IEOR165 Discussion
Week 4

Sheng Liu

University of California, Berkeley

Feb 12, 2016

Outline

Understand the bias and variance

- Bias: the difference between the expected value of the estimator and the true value

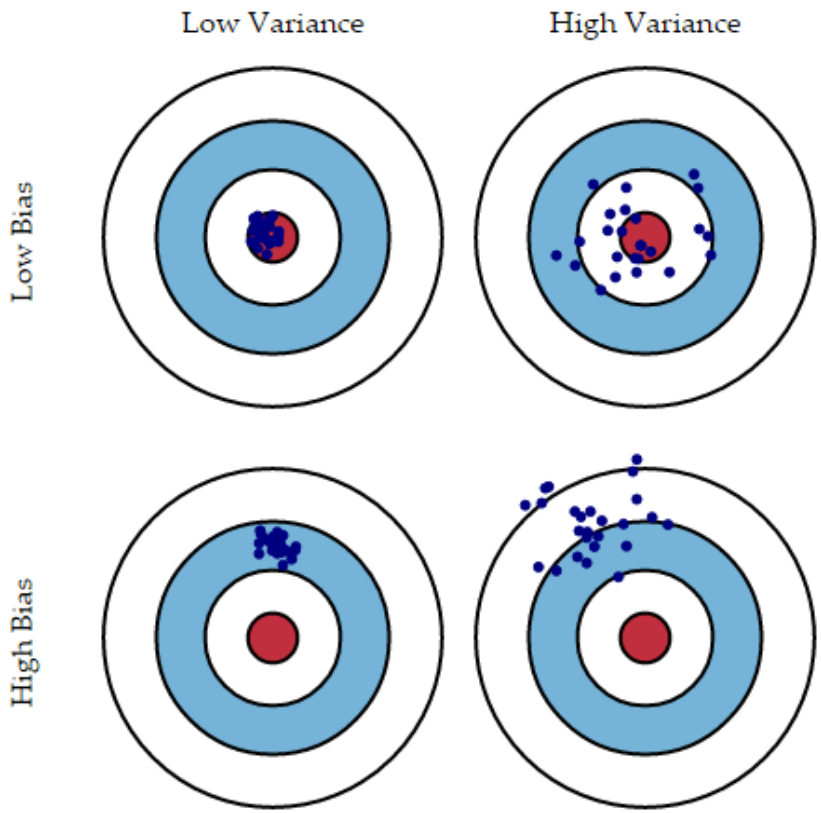
$$E(\hat{\beta}) - \beta$$

- Variance: the variance of the estimator $Var(\hat{\beta})$.
- Keep in mind: the estimator is also a random variable.

In an experiment, we want to estimate a parameter β. Suppose the true value is $\beta = 1$. We adopt two different estimators and repeat the experiment 5 times (each time we derive a data set and calculate the estimates):

- Estimator A: $\hat{\beta}_A = \{0.50, 0.48, 0.45, 0.52, 0.54\}$
- Estimator B: $\hat{\beta}_B = \{1.42, 0.63, 0.73, 1.21, 1.09\}$

Understand the bias and variance



Reference:
`http://scott.fortmann-roe.com/docs/BiasVariance.html`

Squared Loss: Mean Squared Error

Berkeley UNIVERSITY OF CALIFORNIA

How to measure the quality of an estimator?

- A popular way: Mean Squared Error (MSE)

$$\text{MSE} = \mathbb{E}[(\hat{\beta} - \beta)^2]$$

- After some algebra, we have the following important formula:

$$\begin{aligned}\text{MSE} &= (\mathbb{E}(\hat{\beta}) - \beta)^2 + \mathbb{E}[(\hat{\beta} - \mathbb{E}(\hat{\beta}))^2] \\ &= (\text{bias}(\hat{\beta}))^2 + Var(\hat{\beta})\end{aligned}$$

- Minimum MSE is desired
- There is a tradeoff between bias and variance

Evaluate OLS

Now lets look at the OLS estimates for linear regression models[1]:

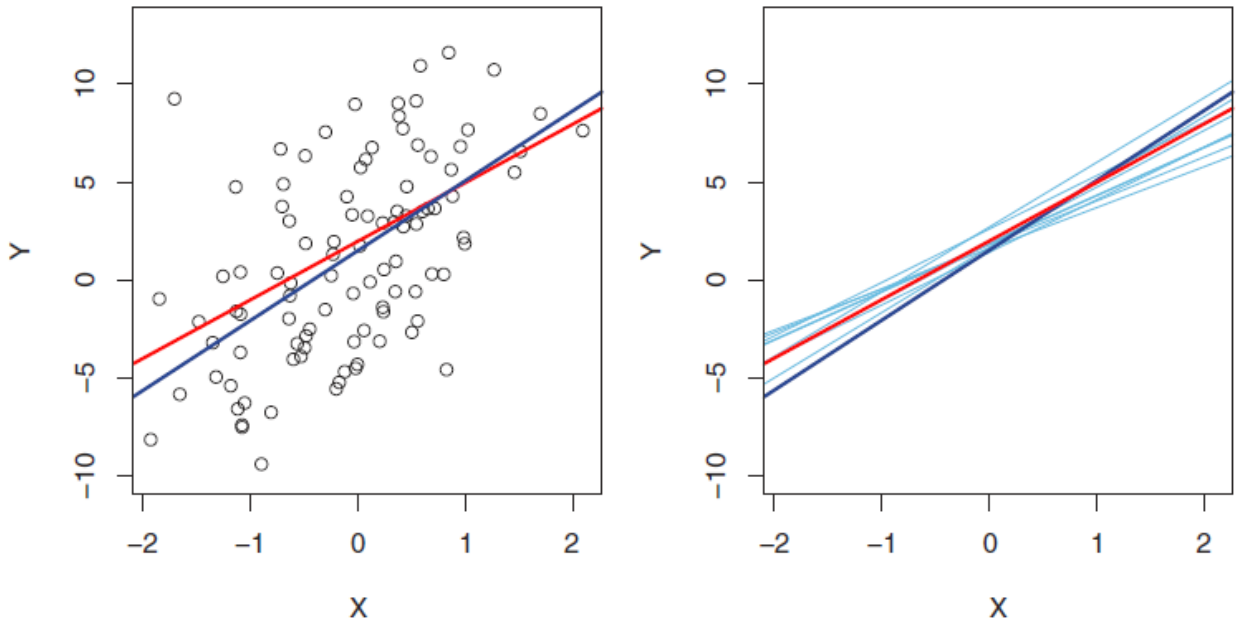


FIGURE 3.3. *A simulated data set.* Left: *The red line represents the true relationship,* $f(X) = 2 + 3X$, *which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for* $f(X)$ *based on the observed data, shown in black.* Right: *The population regression line is again shown in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.*

[1] Gareth, et al. 2013

OLS: MSE

For our OLS estimate of linear model $y = m \cdot x + b + \epsilon$, we have

$$\hat{m} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{\sum_i x_i y_i - \bar{x}\sum_i y_i}{\sum_i x_i^2 - n(\bar{x})^2}$$

$$\hat{b} = \bar{y} - \hat{m}\bar{x} = \frac{1}{n}\sum_i y_i - \hat{m}\bar{x}$$

1, Show that they are unbiased.
2, Calculate the MSE of $\hat{m}$.

Ridge Regression

Ridge regression:

$$(\hat{m}, \hat{b}) = \arg\min \sum_{i=1}^{n} (y_i - m \cdot x_i - b)^2 + \lambda(m^2 + b^2)$$

It introduces bias but reduces the variance, to see this, consider the special case without b, i.e. drop the intercept:

- For OLS,

$$\hat{m}_o = \arg\min \sum_{i=1}^{n} (y_i - m \cdot x_i)^2 \quad \Rightarrow \hat{m}_o = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

- For ridge regression,

$$\hat{m}_r = \arg\min \sum_{i=1}^{n} (y_i - m \cdot x_i)^2 + \lambda m^2 \quad \Rightarrow \hat{m}_r = \frac{\sum_i x_i y_i}{\sum_i x_i^2 + \lambda}$$

Ridge Regression

OLS estimate is unbiased:

$$\mathbb{E}(\hat{m}_o) = m$$

Ridge estimate is biased:

$$E(\hat{m}_r) = \frac{\sum_i x_i^2}{\sum_i x_i^2 + \lambda} m = (1 - \frac{\lambda}{\sum_i x_i^2 + \lambda}) m$$

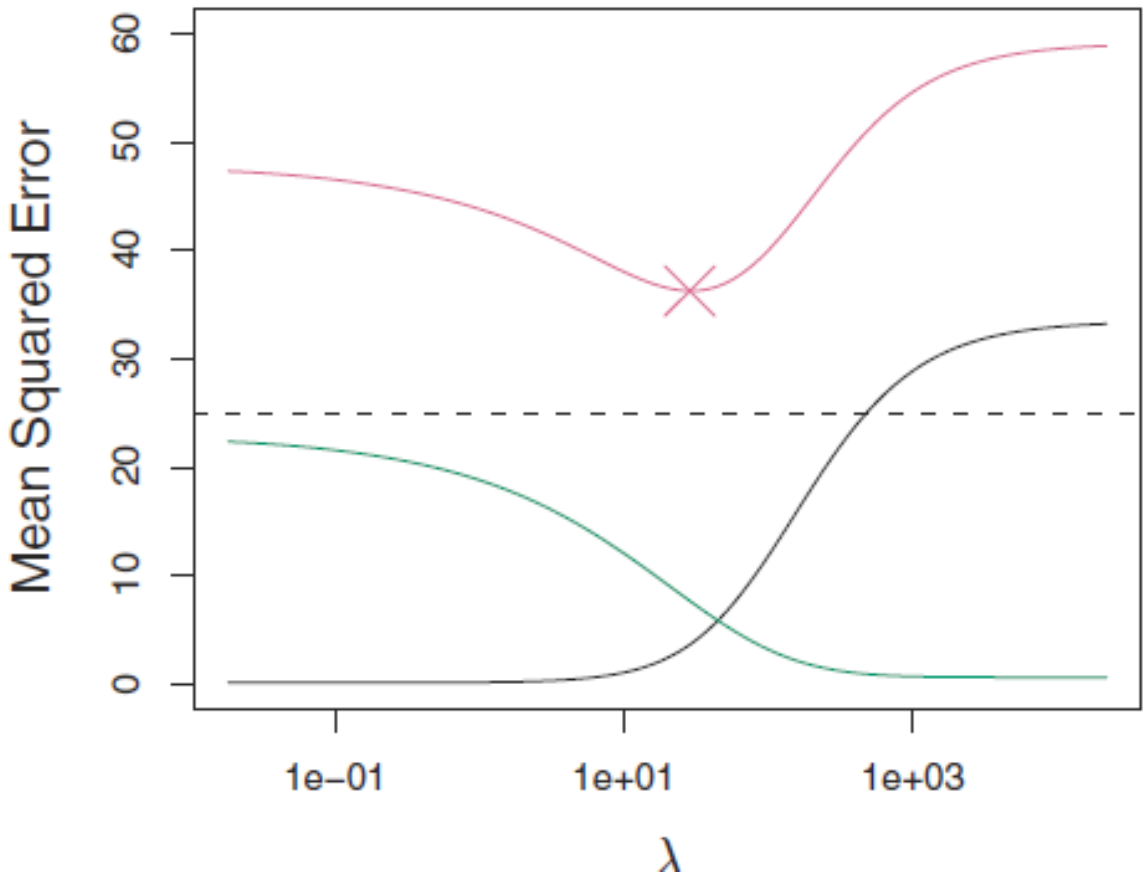
Compare the variance:

$$Var(\hat{m}_o) > Var(\hat{m}_r) \xrightarrow{\lambda \to \infty} 0$$

How about MSE?

Ridge Regression

Squared bias, variance and MSE [2]:



[2] Gareth, et al. 2013

Maximum A Posteriori

- Frequentist: underlying parameters are constant
  - θ is unknown but fixed
  - A random sample is drawn from a population with this θ
  - From the sample we can get knowledge about θ.
- Bayesian: underlying parameters are unknown and follows a distribution
  - θ follows a probability distribution $g(\theta)$, which is subjective (prior)
  - After drawing a sample, we can update the our knowledge about the distribution of θ, the updated distribution is the posterior distribution

$$\begin{aligned}\hat{\theta} &= \arg\max\ f(\theta|X) \\ &= \arg\max\ f(X|\theta)g(\theta)\end{aligned}$$

Maximum A Posteriori: Example

Assume a prior distribution on θ is

$$g(\theta) = \begin{cases} 2\theta & 0 \leq \theta \leq 1 \\ 0 & o.w. \end{cases}$$

And $X|\theta$ follows a Geometric distribution with θ. Suppose we have one observation $X = 3$, find the MAP estimate of θ.