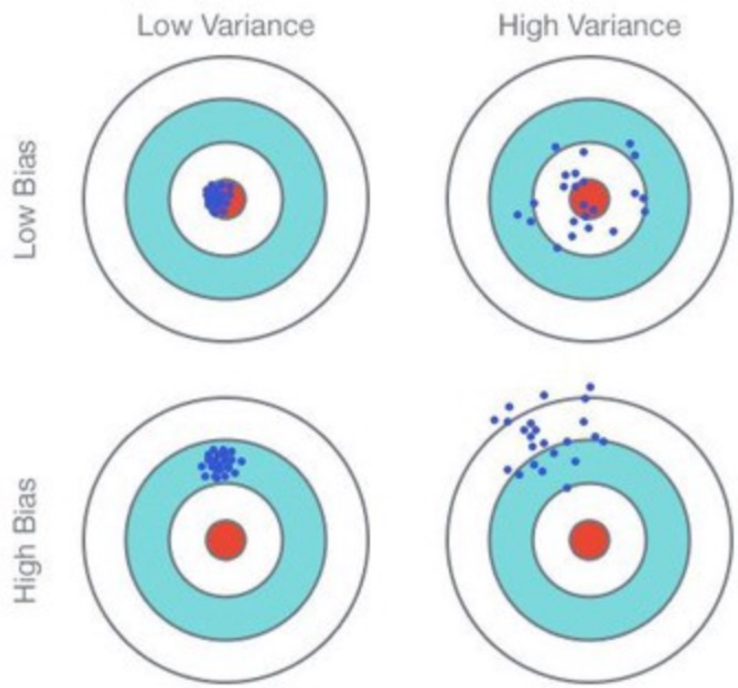
Stat 88: Probability & Mathematical Statistics in Data Science



https://medium.com/@mp32445/understanding-bias-variance-tradeoff-ca59a22e2a83

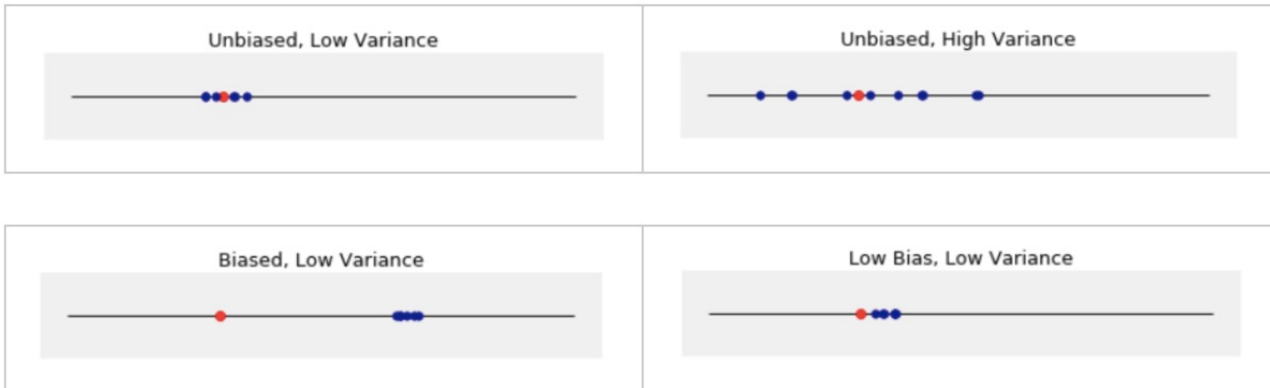
Fig. 1: Graphical Illustration of bias-variance trade-off , Source: Scott Fortmann-Roe., Understanding Bias-Variance Trade-off

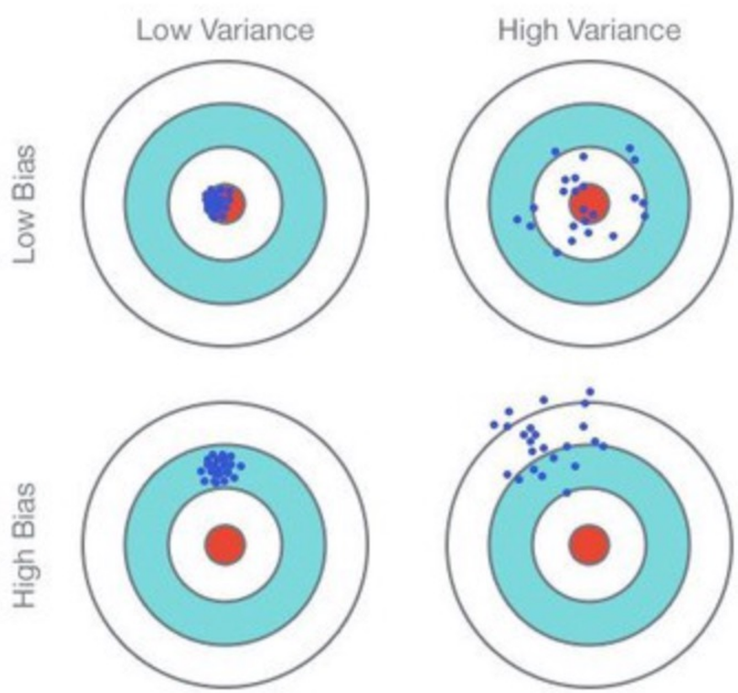
Lecture 36 : 4/21/2021

Chapter 11

Bias, Variance, and Least Squares

Understanding Bias and Variance





T: estimator (rv)

θ: parameter (target, constant)

Say T is *unbiased* if $E(T) = \theta$

$$MSE_\theta(T) = E[(T - \theta)^2]$$

Bias, Variance, and Mean Squared Error

- Bias: $\boldsymbol{B_\theta(T) = E_\theta(T) - \theta}$ (note that $B_\theta(T)$ is a constant)
- Bias is difference between expected value of the estimator and the target.
- Suppose B_θ is positive, what does this mean?

- Deviation (from the mean): $\boldsymbol{D_\theta(T) = T - E_\theta(T)}$ (note that $D_\theta(T)$ is a r.v.)

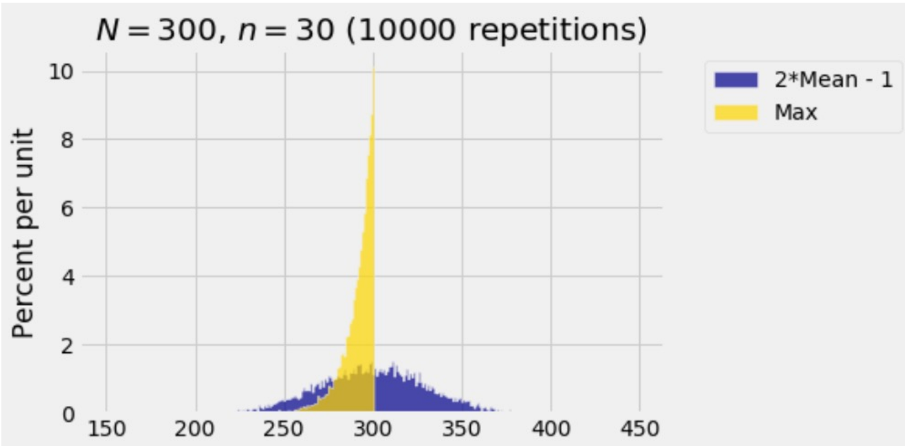
- Error: $\boldsymbol{T - \theta} =$
- Mean Squared Error: $\boldsymbol{MSE_\theta(T) = E[(T - \theta)^2]}$

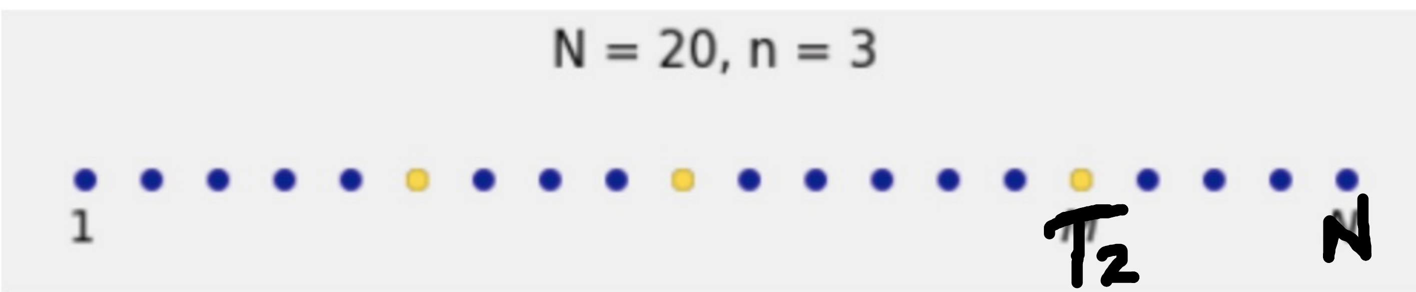
- What is the expected value of $D_\theta(T)$? What about $\left(D_\theta(T)\right)^2$?

Mean Squared Error & the Bias-Variance Decomposition

- $MSE_\theta(T) = E[(T-\theta)^2] =$

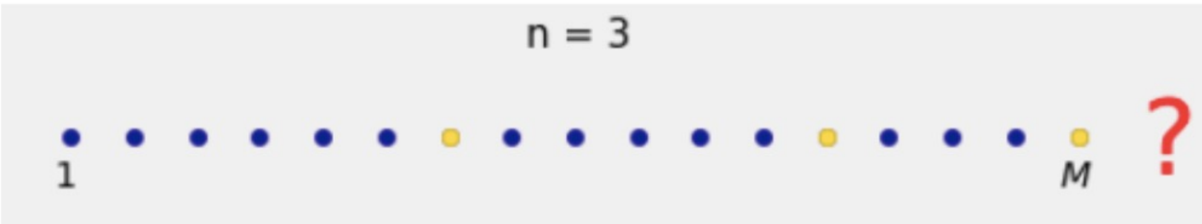
German Tank Problem: $T_1, T_2, \& T_3$





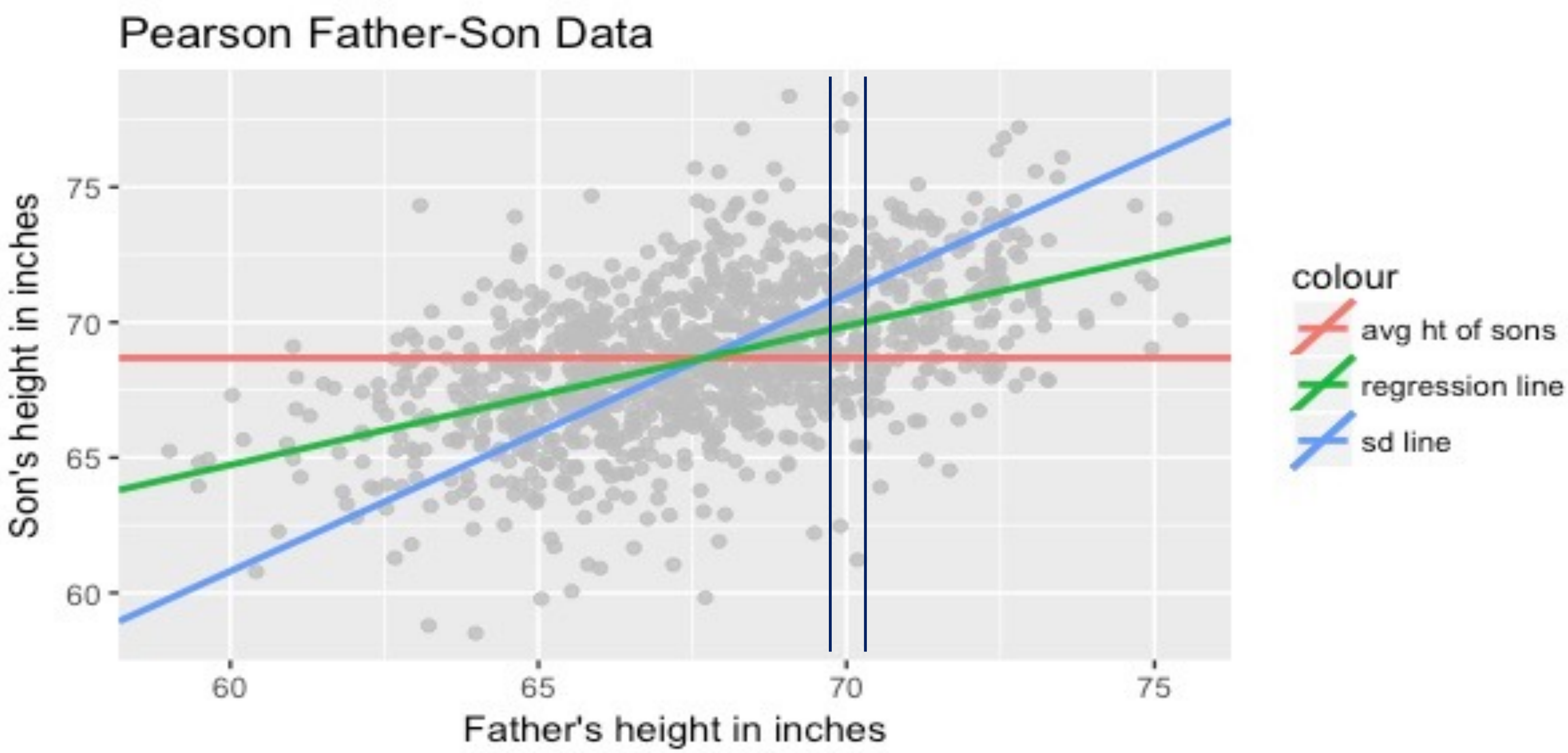
Comparing $MSE(T_2)$ & $MSE(T_3)$

The Augmented Maximum



Simple Linear Regression

- Random pair (X, Y)
- Want to use a linear function of X to estimate Y, say $aX + b$
- Best (in what sense) line for these data.



Want to predict y from x. Could use:

- Average of y (so don't use x at all)
- The SD (diagonal) line: better, but not so good (too steep)
- Much better, if the scatter plot shows a linear relationship, to use the **regression method**, which incorporates the correlation.

The regression method

- The regression method is used to draw the regression line which can be used for prediction.

- It is also called the **least squares line** because it minimizes mean squared error. By *error* we mean the vertical difference between the y-value for some x, and the height of the regression line at that x.

$$e_i = y_i - (ax_i + b), i = 1, 2, \dots, n$$

- From Data 8, do you recall the slope of the regression line? What about the intercept?