Chapter 2: Statistical Learning

- Input variable: predictors, independent variables, features
- Output variable: response, dependent variable
- $Y = f(X) + \epsilon$, where E[ϵ]=0
- Why estimate *f*? Prediction, inference (understanding)
- Let $\hat{Y} = \hat{f}(X)$, then we can examine the prediction error as

$$E\left(Y - \hat{Y}\right)^2 = E\left[f(X) + \epsilon - \hat{f}(X)\right]^2, \text{ let } \left[f(X) - \hat{f}(X)\right] = A$$
$$= E[(A + \epsilon)^2] = E[A^2 + (\epsilon - 0)^2 + 2A\epsilon]$$
$$= \left[f(X) - \hat{f}(X)\right]^2 + Var(\epsilon) + 0$$

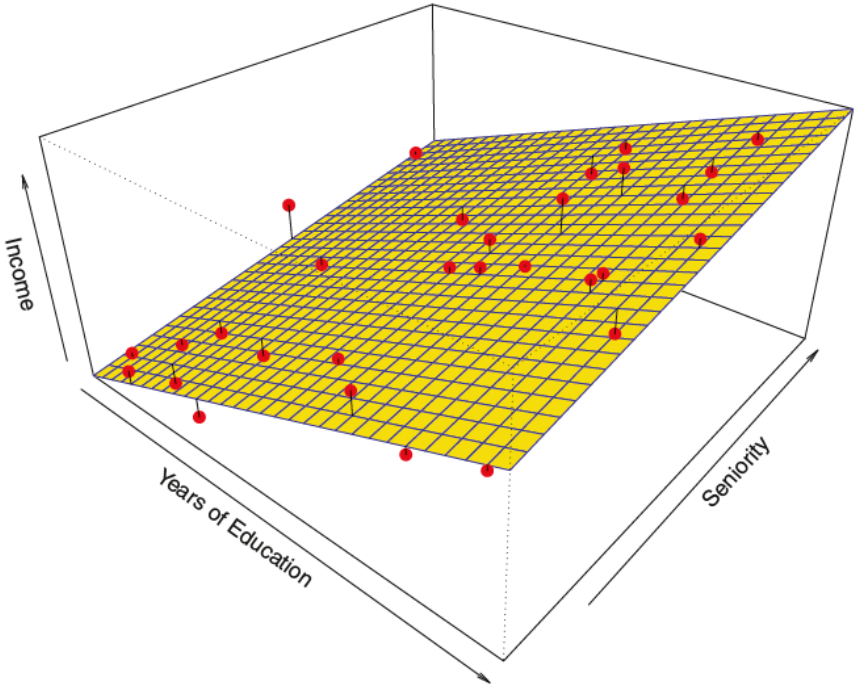
since X and $\hat{f}(X)$ are assumed constant.

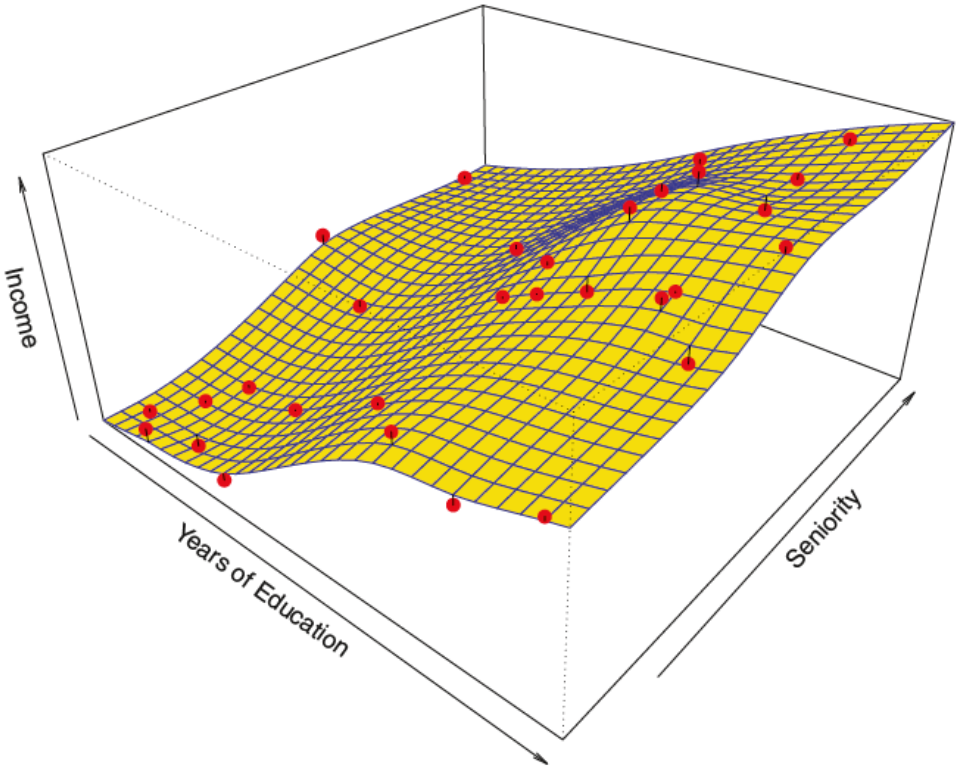
The first term can be reduced by choice of our estimate of *f*.

Inference or Understanding

- Which predictors or features are associated with the response
- What is the relationship between the important predictors and the response

Estimation of f





Linear regression

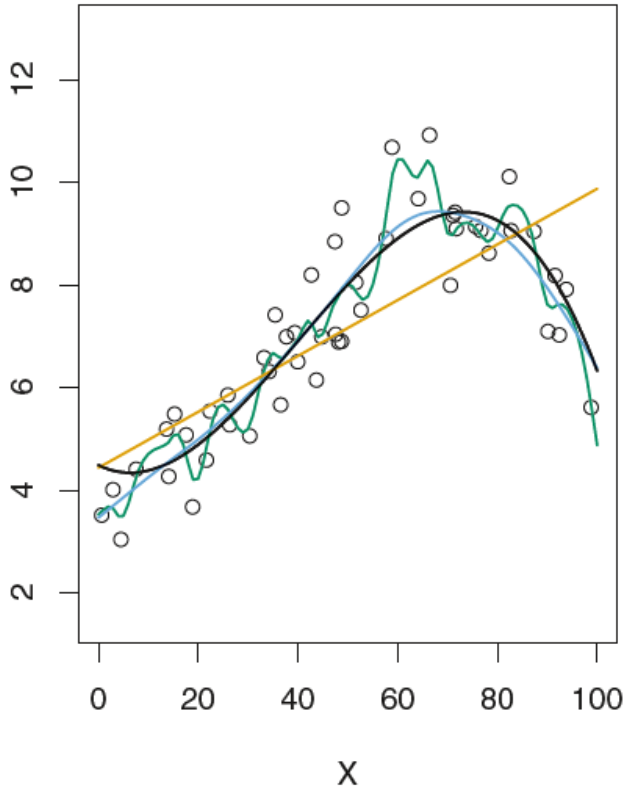
Non-parametric thin plate spline typically will require more data

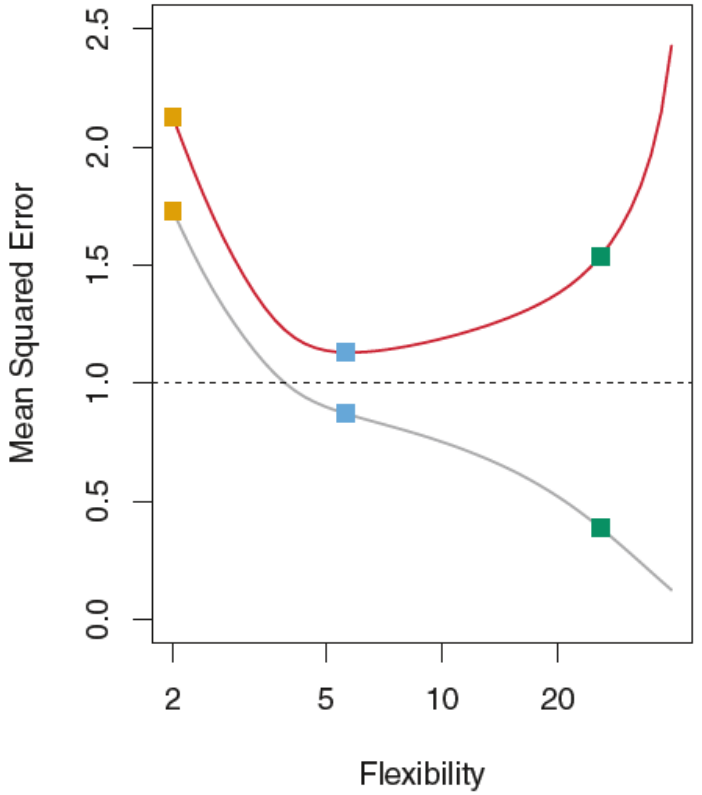
Supervised vs. Unsupervised Learning

- When samples include features and responses we can use supervised learning since we can use the responses as a way of judging the predictions.
- When observations are vectors of measurements but no responses we use unsupervised learning. Example: we have genetic data from many individuals that may be from a single species or multiple sibling species. Clustering techniques, principal components may be used to determine the relationships.

Assessing Model Accuracy

- Mean squared error:
  $MSE = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}(x_i)\right]^2$
- If the same data is used to estimate $\hat{f}$ and MSE, then the estimate of MSE will be too small.
- To avoid this bias the data is divided into a training set (to estimate $\hat{f}$) and a test set to estimate MSE.





Bias vs. Variance Trade-Off

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

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = E\left[f(x_0) - E\left[\hat{f}(x_0)\right] + E\left[\hat{f}(x_0)\right] - \hat{f}(x_0) + \epsilon\right]^2$$

$$(-1)^2 E\left\{\left[E\left[\hat{f}(x_0)\right] - f(x_0)\right] + \left[\hat{f}(x_0) - E\left[\hat{f}(x_0)\right]\right] - \epsilon\right\}^2$$

$$recall\ that f(x_0)\ and\ E\left[\hat{f}(x_0)\right]\ are\ constants, let\ the\ bias = a$$

$$= E\left[\hat{f}(x_0)\right] - f(x_0)$$

Bias vs. Variance Trade-Off (cont)

- $MSE = E\left\{a + \left[\hat{f}(x_0) - E\left[\hat{f}(x_0)\right]\right] - \epsilon\right\}^2$

  $= E\left\{a + \left[\hat{f}(x_0) - E\left[\hat{f}(x_0)\right]\right]\right\}^2 + E(\epsilon^2) + 2E\left[a\epsilon + \epsilon\left[\hat{f}(x_0) - E\left[\hat{f}(x_0)\right]\right]\right]$

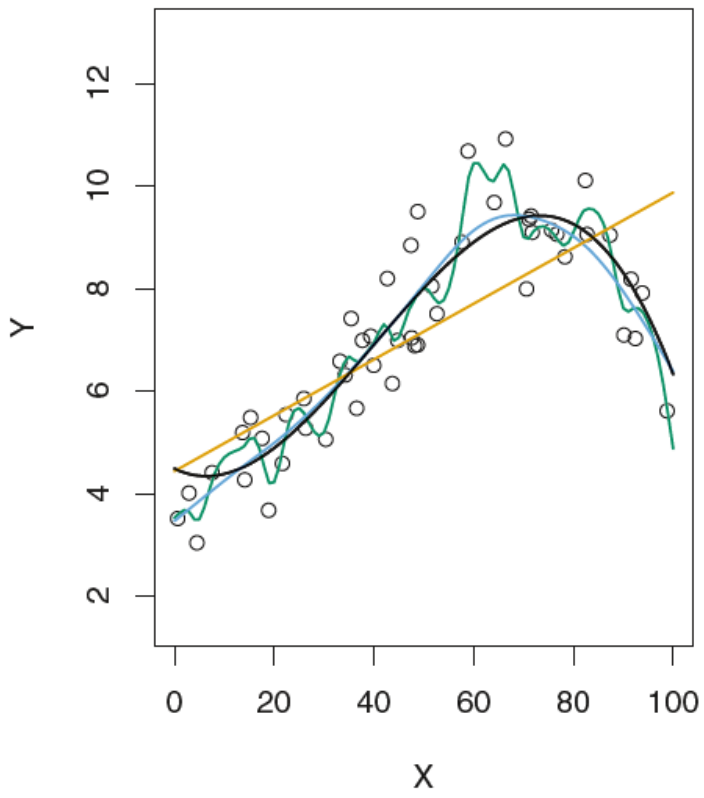
  let $z = \hat{f}(x_0) - E\left[\hat{f}(x_0)\right]$ then

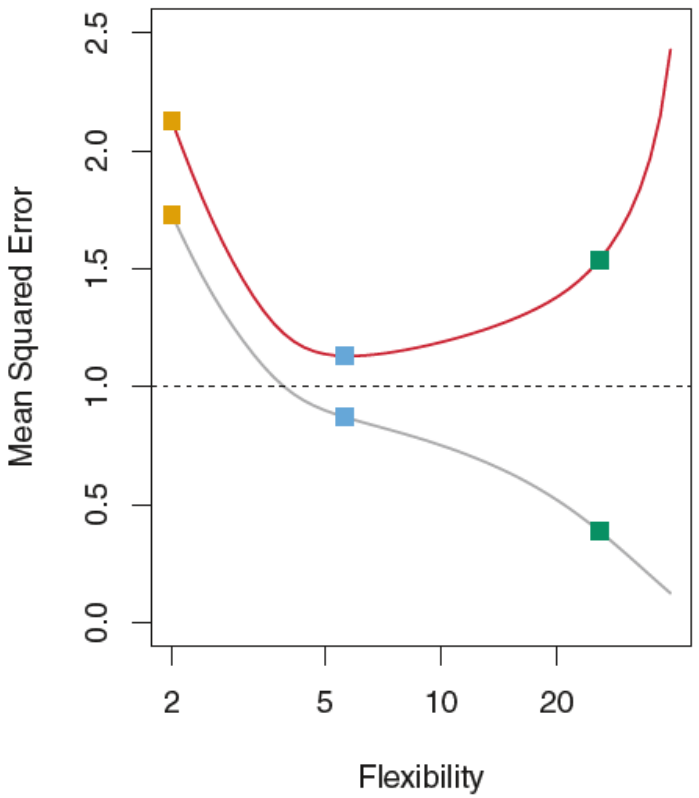
  $\text{MSE} = E\{a + z\}^2 + Var(\epsilon) + 0, since\ \epsilon\ and\ \hat{f}(x_0)\ are\ independent$

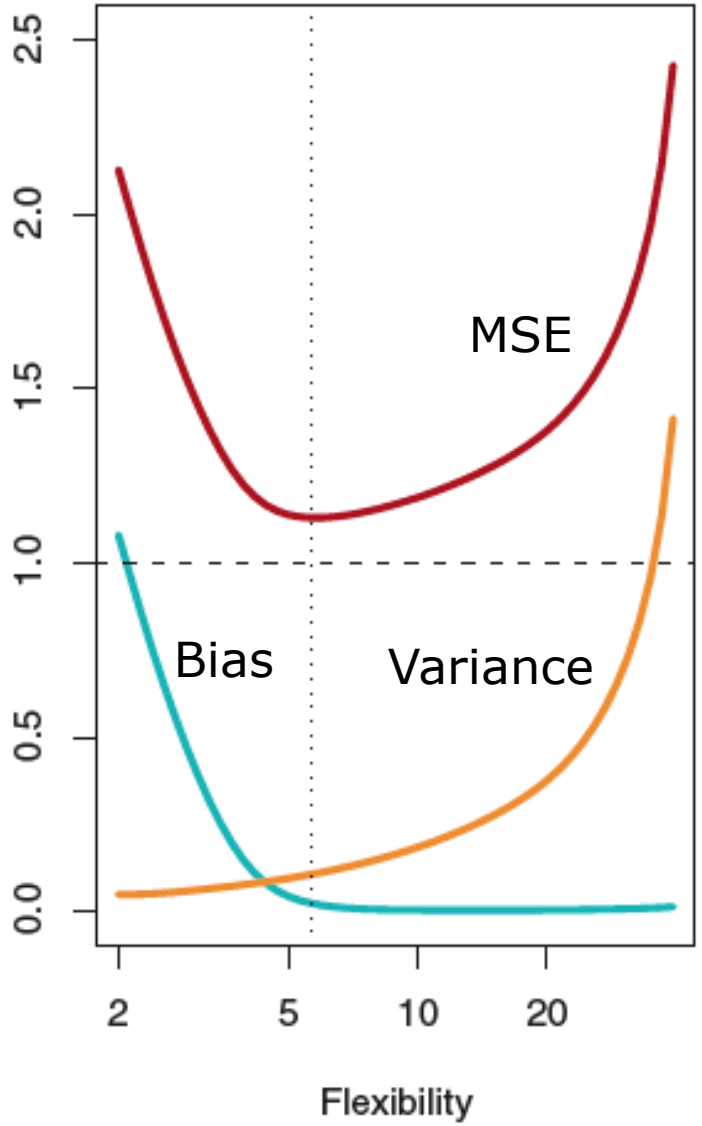
  $E(a^2) + E(z^2) + 2aE(z) + Var(\epsilon), but\ E(z) = 0$

  $bias^2 + Var\left(\hat{f}(x_0)\right) + Var(\epsilon)$

Bias vs. Variance Trade-Off





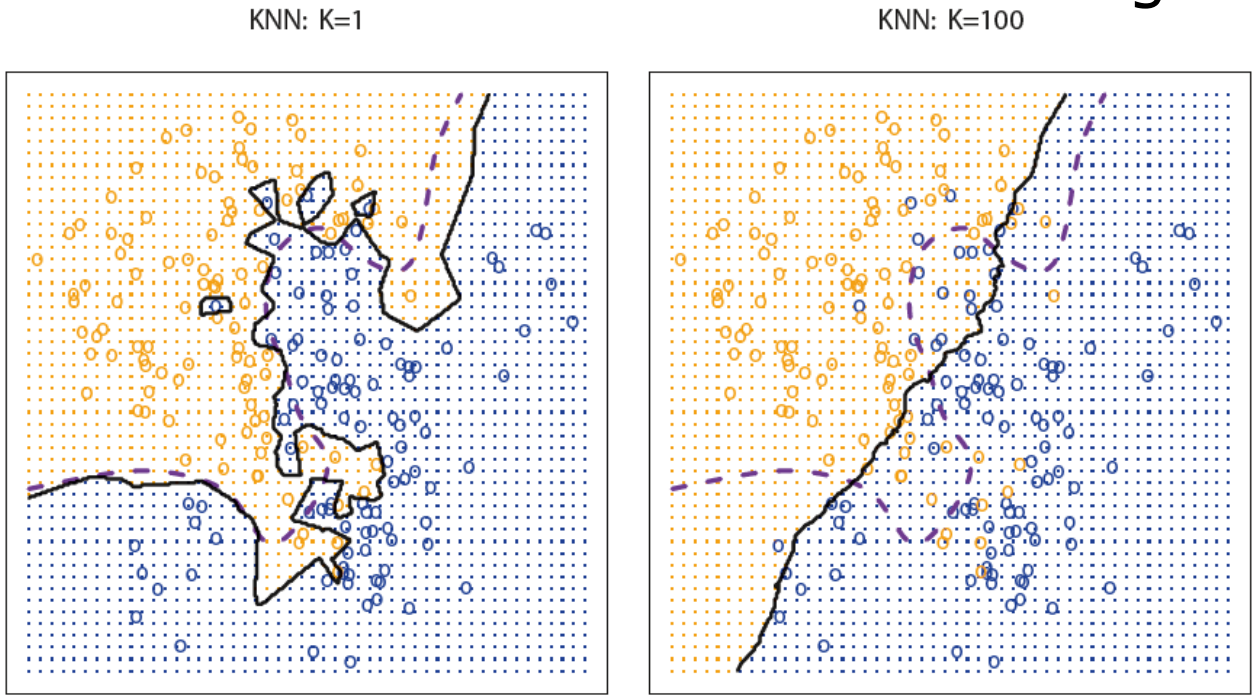


Classification Problems

- The response variables are 2 or more qualitative variables, e.g. *D. melanogaster* or *D. simulans*; coniferous forest or deciduous forest
- Quantify the accuracy of a model by computing with a test set of n observations, $\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$ this is the proportion of incorrectly predicted responses.

K nearest neighbors

- *K* is a positive integer and for any point, x_0, defines a , N_0, of observations that are closest to x_0. Then for each response variable (j) calculate, $\Pr(Y = j|X = x_0)\frac{1}{K}\sum_{i\in \mathrm{N}_0} I(y_i = j)$
- Choose the class which has the highest conditional probability.



Predict membership in "blue" or "orange" based on x_1 and x_2.

With k=1, there is low bias but high Variance. At k=100, the variance is Reduced but the bias is much higher.