Risk Optimization and Bias-Variance Trade-Off

Dudoit

Bias-Variance Trade-Off

Regression: MSE and MSPE

Risk Optimization and Bias-Variance Trade-Off

Data 100: Principles and Techniques of Data Science

Sandrine Dudoit

Department of Statistics and Division of Biostatistics, UC Berkeley

Spring 2019

Outline

1. Bias-Variance Trade-Off

2. Regression: MSE and MSPE

Version: 19/03/2019, 18:03

Regression Example

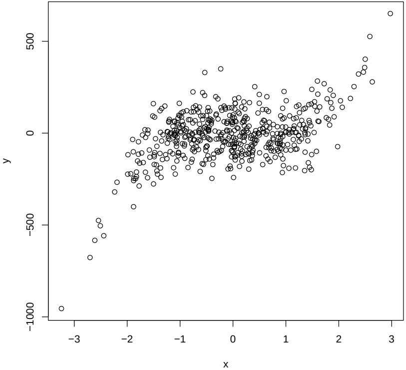

Figure 1: *Regression.* Scatterplot of 500 covariate-outcome pairs from an unknown data generating distribution. What is the regression function?

Regression Example

- Suppose we have a learning set $\mathcal{L}_n = \{(X_i, Y_i) : i = 1, \dots, n\}$ of $n = 500$ independent and identically distributed (IID) covariate-outcome pairs from an unknown data generating distribution P.
- How can we use these data to estimate the regression function of Y on X: $\theta(X) = \mathrm{E}_P[Y|X]$?
- Based on the scatterplot of Y vs. X, it seems that the regression function is non-linear in X, i.e., a constant or linear (in X) regression function would be too simple to capture the patterns/trends suggested by the plot.
- We could try fitting polynomials in X of higher degrees. The higher the degree of the polynomial, the better the fit on the learning set.

Regression Example

- However, by arbitrarily increasing the polynomial degree, we risk fitting the noise, as opposed to the actual signal, in the learning data.

Regression Example: Model Complexity

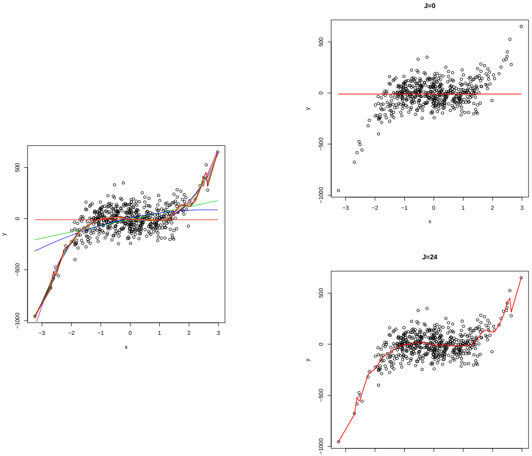

Figure 2: *Linear regression complexity.* Linear regression fits for polynomials of degree 0 to 24.

Regression Example: Model Complexity

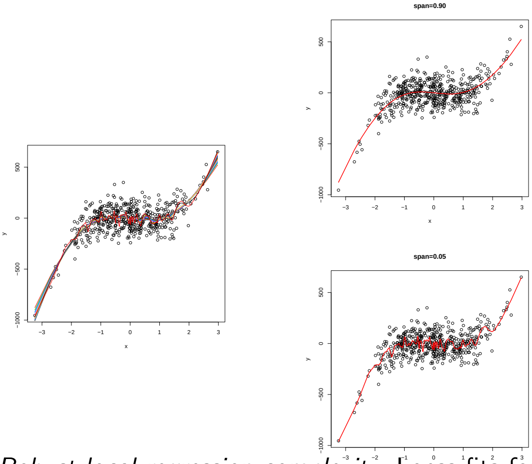

Figure 3: *Robust local regression complexity.* Loess fits for spans ranging from 0.05 to 0.90.

Bias, Variance, and Accuracy

- In some cases, we may choose a model that is too simple to represent the underlying data generation mechanism, i.e., misses the signal in the learning data.
  E.g. Fitting a constant regression function, when there is in fact a non-linear relationship between the outcome and the covariate.
- In others, we may choose a model that is too complex, i.e., fits the noise in the learning data.
  E.g. Fitting a regression function that is a high-degree polynomial of the covariate, when there is in fact a simple linear relationship between the outcome and the covariate.
- These two situations are referred to, respectively, as underfitting and overfitting the learning data.

Bias, Variance, and Accuracy

- The phenomenon of overfitting/underfitting is related to the bias of an estimator, i.e., how close its average is to the parameter of interest, and to its variance or precision, i.e., how variable it is around its expected value (not necessarily the parameter, unless the estimator is unbiased).
- Ideally, we'd like to minimize both bias and variance.
- However, this is not possible, as there is a trade-off between bias and variance: Decreasing bias is typically associated with an increase in variance and vice versa.
- In general, the more complex a model, the less biased and more variable an estimator.
- The complexity of a model or estimator can be measured in various ways.

Bias, Variance, and Accuracy

- The number of covariates for a regression function.
- The polynomial degree for a regression function.
- The number of leaf nodes for a classification or regression tree.
- The span for robust local regression (i.e., loess) and the bandwidth for kernel density estimation, i.e., how "local" a smoother is.
- The penalty parameter for regularized regression, e.g., ridge regression.
- The number of input nodes and layers for a neural network.

- Note also that, in general, variance decreases with increasing sample size, but not bias. As seen in our discussion of survey sampling, one can become more and more precise about a completely wrong answer!

Bias, Variance, and Accuracy

- Instead of attempting to simultaneously minimize both bias and variance, one seeks to minimize risk or maximize accuracy, i.e., the average "distance" between an estimator and the parameter of interest.
- Risk for the squared error loss function, i.e., mean squared error (MSE), can be decomposed in terms of bias and variance components. That is, given an estimator $\hat{\theta}$ of a parameter θ,

$$\begin{aligned} \mathrm{MSE}_P[\hat{\theta}, \theta] &\equiv \mathrm{E}_P[(\hat{\theta} - \theta)^2] \\ &= \mathrm{E}_P[(\hat{\theta} - \mathrm{E}_P[\hat{\theta}])^2] + (\mathrm{E}_P[\hat{\theta}] - \theta)^2 \\ &= \mathrm{Var}_P[\hat{\theta}] + (\mathrm{Bias}_P[\hat{\theta}, \theta])^2. \end{aligned} \tag{1}$$

In short,

$$\mathrm{MSE} = \mathrm{Variance} + \mathrm{Bias}^2.$$

Bias, Variance, and Accuracy

Proof.

$$
\begin{aligned}
\mathrm{E}_P[(\hat{\theta}-\theta)^2] &= \mathrm{E}_P[(\hat{\theta}-\mathrm{E}_P[\hat{\theta}]+\mathrm{E}_P[\hat{\theta}]-\theta)^2] \\
&= \mathrm{E}_P[(\hat{\theta}-\mathrm{E}_P[\hat{\theta}])^2]+\mathrm{E}_P[(\mathrm{E}_P[\hat{\theta}]-\theta)^2] \\
&\quad +2\,\mathrm{E}_P[(\hat{\theta}-\mathrm{E}_P[\hat{\theta}])(\mathrm{E}_P[\hat{\theta}]-\theta)] \\
&= \mathrm{Var}_P[\hat{\theta}]+(\mathrm{E}_P[\hat{\theta}]-\theta)^2 \\
&\quad +2(\mathrm{E}_P[\hat{\theta}]-\theta)\,\mathrm{E}_P[(\hat{\theta}-\mathrm{E}_P[\hat{\theta}])] \\
&= \mathrm{Var}_P[\hat{\theta}]+(\mathrm{Bias}_P[\hat{\theta},\theta])^2,
\end{aligned}
$$

where the third equality follows by noting that $\mathrm{E}_P[\hat{\theta}]-\theta$ is a constant and the fourth by
$\mathrm{E}_P[\hat{\theta}-\mathrm{E}_P[\hat{\theta}]]=\mathrm{E}_P[\hat{\theta}]-\mathrm{E}_P[\hat{\theta}]=0$. □

Bias, Variance, and Accuracy

- Note that expected values and variances refer to the sampling distribution of an estimator, i.e., its distribution over repeated random sampling from the population of interest. Specifically, these quantities are computed with respect to the unknown data generating/population distribution P.

Bias, Variance, and Accuracy

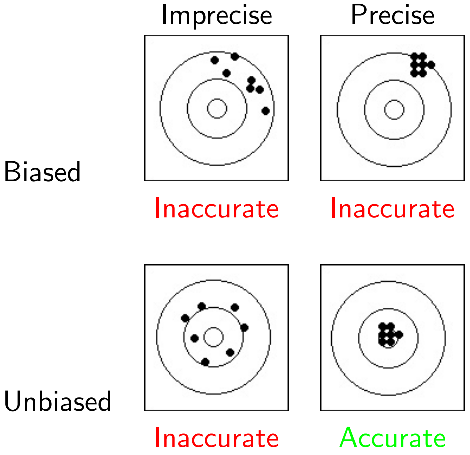

Figure 4: *Bias, variance, and accuracy.*

Bias-Variance Trade-Off

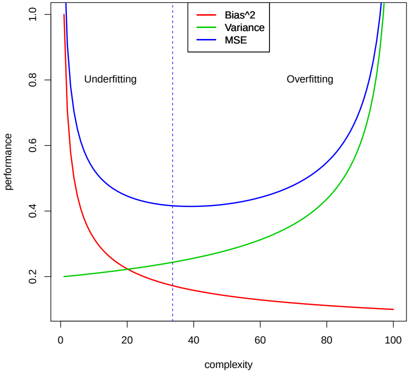

Figure 5: *Bias-variance trade-off.* Schematic representation of bias-variance trade-off as a function of model complexity.

Bias-Variance Trade-Off

Table 1: *Bias-variance trade-off.* Effect of model complexity and of sample size on bias and variance.

	Bias	Variance
Complexity $\uparrow$	$\downarrow$	$\uparrow$
Sample size $\uparrow$	?	$\downarrow$

Bias-Variance Trade-Off

- Figure 4 is a cartoon illustration of the notions of bias, variance/precision, and accuracy. While helpful, it does not illustrate the bias-variance trade-off related to model complexity.
- Figure 5 illustrates the bias-variance trade-off as it relates to model complexity. This figure is also an idealized representation of this phenomenon.
  - The term "complexity" is vague and needs to be precisely defined. Complexity means different things depending on the type of model/estimator, e.g., polynomial degree for linear regression, smoother span for loess.
  - In practice, bias and variance can be on very different scales.
  - In practice, the decay/increase of bias/variance with complexity is not always smooth.

Regression Example: Model Complexity

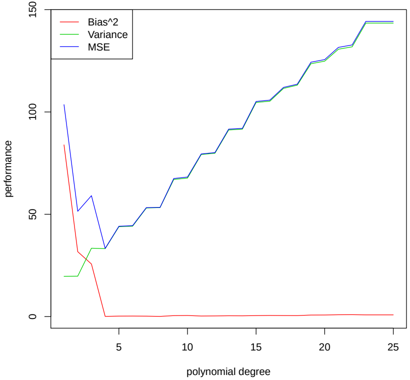

Figure 6: *Bias-variance trade-off: Linear regression.* Bias, variance, and MSE for linear regression fits for polynomials of degree 0 to 24.

Regression Example: Model Complexity

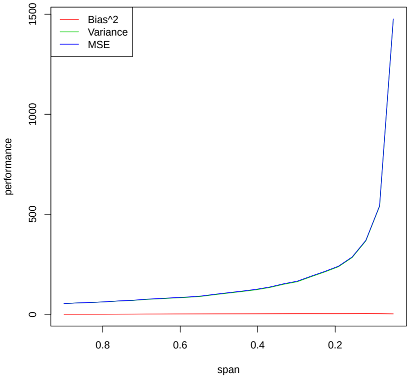

Figure 7: *Bias-variance trade-off: Robust local regression.* Bias, variance, and MSE for loess fits for spans ranging from 0.05 to 0.90.

Regression Example: Sample Size

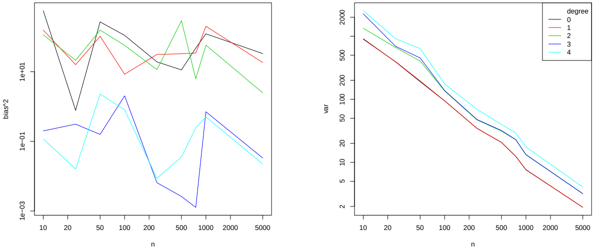

Figure 8: *Effect of sample size on bias and variance: Linear regression.* Bias and variance for linear regression fits vs. sample size n, for polynomials of degree 0 to 4.

Regression Example: Sample Size

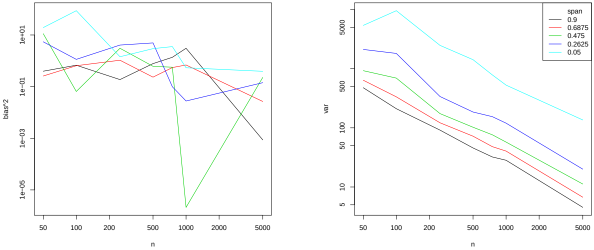

Figure 9: *Effect of sample size on bias and variance: Robust local regression.* Bias and variance for loess fits vs. sample size n, for spans ranging from 0.05 to 0.90.

Regression Example: True Regression Function

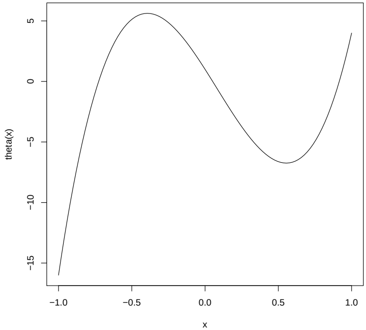

Figure 10: *Regression.* True regression function $\theta(x) = \mathrm{E}_P[Y|X=x] = 1 - 19x - 7x^2 + 29x^3$. $\mathrm{Var}_P[Y|X] = \sigma^2 = 100^2$. $X \sim \mathrm{N}(0,1)$.

Regression

- In the context of regression, the data structure is (X, Y), where $X \in \mathbb{R}^J$ is a J-dimensional column vector of covariates and $Y \in \mathbb{R}$ a scalar outcome.
- The parameter of interest is the regression function, i.e., the conditional expected value $\theta(X) \equiv \mathrm{E}_P[Y|X]$ of the outcome given the covariates.
- A natural loss function is the squared error or L_2 loss function

$$L_2((X, Y), \theta) \equiv (Y - \theta(X))^2. \tag{2}$$

Regression

- The population regression function (an unknown parameter) minimizes risk, i.e., MSE, computed with respect to the unknown population distribution P,

$$\theta(X) \equiv \mathrm{E}_P[Y|X] = \mathrm{argmin}_{\theta' \in \Theta} \mathrm{E}_P[(Y - \theta'(X))^2], \quad (3)$$

  where no restrictions are placed on the parameter space Θ for θ. That is, θ could be any function from $\mathbb{R}^J$ to $\mathbb{R}$.
- In practice, when seeking to estimate θ, one does not have access to the population distribution P, but only to the empirical distribution P_n corresponding to a random sample drawn from that population, i.e., a learning set, $\mathcal{L}_n = \{(X_i, Y_i) : i = 1, \ldots, n\}$.

Regression

- It is them customary to estimate the regression function θ by minimizing the empirical risk over a subset of the parameter space, $\hat{\Theta} \subseteq \Theta$,

$$\hat{\theta}_n(X) \equiv \text{argmin}_{\theta' \in \hat{\Theta}} \mathrm{E}_{P_n}[(Y - \theta'(X))^2]. \tag{4}$$

- Subsets $\hat{\Theta}$ of the parameter space Θ correspond to models for the regression function.
- As seen in a previous lecture, one popular model is the linear regression model,

$$\mathrm{E}[Y|X] = X^\top \beta = \sum_{j=1}^{J} \beta_j X_j = \beta_1 X_1 + \ldots + \beta_J X_J, \tag{5}$$

Regression

where the column vector $\beta = (\beta_j : j = 1, \ldots, J) \in \mathbb{R}^J$ contains the parameters of the model, referred to as regression coefficients.

- The least squares estimator (LSE) of the regression coefficients β is a solution to the normal equations

$$\mathbf{X}_n^\top \mathbf{Y}_n = \mathbf{X}_n^\top \mathbf{X}_n \beta, \tag{6}$$

where $\mathbf{X}_n$ is the $n \times J$ design matrix or model matrix, with ith row corresponding to the ith covariate vector X_i, and $\mathbf{Y}_n$ is the n-dimensional column outcome vector, with ith element corresponding to the ith outcome Y_i, $i = 1, \ldots, n$.

- When the design matrix is of full column rank, i.e., $\mathbf{X}_n^\top \mathbf{X}_n$ is invertible, the normal equations have a unique solution

$$\hat{\beta}_n = (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{X}_n^\top \mathbf{Y}_n. \tag{7}$$

Regression

- The regression function $\theta(x_0)$, evaluated at a particular covariate value x_0, can be estimated by $\hat{\theta}_n(x_0) = x_0^\top \hat{\beta}_n$.
- Note that all inference is conditional on the covariates, i.e., as if the design matrix $\mathbf{X}_n$ were fixed.

Regression: Estimating the Regression Function

- Suppose one is interested in estimating the conditional expected value $\theta(x_0) = \mathrm{E}_P[Y|X = x_0]$ of an outcome given the covariate value $X = x_0$.
- Let $\hat{\theta}_n(x_0)$ denote a particular estimator of $\theta(x_0)$, e.g., from LSE for a linear regression model that is quadratic in X or from loess with span 0.5.

Regression: Estimating the Regression Function

- Then, the risk for the squared error loss function, i.e., MSE, is the sum of the variance and of the square of the bias of $\hat{\theta}_n(x_0)$,

$$
\begin{aligned}
& \mathrm{E}_P[(\hat{\theta}_n(x_0) - \theta(x_0))^2 | \mathbf{X}_n] \qquad (8) \\
&= \mathrm{E}_P\left[\left(\hat{\theta}_n(x_0) - \mathrm{E}_P[\hat{\theta}_n(x_0) | \mathbf{X}_n] \right)^2 | \mathbf{X}_n \right] \\
&\quad + \left(\mathrm{E}_P[\hat{\theta}_n(x_0) | \mathbf{X}_n] - \theta(x_0) \right)^2 \\
&= \mathrm{Var}_P[\hat{\theta}_n(x_0) | \mathbf{X}_n] + (\mathrm{Bias}_P[\hat{\theta}_n(x_0), \theta(x_0) | \mathbf{X}_n])^2.
\end{aligned}
$$

In short,

$$
\mathrm{MSE} = \text{Variance of } \hat{\theta} + (\text{Bias of } \hat{\theta})^2.
$$

Regression: Estimating the Regression Function

Proof.

$$
\begin{aligned}
&\mathrm{E}_P[(\hat{\theta}_n(x_0) - \theta(x_0))^2|\mathbf{X}_n] \\
&\quad = \mathrm{E}_P\Big[\big(\hat{\theta}_n(x_0) - \mathrm{E}_P[\hat{\theta}_n(x_0)|\mathbf{X}_n] \\
&\quad\quad + \mathrm{E}_P[\hat{\theta}_n(x_0)|\mathbf{X}_n] - \theta(x_0)\big)^2|\mathbf{X}_n\Big] \\
&\quad = \mathrm{E}_P\Big[\big(\hat{\theta}_n(x_0) - \mathrm{E}_P[\hat{\theta}_n(x_0)|\mathbf{X}_n]\big)^2|\mathbf{X}_n\Big] \\
&\quad\quad + \mathrm{E}_P\Big[\big(\mathrm{E}_P[\hat{\theta}_n(x_0)|\mathbf{X}_n] - \theta(x_0)\big)^2|\mathbf{X}_n\Big] \\
&\quad\quad + 2\,\mathrm{E}_P\Big[\big(\hat{\theta}_n(x_0) - \mathrm{E}_P[\hat{\theta}_n(x_0)|\mathbf{X}_n]\big)\big(\mathrm{E}_P[\hat{\theta}_n(x_0)|\mathbf{X}_n] - \theta(x_0 \\
&\quad = \mathrm{Var}_P[\hat{\theta}_n(x_0)|\mathbf{X}_n] + (\mathrm{Bias}_P[\hat{\theta}_n(x_0), \theta(x_0)|\mathbf{X}_n])^2 \\
&\quad\quad + 2\big(\mathrm{E}_P[\hat{\theta}_n(x_0)|\mathbf{X}_n] - \theta(x_0)\big)\,\mathrm{E}_P\Big[\big(\hat{\theta}_n(x_0) - \mathrm{E}_P[\hat{\theta}_n(x_0)|\mathbf{X}_n \\
&\quad = \mathrm{Var}_P[\hat{\theta}_n(x_0)|\mathbf{X}_n] + (\mathrm{Bias}_P[\hat{\theta}_n(x_0), \theta(x_0)|\mathbf{X}_n])^2,
\end{aligned}
$$

where the third equality follows by noting that

Regression: Estimating the Regression Function

- Note that Equation (8) extends the simpler result of Equation (1) to the case where the parameter of interest is a regression function and one has to be mindful of conditioning on the covariates.

Regression: Predicting an Outcome

- Now suppose one is interested in predicting the actual value of an outcome Y for which the covariates are $X = x_0$ and where (X, Y) are independent from the learning set $\mathcal{L}_n$.
- A natural predictor is the estimator for the conditional expected value of Y given $X = x_0$, i.e., $\hat{\theta}_n(x_0)$.
- Then, the risk for the squared error loss function, i.e., MSE, is the sum of the variance of the outcome given the covariates and of the variance and square of the bias of $\hat{\theta}_n(x_0)$,

$$\begin{aligned} & \mathrm{E}_P[(Y - \hat{\theta}_n(x_0))^2 | \mathbf{X}_n, X = x_0] \qquad (9) \\ & = \mathrm{Var}_P[Y|X = x_0] + \mathrm{Var}_P[\hat{\theta}_n(x_0)|\mathbf{X}_n] + \mathrm{Bias}_P[\hat{\theta}_n(x_0), \theta(\end{aligned}$$

Risk Optimization and Bias-Variance Trade-Off

Dudoit

Bias-Variance Trade-Off

Regression: MSE and MSPE

In short,

$$\text{MSE} = \text{Variance of } Y + \text{Variance of } \hat{\theta} + (\text{Bias of } \hat{\theta})^2.$$

Proof.

$$\begin{aligned}
&\mathrm{E}_P[(Y - \hat{\theta}_n(x_0))^2 | \mathbf{X}_n, X = x_0] \\
&\quad = \mathrm{E}_P[(Y - \theta(x_0) + \theta(x_0) - \hat{\theta}_n(x_0))^2 | \mathbf{X}_n, X = x_0] \\
&\quad = \mathrm{E}_P[(Y - \theta(x_0))^2 | \mathbf{X}_n, X = x_0] \\
&\quad\ \ + \mathrm{E}_P[(\theta(x_0) - \hat{\theta}_n(x_0))^2 | \mathbf{X}_n, X = x_0] \\
&\quad\ \ + 2\,\mathrm{E}_P[(Y - \theta(x_0))(\theta(x_0) - \hat{\theta}_n(x_0)) | \mathbf{X}_n, X = x_0] \\
&\quad = \mathrm{Var}_P[Y | X = x_0] + \mathrm{E}_P[(\theta(x_0) - \hat{\theta}_n(x_0))^2 | \mathbf{X}_n] \\
&\quad\ \ + 2\,\mathrm{E}_P[(Y - \theta(x_0)) | X = x_0]\,\mathrm{E}_P[(\theta(x_0) - \hat{\theta}_n(x_0)) | \mathbf{X}_n] \\
&\quad = \mathrm{Var}_P[Y | X = x_0] + \mathrm{Var}_P[\hat{\theta}_n(x_0) | \mathbf{X}_n] + (\mathrm{Bias}_P[\hat{\theta}_n(x_0), \theta
\end{aligned}$$

where the third equality follows by independence of (X, Y) from the learning set and the fourth from Equation (8) and the fact that $E_P[(Y - \theta(x_0))|X = x_0] = 0$. $\square$

- Note that although $\hat{\theta}_n(x_0)$ is used both as an estimator of $\theta(x_0)$ and as a predictor of Y given $X = x_0$, the MSE is different.
- When estimating $\theta(x_0) = E_P[Y|X = x_0]$, the MSE compares $\hat{\theta}_n(x_0)$ to $\theta(x_0)$ and is defined as

$$E_P[(\hat{\theta}_n(x_0) - \theta(x_0))^2|\mathbf{X}_n].$$

Regression: Predicting an Outcome

Risk Optimization and Bias-Variance Trade-Off

Dudoit

Bias-Variance Trade-Off

Regression: MSE and MSPE

- When predicting an outcome value Y given covariates $X = x_0$, the MSE, sometimes referred to as mean squared prediction error (MSPE), compares $\hat{\theta}_n(x_0)$ to Y and is defined as

$$\mathrm{E}_P[(Y - \hat{\theta}_n(x_0))^2|\mathbf{X}_n, X = x_0],$$

  to account for the variance of the outcome given the covariates $\mathrm{Var}[Y|X = x_0]$, i.e., the additional variation of the outcome around its expected value $\theta(x_0)$.
- A common assumption is that of constant variance, $\mathrm{Var}[Y|X] = \sigma^2$.