Introduction to Multiple Linear Regression

DS 301

Iowa State University

Today's Agenda

- HW 1 posted later today. Due Wednesday Feb. 2 at 11:59 pm on Canvas.
- Introduction to Multiple Linear Regression
- Optional reading: Chapter 3 (ISLR)

Recap

Bias-variance tradeoff

estimate using test MSE $= \frac{1}{m}\sum_{i=1}^{m}(y_i' - \hat{f}(x_i'))^2$

↓

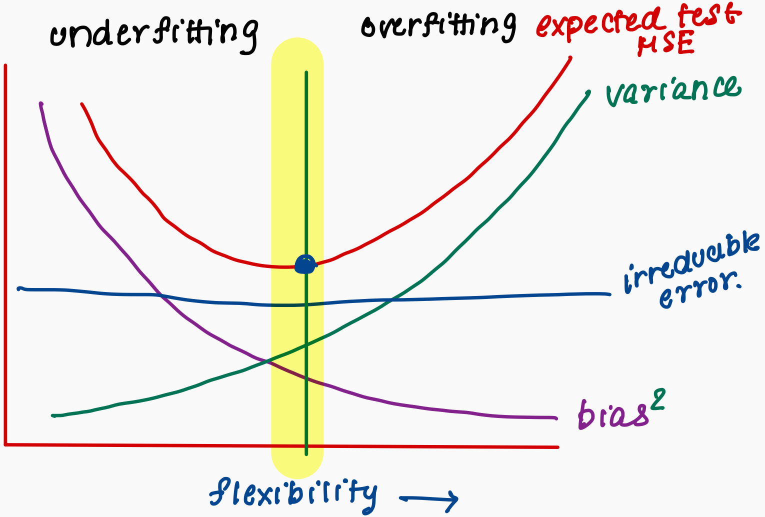
The expected test MSE depends on two competing quantities:

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

assumes we know true $f(x)$.

- **General rule**: More flexible methods (complex) ⇒ $\underbrace{\text{higher variance}}_{\text{overfitting}}$ and lower bias.

Bias-variance tradeoff



Implications of the bias-variance tradeoff

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

- The expected test MSE is never smaller than the irreducible error.
- Easy to find a zero variance estimate with high bias.
- Easy to find a low bias estimate with high variance.
- In practice, the best expected test MSE is achieved by allowing some bias to decrease variance and vice-versa: this is the bias-variance trade-off.
- **General rule**: More flexible methods ⇒ higher variance and lower bias.

Main points

- In supervised learning, the challenge lies in finding a method for which both the variance and the squared bias are low.

Main points

- In supervised learning, the challenge lies in finding a method for which both the variance and the squared bias are low.
- In a real life situation, it is generally not possible to explicitly compute the expected test MSE, bias, or variance for a method (think about why.).

Main points

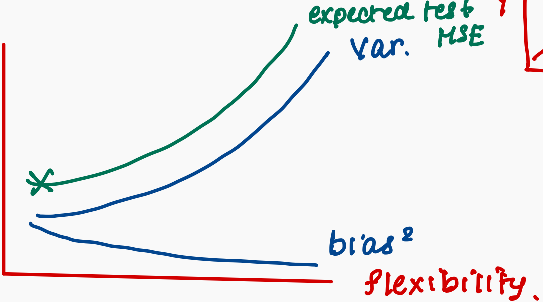
- In supervised learning, the challenge lies in finding a method for which both the variance and the squared bias are low.
- In a real life situation, it is generally not possible to explicitly compute the expected test MSE, bias, or variance for a method (think about why.).
- Despite this, we should always keep the bias-variance trade-off in mind. It gives us insight into how our models behave and will be a recurring theme throughout the course.

Main points

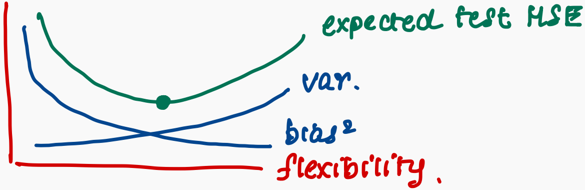
- In supervised learning, the challenge lies in finding a method for which both the variance and the squared bias are low.
- In a real life situation, it is generally not possible to explicitly compute the expected test MSE, bias, or variance for a method (think about why.).
- Despite this, we should always keep the bias-variance trade-off in mind. It gives us insight into how our models behave and will be a recurring theme throughout the course.

Tradeoff depends on the true model

- As we move through the course, we will explore methods that are extremely flexible and can essentially eliminate bias. However, this **does not** guarantee that they will outperform a much simpler method such as linear regression.
- Consider an extreme example: suppose the true f is linear.
  - In this situation linear regression will have no bias, making it very hard for a more flexible method to compete.



Tradeoff depends on the true model (f)



- On the other end of the spectrum, suppose the true f is highly non-linear and we have an ample number of training observations, then we may do better using a highly flexible approach.

Multiple Linear Regression

Multiple Linear Regression

Motivation:

1. Simple, can provide an interpretable description of the relationship between Y and X.
   - Inference is well-studied in this setting.
2. Widely used.
3. In terms of prediction, can often outperform more complicated models.
4. The fundamentals covered here are the building blocks for more complicated models.

Recall regression setup:

$$Y_i = f(X_i) + \epsilon_i, \quad i = 1 \ldots, n.$$

If we are willing to make the assumption that the relationship between X and Y is *approximately* linear, then

$$f(X_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}.$$

parameters (unknown)

Population regression line:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i, \quad i = 1 \ldots, n.$$

How to obtain estimate for $\hat{\beta}$? $\hat{B_0}, \hat{B_1}, \hat{B_2}, \ldots, \hat{B_p}$

Let's start with the simple linear regression case (we only have 1 predictor X_1). $y_i = B_0 + B_1 X_{1i} + \varepsilon_i \Longrightarrow \hat{y_i} = \hat{B_0} + \hat{B_1} X_{1i}$

- Our goal is to find coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model fits the data well ($y_i \approx \hat{\beta}_0 + \hat{\beta}_1 X_{1i}$).
- In other words, we want to the line to be as close as possible to the data points.
- How do we define closeness here? The most common approach involves *least squares criterion.* > find $\hat{B_0}, \hat{B_1}$ minimizes RSS

(residual sum of squares)

$$RSS = \sum_{i=1}^{n} \Big(\underbrace{y_i - (\hat{B_0} + \hat{B_1} X_{1i})} \Big)^2$$

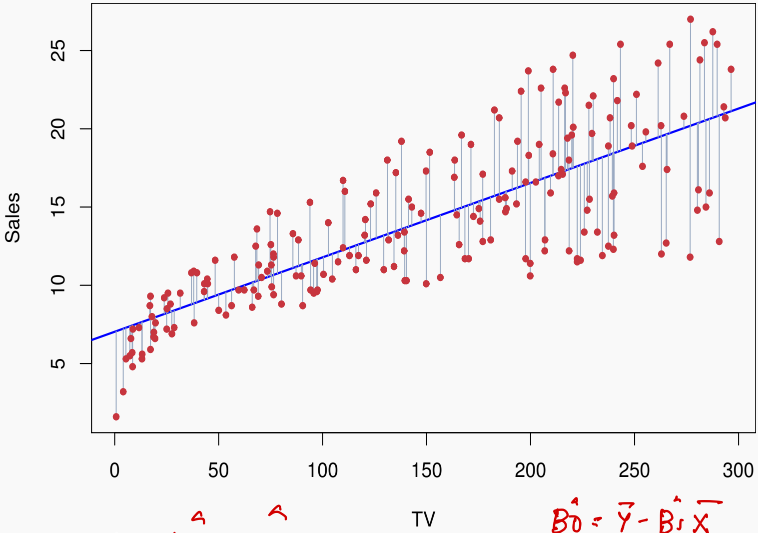
$$\sum_{i=1}^{n} (\underbrace{y_i - \hat{y_i}}_{e_i})^2$$

e_i: residual

$y_i - \hat{y_i}$

$y_i - (\hat{B_0} + \hat{B_1} X_i)$

Least squares estimation $\rightarrow$ minimizing RSS



$\hat{B}_0 , \hat{B}_1$

$$\hat{B}_0 = \bar{Y} - \hat{B}_1 \bar{X}$$

$$\hat{B}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Extending to multiple linear regression

$X_1, X_2, X_3, \ldots, X_p$.

$$y_i = \underline{\beta_0} + \underline{\beta_1} x_{i1} + \underline{\beta_2} x_{i2} + \cdots + \underline{\beta_p} x_{ip} + \varepsilon_i$$

find $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ such that

$$RSS = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}) \right)^2$$

is minimized.

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (\text{fyi}).$$

Implementation in `R`

See R script `IntroMLR.R`

Questions?