On Bias and Variance for Regression

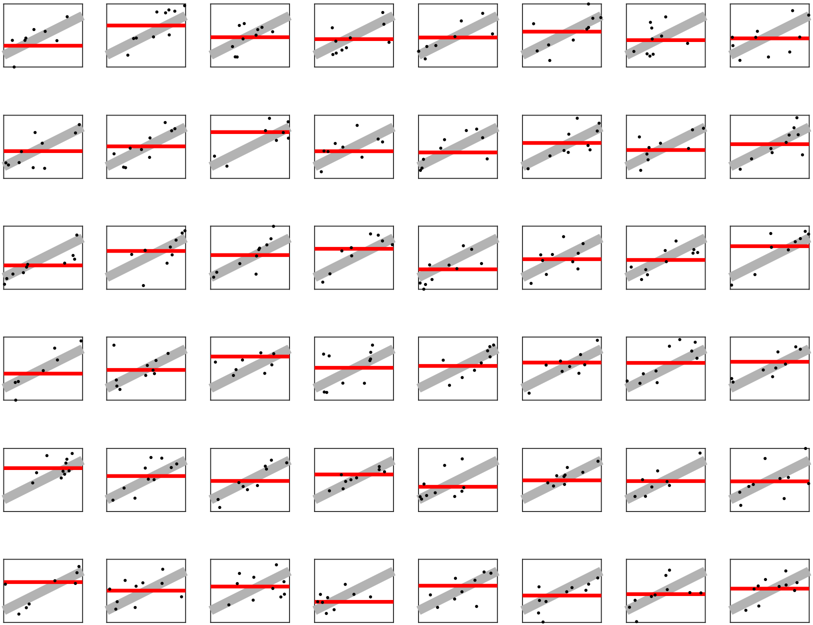
CS189/289A: Introduction to Machine Learning

Stella Yu

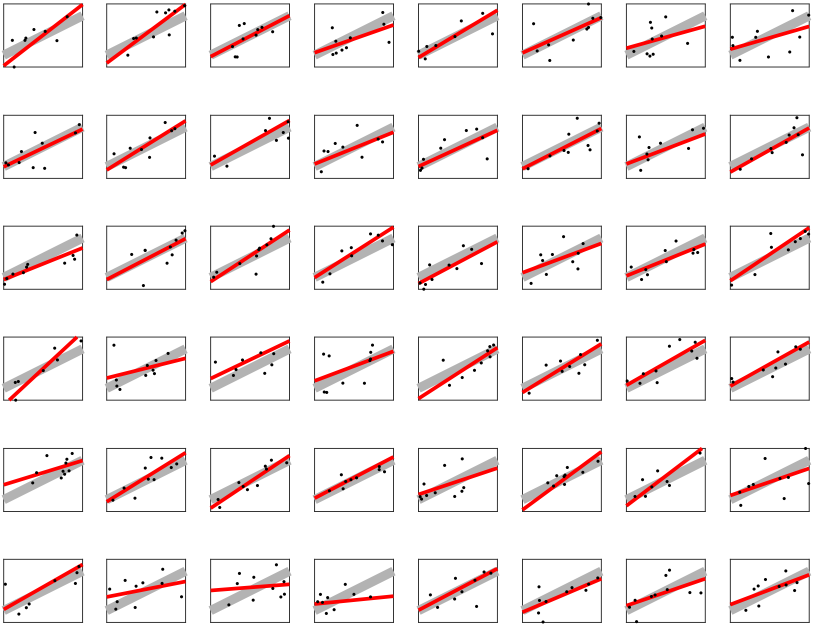
UC Berkeley

12 September 2017

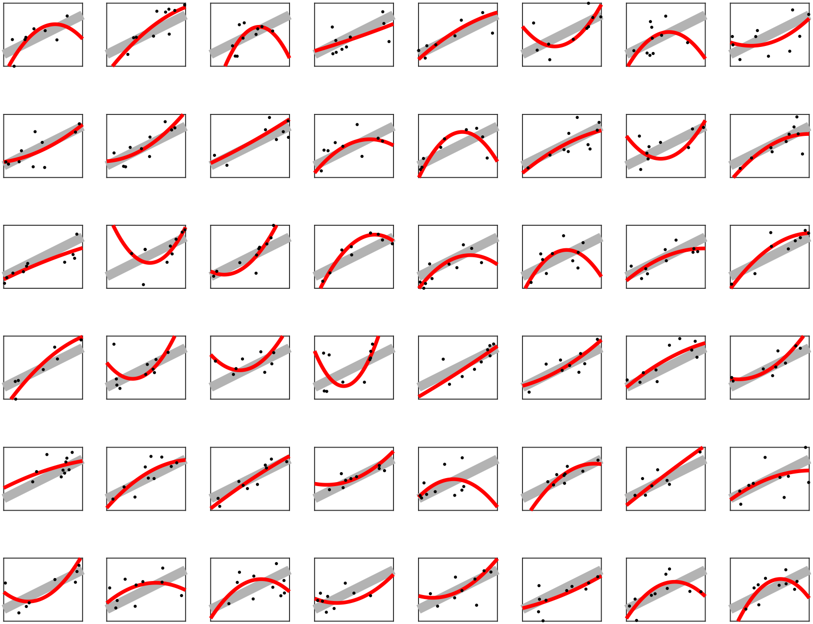
Fitting A Model over Multiple Datasets: $p = 0$



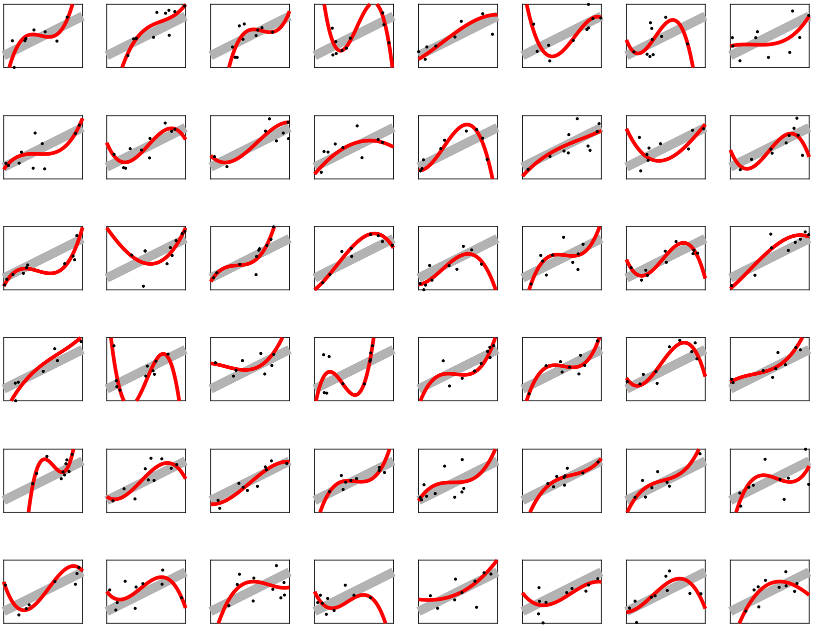
Fitting A Model over Multiple Datasets: $p = 1$



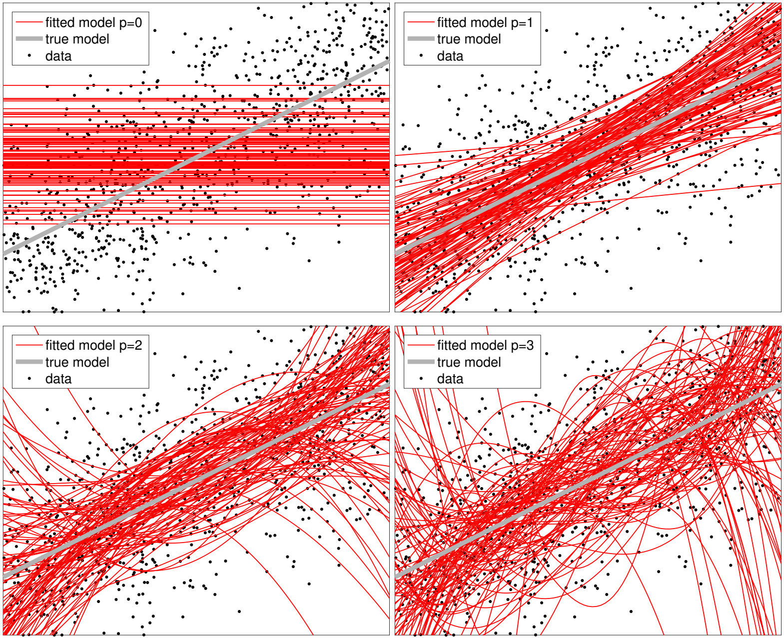
Fitting A Model over Multiple Datasets: $p = 2$



Fitting A Model over Multiple Datasets: $p = 3$



Model Quality Assessment and Model Selection



Task: Regressor Variation upon Data Selection

- Observation model of random variables (X, Y):

$$Y = f(X) + N \tag{1}$$

  The true function $f(x)$ is fixed but unknown, the noise N is zero-mean, independent and identically distributed.

- Observation dataset D of n random samples:

$$D = (X_1, Y_1; X_2, Y_2; \ldots; X_n, Y_n). \tag{2}$$

- Regressor h tries to approximate f and is dependent upon D:

$$h(x; D) \approx f(x). \tag{3}$$

- For an arbitrary point x, not necessarily in D, its Y is random:

$$Y = f(x) + N. \tag{4}$$

- How much deviation is expected from h's prediction of Y?

$$E[(h(x; D) - Y)^2] = ? \tag{5}$$

Metric: Prediction Error of A Regressor

- How good is $h(x;D)$ at estimating response Y?

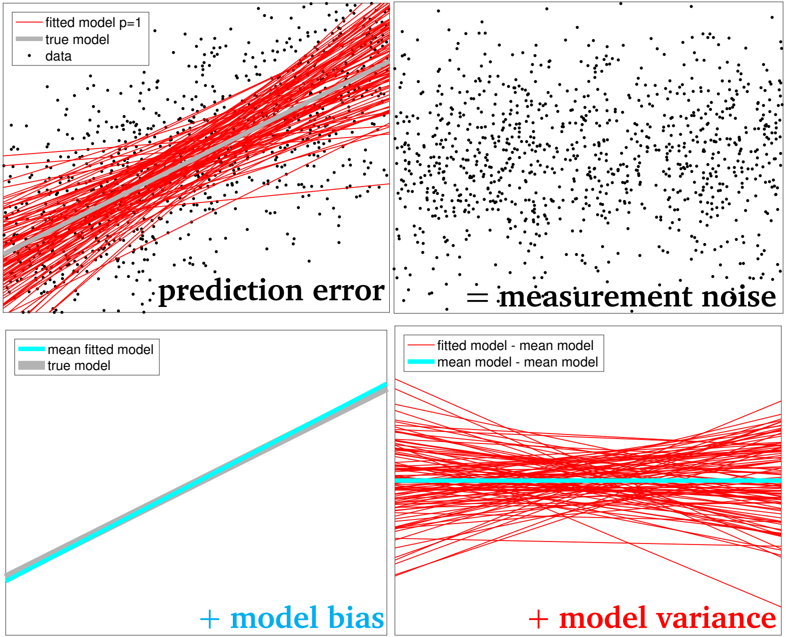
$$\varepsilon(x;h) = E[(h(x;D) - Y)^2] \tag{6}$$

- It measures the difference between the regressor prediction and the measured data, averaged over all possible training data sets of a particular sample size n.

- The metric varies with x, the location in the data space.

- The randomness comes from the dataset D used to estimate h and the inherent uncertainty in the target measurement Y.

Bias and Variance Decomposition Aspects



Bias-Variance Decomposition: Basics

- N and Y at any given x:

$$E[N] = 0 \tag{7}$$
$$E[Y] = E[f(x) + N] = f(x) \tag{8}$$
$$V[Y] = V[f(x) + N] = V[N] \tag{9}$$

- For any random variable X, the following equation holds:

$$V[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2 \tag{10}$$
$$\Downarrow$$
$$E[X^2] = V[X] + E[X]^2 \tag{11}$$

Bias-Variance Decomposition Equation

Take expectation over random dataset D and noisy Y at given x:

$$
\begin{aligned}
& E[(h(x;D)-Y)^2] && (12)\\
= \quad & E[h(x;D)^2]+E[Y^2]-2\cdot E[h(x;D)\cdot Y] && (13)\\
= \quad & E[h(x;D)]^2+V[h(x;D)] && (14)\\
+ \quad & E[Y]^2+V[Y] && (15)\\
- \quad & 2\cdot E[h(x;D)]\cdot E[Y] \quad \Leftarrow h(x;D), Y \text{ are independent} && (16)\\
= \quad & (E[h(x;D)]-E[Y])^2+V[h(x;D)]+V[Y] && (17)\\
= \quad & \underbrace{(E[h(x;D)]-f(x))^2}_{\text{bias}^2\text{ of method}}+\underbrace{V[h(x;D)]}_{\text{variance of method}}+\underbrace{V[N]}_{\text{irreducible error}} && (18)
\end{aligned}
$$

Bias-Variance Decomposition Intuitions

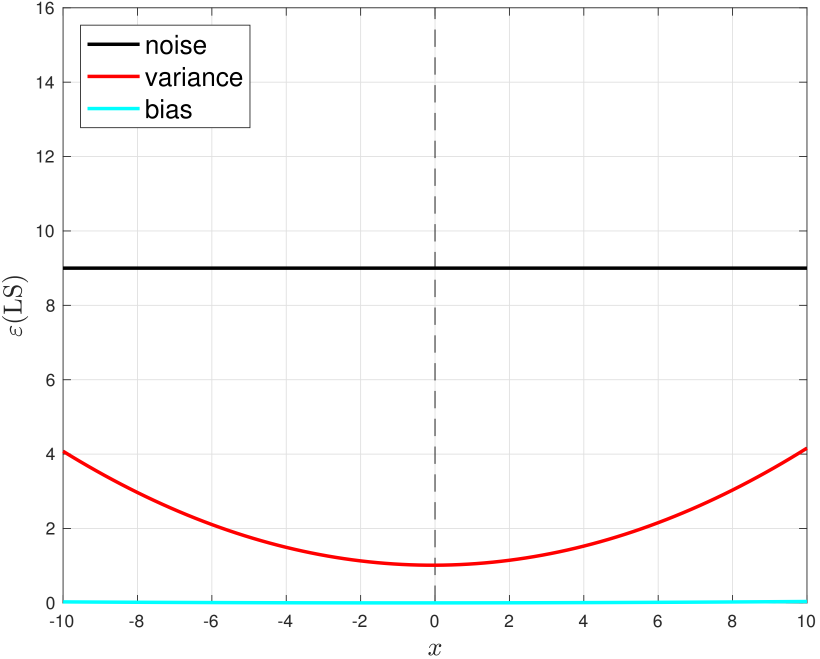
The prediction metric can be decomposed into three terms.

1. The **bias of the method** measures the **accuracy** of the prediction with respect to the true value. A low bias means that on average the regressor $h(x)$ accurately estimates $f(x)$.

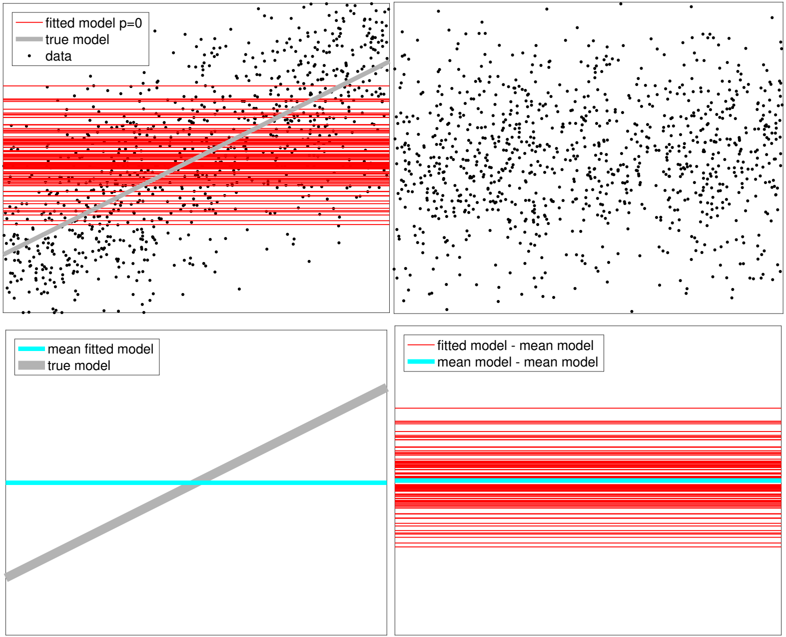
2. The **variance of the method** measures the **stability** of the prediction. A low variance means that the prediction does not change much as the training set varies. An un-biased method (bias $= 0$) could have a large variance.

3. The **irreducible error** measures the **precision** of the prediction. A low irreducible error means that the prediction can be very precise to the true value.

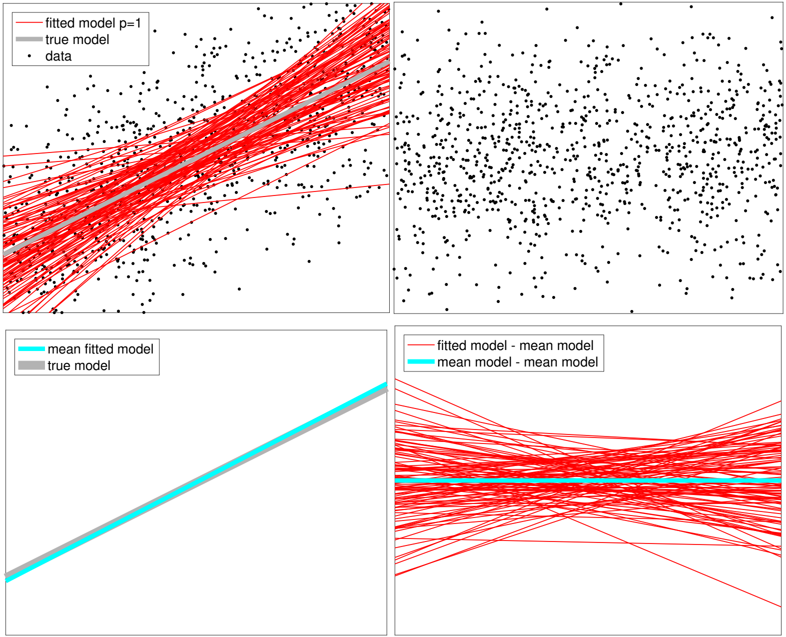
Bias vs. Variance, Interpolation vs. Extrapolation



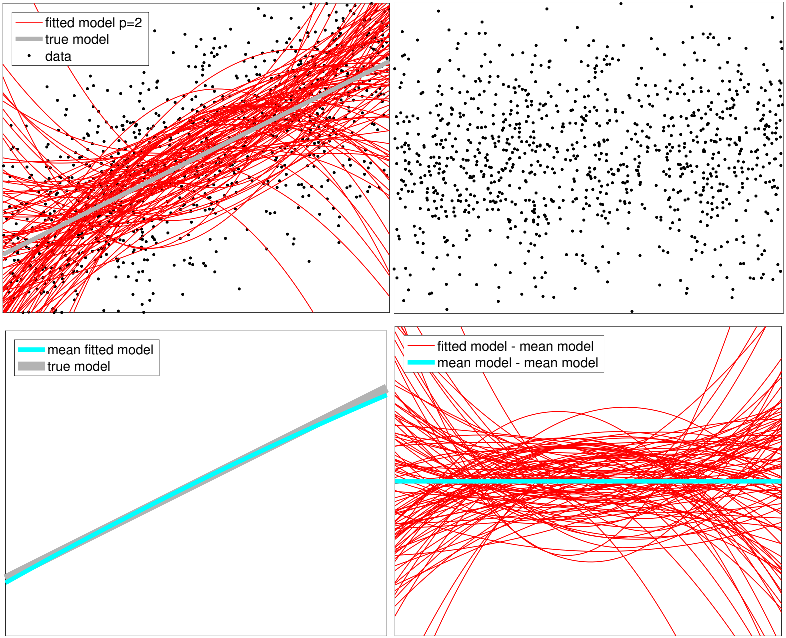
Bias and Variance in Model Selection: $p = 0$



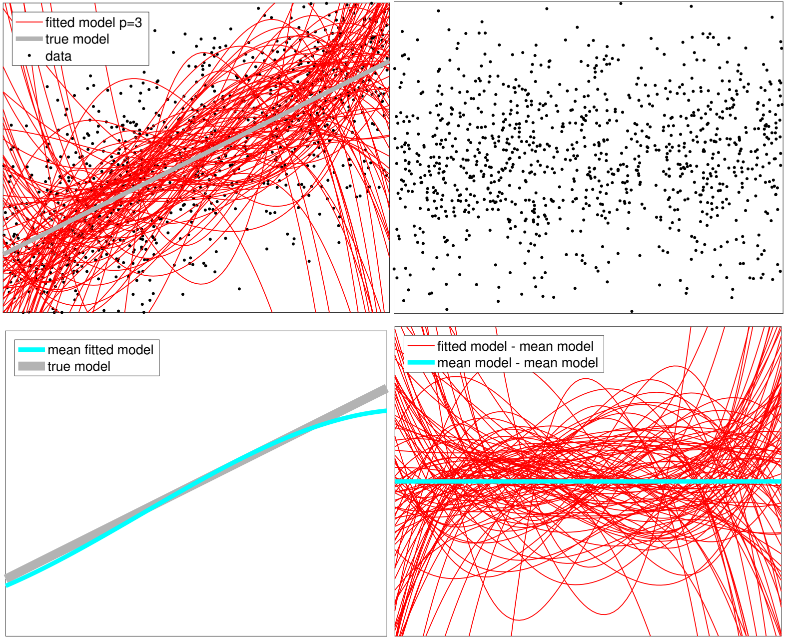
Bias and Variance in Model Selection: $p = 1$



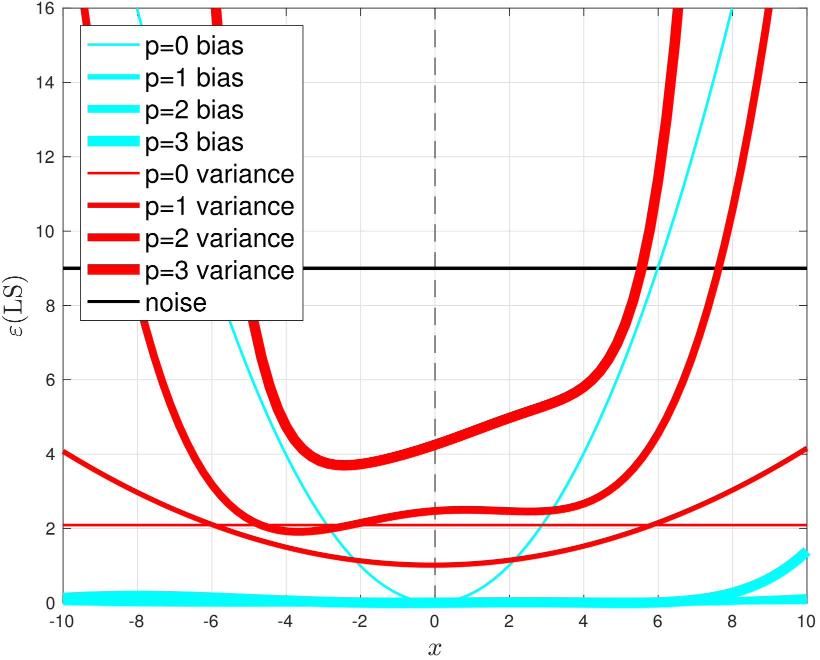
Bias and Variance in Model Selection: $p = 2$

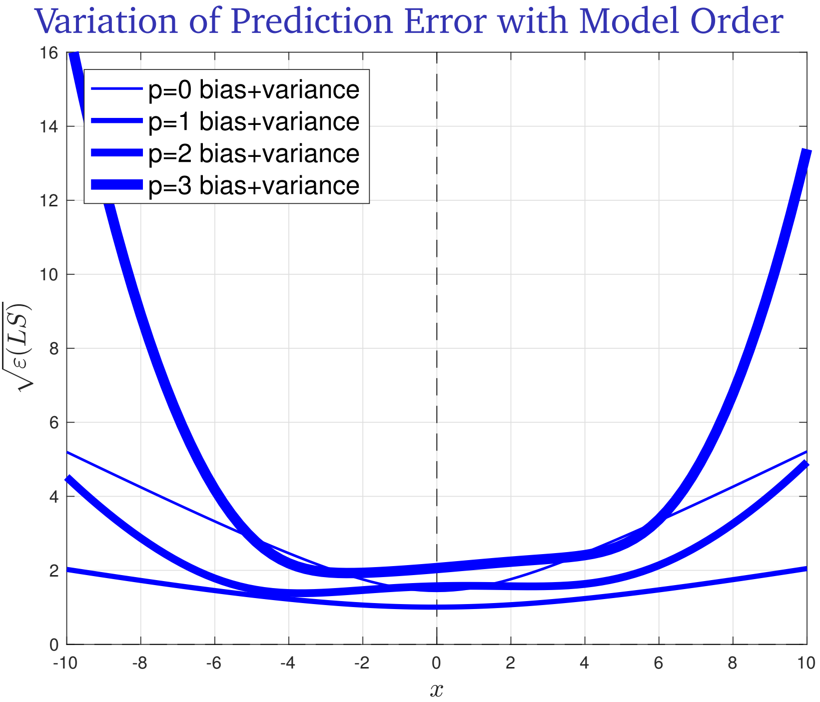


Bias and Variance in Model Selection: $p = 3$

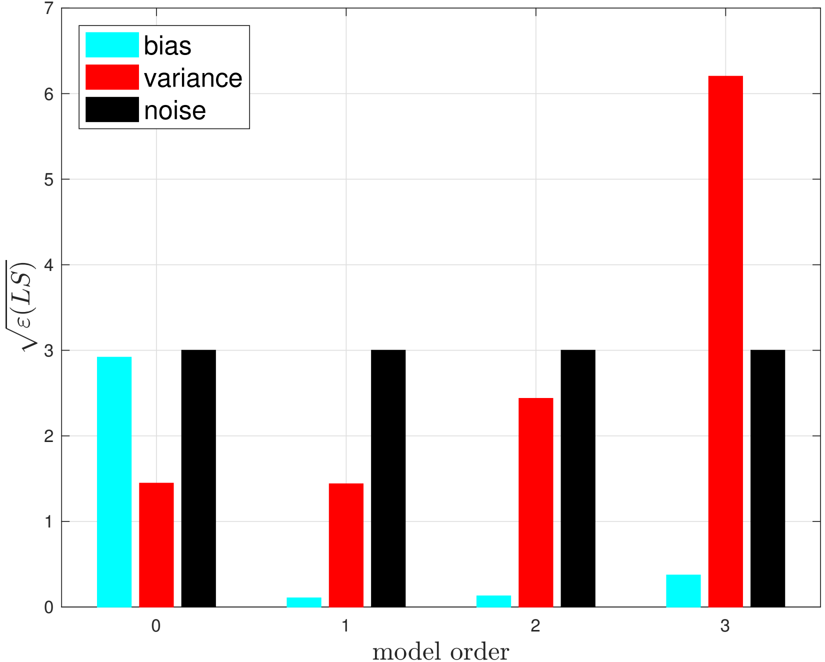


Variation of Bias/Variance Over Model Order

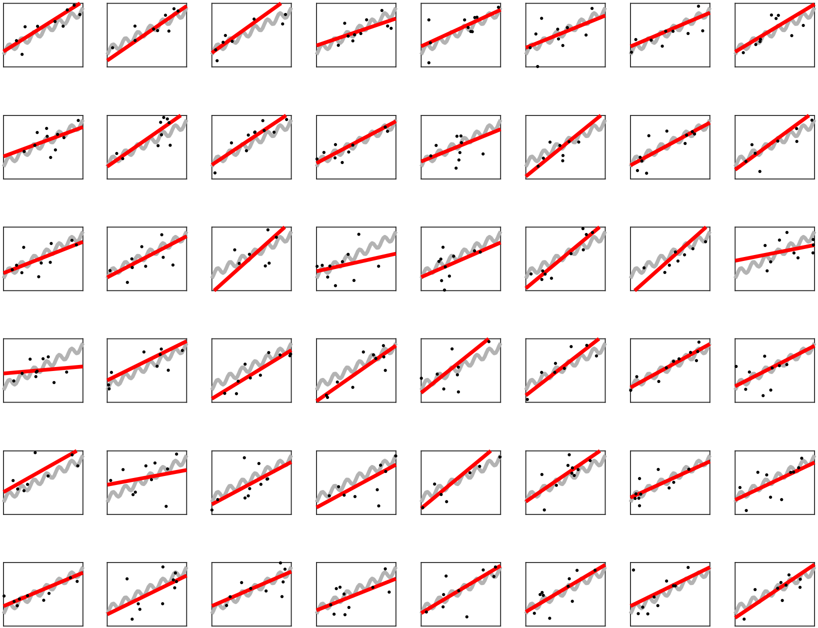



Variation of Prediction Error with Model Order
p=0 bias+variance
p=1 bias+variance
p=2 bias+variance
p=3 bias+variance
$\sqrt{\varepsilon(LS)}$
x
0
2
4
6
8
10
12
14
16
-10
-8
-6
-4
-2
0
2
4
6
8
10

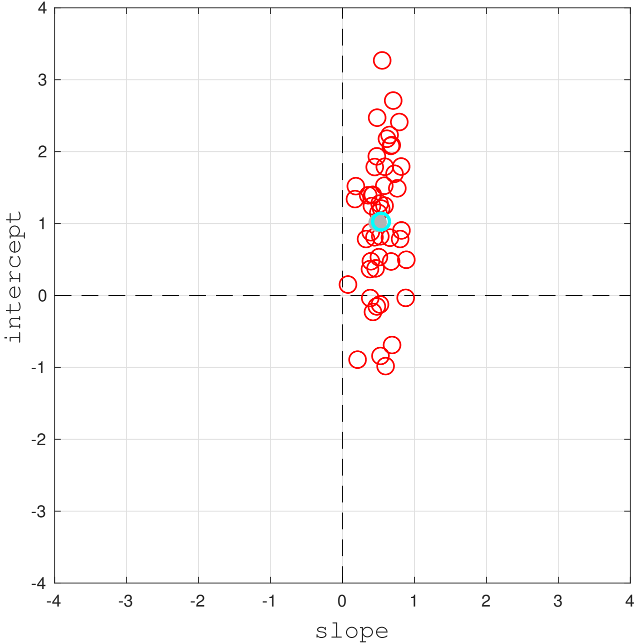
Bias and Variance: Underfitting vs. Overfitting



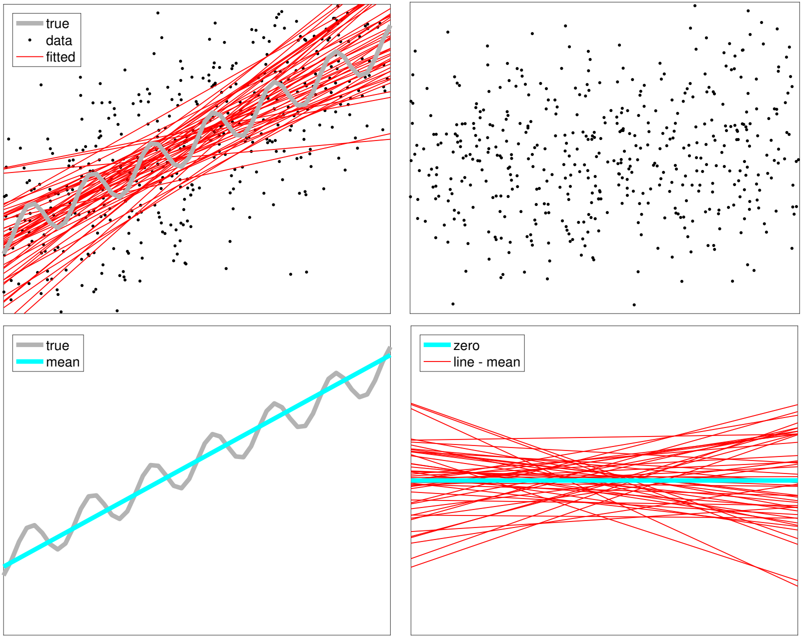
Fitting A Line Over 10 Points in 48 Random Sets



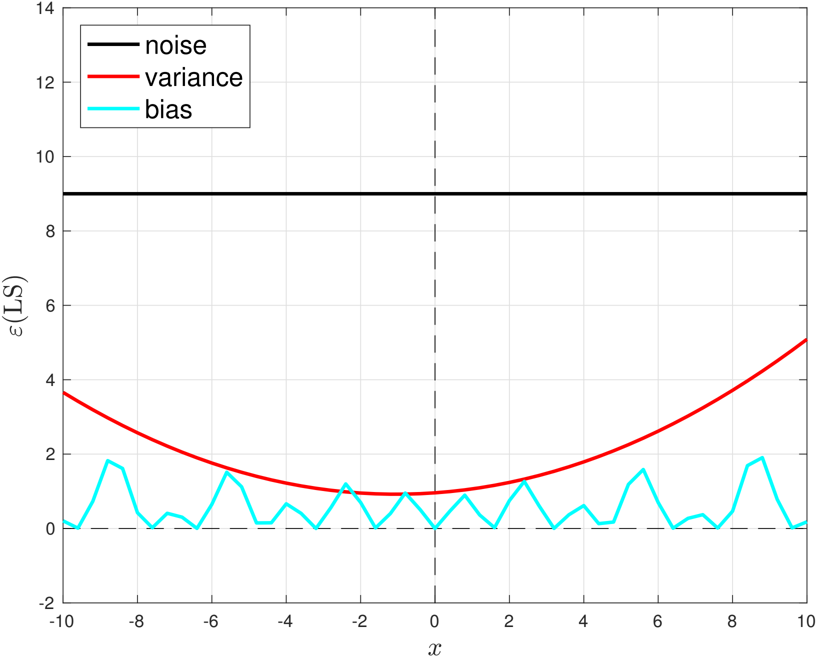
Random Variable h in the Model Space: $n = 10$



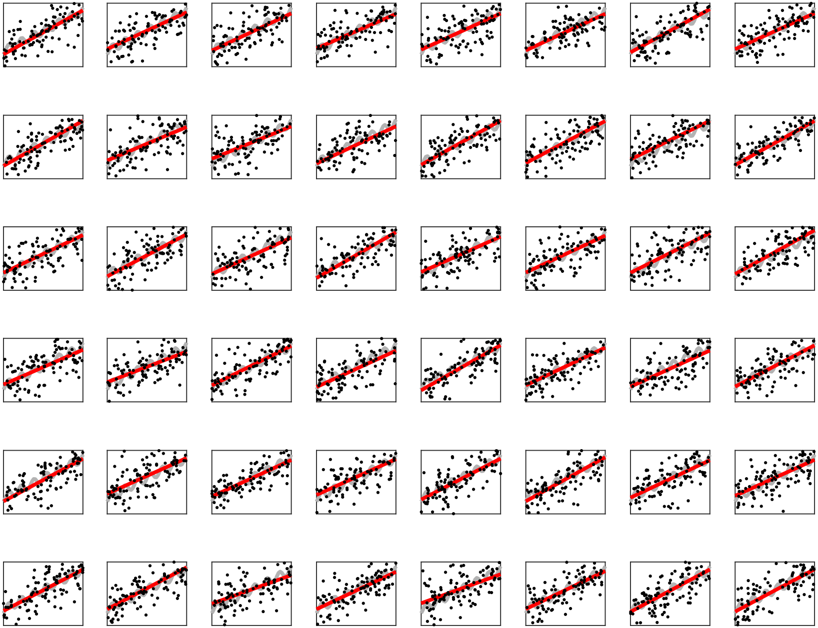
Bias-Variance Decomposition Aspects: $n = 10$



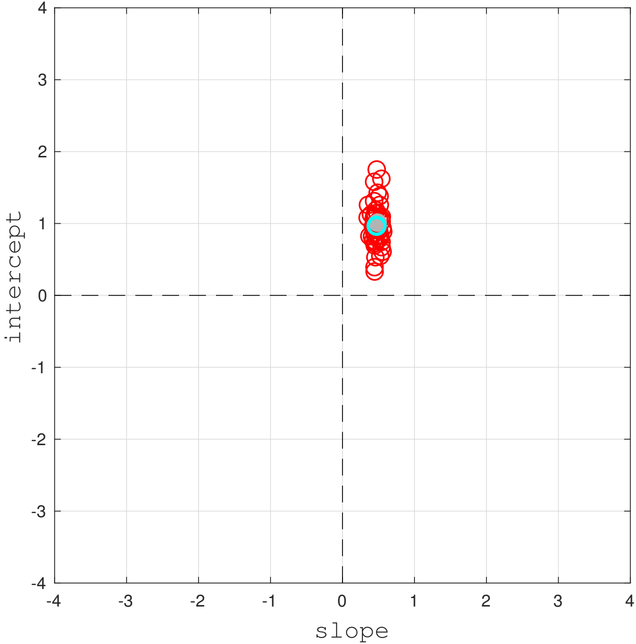
Bias-Variance Decomposition Results: $n = 10$



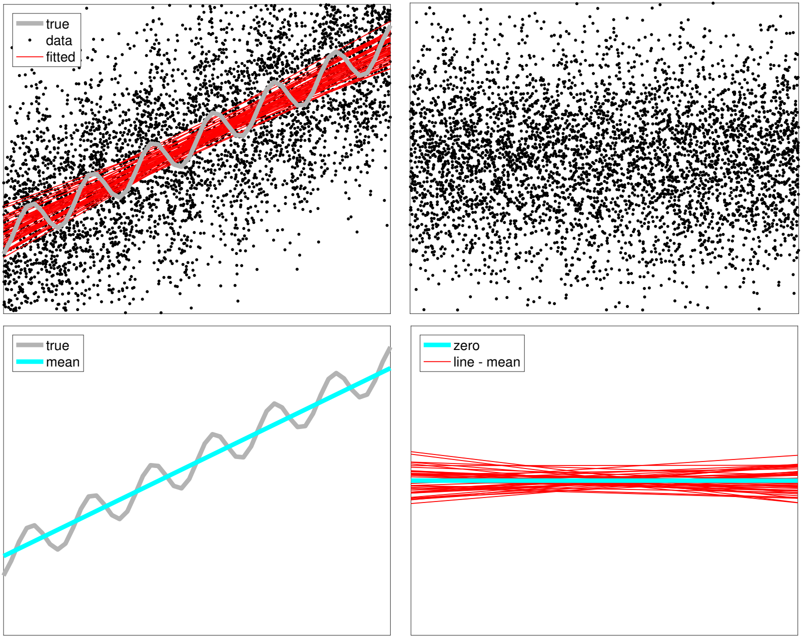
Fitting A Line Over 100 Points in 48 Random Sets



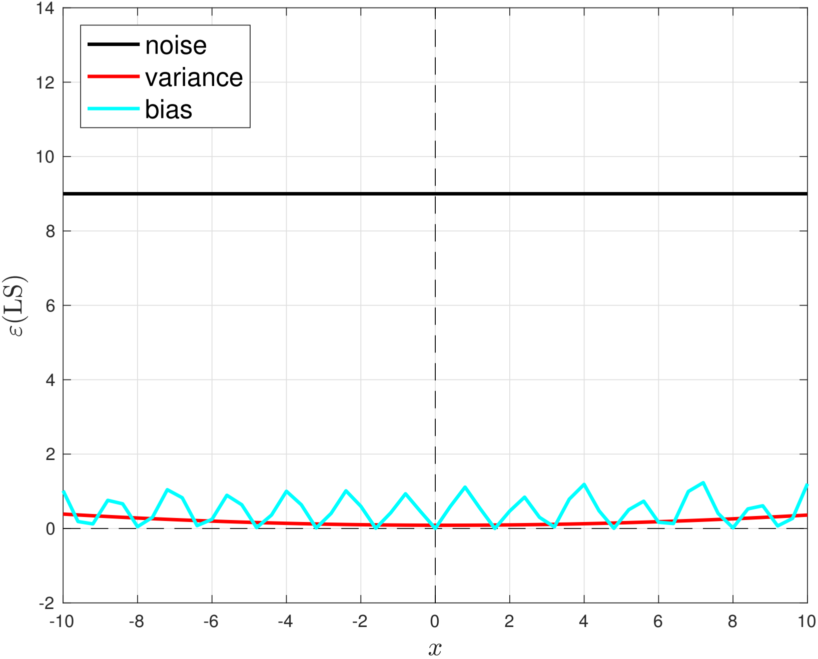
Random Variable h in the Model Space: $n = 100$



Bias-Variance Decomposition Aspects: $n = 100$



Bias-Variance Decomposition Results: $n = 100$



Understanding Model Estimation

1. Under-fitting = much bias; most overfitting = much variance
2. Training error reflects bias but not variance; test error reflects both; low training error can fool you when you've overfitted
3. Variance $\to 0$ as $n \to \infty$
4. If h can fit f exactly, for many distributions, bias $\to 0$ as $n \to \infty$. If h cannot fit f well, bias is large at most points
5. Adding a good feature reduces bias
   Adding a bad feature rarely increases it
6. Adding a feature usually increases variance
   Don't add a feature unless it reduces bias more
7. Can't reduce irreducible error, hence its name
8. Noise in test set affects only $V[N]$
   Noise in training set affects bias and variance.
9. For real-world data, f is rarely knowable and noise model might be wrong, so we can't actually calculate bias and variance. But we can test algorithms on synthetic data.