{F}$ is a class of choices of f.
- The OLS approach corresponds to $L(a, b) = (a - b)^2$ and $\mathcal{F}(x) = \{\beta : f(x) = x^\top \beta\}$.
- In general, the difficulty of solving (1) varies across problems. But, the OLS approach has a unique, closed-form solution:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

  whenever $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$ has full column rank.

An example for $p = 1$: Advertising Data

Y : **Sales**, X : **TV** budget, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 7.03 + 0.05 x_i$.



- Each grey segment represents an error. The fitted model compromises by averaging the squared errors.
- A linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

Some important considerations

- Estimation of β:
  - How close is the point estimation $\hat{\beta}$ to β?
  - (Inference) Can we provide confidence interval / conduct hypothesis testing of β?
- Prediction of Y at $X = x$:
  - How accurate is the point prediction $\hat{y} = x^\top \hat{\beta}$?
  - (Inference) Can we further provide confidence interval of Y?
- Variable (Model) selection:
  - Do all the predictors help to explain Y, or is there only a subset of the predictors useful?
  - Later in Lecture 3

Property of $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$

Take the *design matrix* $\mathbf{X}$ to be deterministic with full column rank. Assume $\epsilon_1, \ldots, \epsilon_n$ are independent with $\mathbb{E}[\epsilon_i] = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$.

- Unbiasedness: $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$
- The covariance matrix of $\hat{\boldsymbol{\beta}}$ is:

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

- The above two properties imply the ℓ_2 estimation error

$$\mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \right] = \sigma^2 \,\text{Tr}\left[(\mathbf{X}^\top \mathbf{X})^{-1} \right]$$

When $\mathbf{X}^\top \mathbf{X} = n\, \mathbf{I}_{p+1}$,

$$\mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \right] = \frac{\sigma^2 (p+1)}{n}.$$

The MSE of estimating β increases as p gets larger.

Inference on β

- The *unknown* variance σ^2 may be estimated by

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\left(y_i - x_i^\top \hat{\beta}\right)^2.$$

- A 95% confidence interval of β_j has the form of

$$\left[\hat{\beta}_j - 1.96 \cdot SE(\hat{\beta}_j),\quad \hat{\beta}_j + 1.96 \cdot SE(\hat{\beta}_j)\right],$$

where

$$SE(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}$$

- Hypothesis testing

$H_0 : \beta_j = 0$ (There is no linear relationship between Y and X_j)

vs

$H_1 : \beta_j \neq 0$ (There is linear relationship between Y and X_j)

Inference on β

We base on the t-statistic

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

- This has a t-distribution with $n - p - 1$ degrees of freedom, when $\beta_j = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the ***p*-value**.
- In most applications, we reject the null hypothesis if the p-value ≤ 0.05.
- Can be generalized to

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$

via F-statistics. (c.f. pp 75-78 of the textbook.)

Results for Advertising Data

Y : **Sales**, X : **TV** budget, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 7.0325 + 0.0475 x_i$.

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

- The p-value for **TV** is smaller than 0.05, so that we reject the null hypothesis $\beta_1 = 0$.
- This indicates that **TV** is significant for predicting **Sales**.

Property of the prediction $\hat{y} = x^\top \hat{\beta}$ at $X = x$

The property of $\hat{\beta}$ can be used to analyze the prediction $\hat{y}$ of y.

- Expectation

$$\mathbb{E}[\hat{y} \mid X = x] = x^\top \mathbb{E}[\hat{\beta}] = x^\top \beta$$

- Variance

$$\text{Var}[\hat{y} \mid X = x] = x^\top \text{Cov}(\hat{\beta}) x = \sigma^2 x^\top (\mathbf{X}^\top \mathbf{X})^{-1} x.$$

- MSE

$$\mathbb{E}[(y - \hat{y})^2 \mid X = x] = \sigma^2 + \sigma^2 x^\top (\mathbf{X}^\top \mathbf{X})^{-1} x.$$