



Linear Regression

Machine Learning – CSE546

Kevin Jamieson

University of Washington

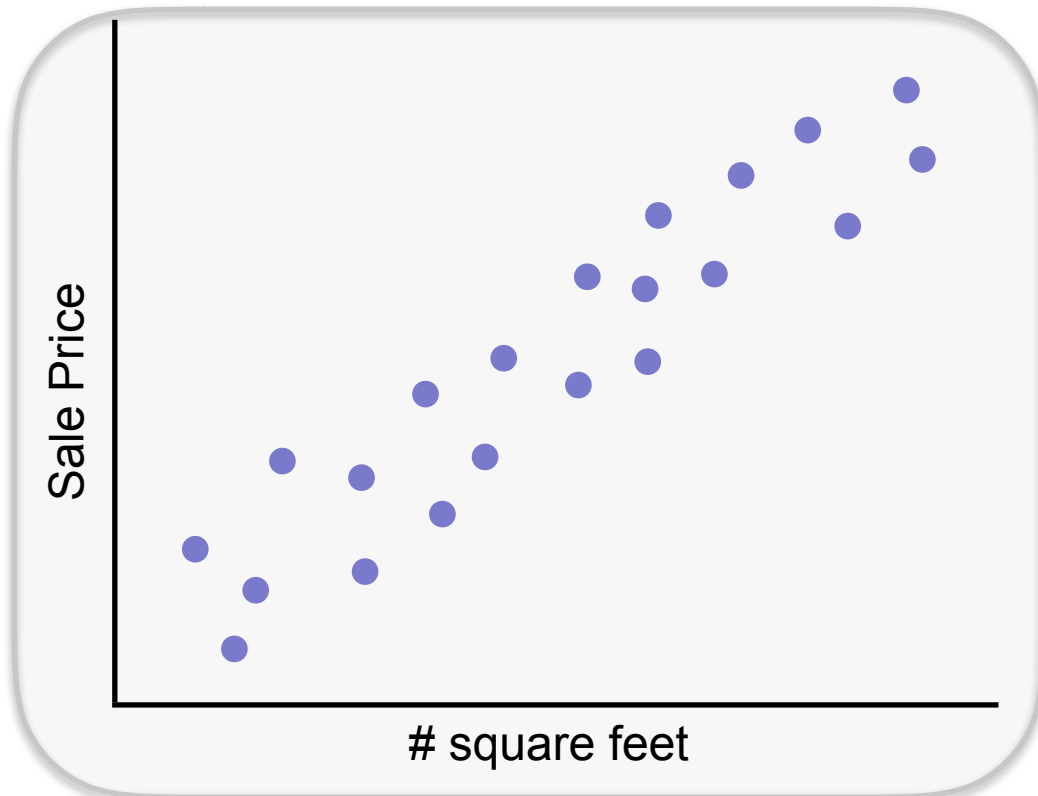
Oct 2, 2018

The regression problem

Given past sales data on [zillow.com](https://www.zillow.com), predict:

$y =$ **House sale price** *from*

$x =$ **{# sq. ft., zip code, date of sale, etc.}**



Training Data:

$$\{(x_i, y_i)\}_{i=1}^n$$

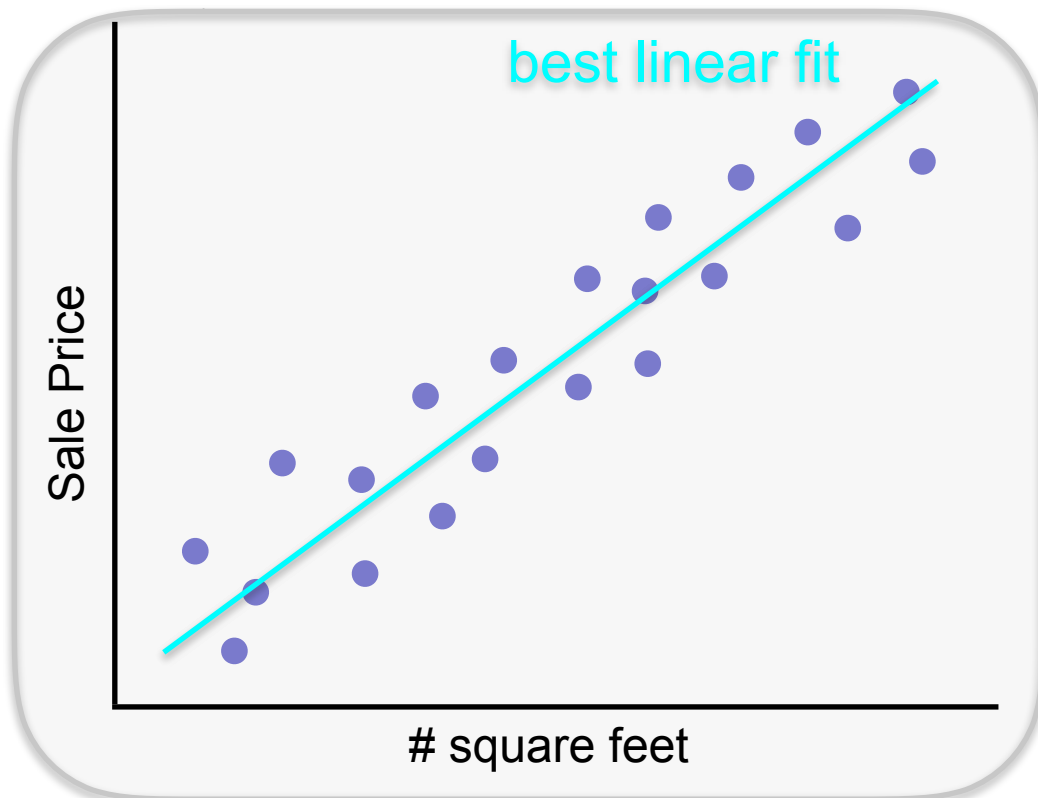
$$x_i \in \mathbb{R}^d$$
$$y_i \in \mathbb{R}$$

The regression problem

Given past sales data on [zillow.com](https://www.zillow.com), predict:

$y =$ House sale price *from*

$x = \{\# \text{ sq. ft.}, \text{ zip code}, \text{ date of sale}, \text{ etc.}\}$



Training Data:

$$\{(x_i, y_i)\}_{i=1}^n$$

$$x_i \in \mathbb{R}^d$$
$$y_i \in \mathbb{R}$$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

The regression problem in matrix notation

$$\begin{aligned}\hat{w}_{LS} &= \arg \min_w \sum_{i=1}^n (y_i - x_i^T w)^2 \\ &= \arg \min_w (\mathbf{y} - \mathbf{X}w)^T (\mathbf{y} - \mathbf{X}w)\end{aligned}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

The regression problem in matrix notation

$$\begin{aligned}\hat{w}_{LS} &= \arg \min_w \|\mathbf{y} - \mathbf{X}w\|_2^2 \\ &= \arg \min_w (\mathbf{y} - \mathbf{X}w)^T (\mathbf{y} - \mathbf{X}w)\end{aligned}$$

The regression problem in matrix notation

$$\begin{aligned}\hat{w}_{LS} &= \arg \min_w \|\mathbf{y} - \mathbf{X}w\|_2^2 \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}\end{aligned}$$

What about an offset?

$$\begin{aligned}\hat{w}_{LS}, \hat{b}_{LS} &= \arg \min_{w,b} \sum_{i=1}^n (y_i - (x_i^T w + b))^2 \\ &= \arg \min_{w,b} \|\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)\|_2^2\end{aligned}$$

Dealing with an offset

$$\hat{w}_{LS}, \hat{b}_{LS} = \arg \min_{w, b} \|\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)\|_2^2$$

Dealing with an offset

$$\hat{w}_{LS}, \hat{b}_{LS} = \arg \min_{w, b} \|\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)\|_2^2$$

$$\mathbf{X}^T \mathbf{X} \hat{w}_{LS} + \hat{b}_{LS} \mathbf{X}^T \mathbf{1} = \mathbf{X}^T \mathbf{y}$$

$$\mathbf{1}^T \mathbf{X} \hat{w}_{LS} + \hat{b}_{LS} \mathbf{1}^T \mathbf{1} = \mathbf{1}^T \mathbf{y}$$

If $\mathbf{X}^T \mathbf{1} = 0$ (i.e., if each feature is mean-zero) then

$$\hat{w}_{LS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

$$\hat{b}_{LS} = \frac{1}{n} \sum_{i=1}^n y_i$$

The regression problem in matrix notation

$$\begin{aligned}\hat{w}_{LS} &= \arg \min_w \|\mathbf{y} - \mathbf{X}w\|_2^2 \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}\end{aligned}$$

But why least squares?

Consider $y_i = x_i^T w + \epsilon_i$ where $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$P(y|x, w, \sigma) =$$

Maximizing log-likelihood

Maximize:

$$\log P(\mathcal{D}|w, \sigma) = \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^n e^{-\frac{(y_i - x_i^T w)^2}{2\sigma^2}}$$

MLE is LS under linear model

$$\hat{w}_{LS} = \arg \min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

$$\hat{w}_{MLE} = \arg \max_w P(\mathcal{D}|w, \sigma)$$

$$\text{if } y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\hat{w}_{LS} = \hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Analysis of error

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Analysis of error

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\begin{aligned}\hat{w}_{MLE} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X}w + \epsilon) \\ &= w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon\end{aligned}$$

$$\text{Cov}(\hat{w}_{MLE}) = \mathbb{E}[(\hat{w} - \mathbb{E}[\hat{w}])(\hat{w} - \mathbb{E}[\hat{w}])^T] = (\mathbf{X}^T \mathbf{X})^{-1}$$

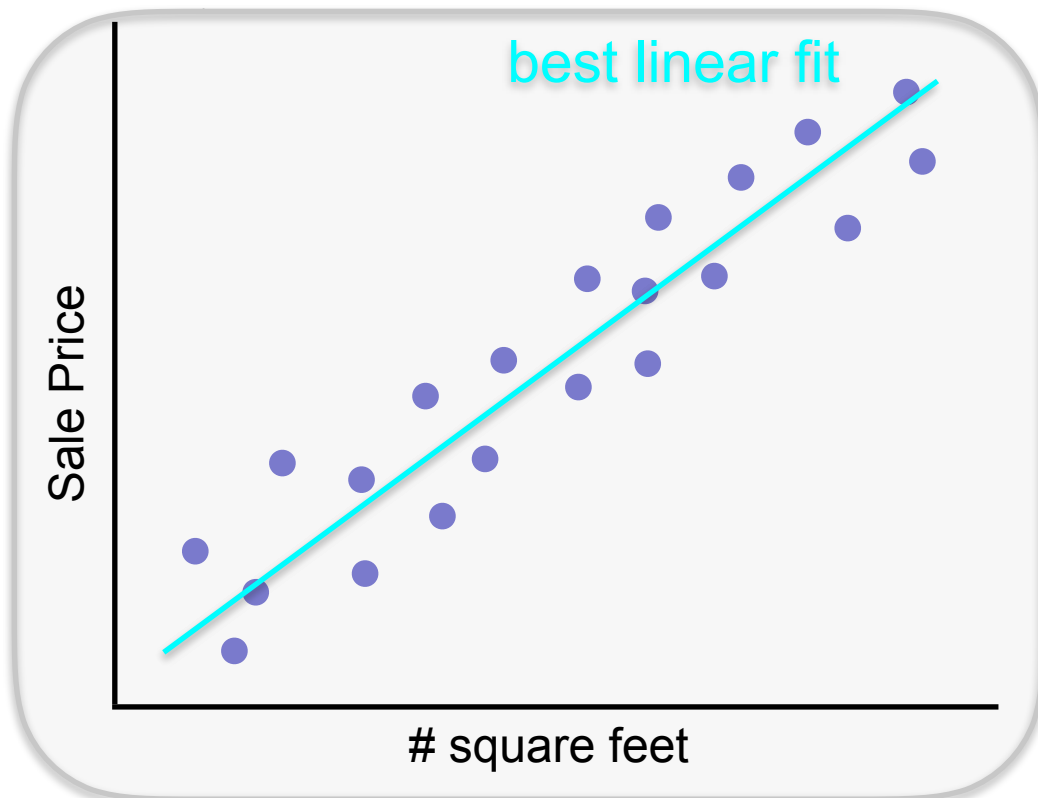
$$\hat{w}_{MLE} \sim \mathcal{N}(w, (\mathbf{X}^T \mathbf{X})^{-1})$$

The regression problem

Given past sales data on [zillow.com](https://www.zillow.com), predict:

$y =$ House sale price *from*

$x = \{\# \text{ sq. ft.}, \text{ zip code}, \text{ date of sale}, \text{ etc.}\}$



Training Data:

$$\{(x_i, y_i)\}_{i=1}^n$$

$$x_i \in \mathbb{R}^d$$
$$y_i \in \mathbb{R}$$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

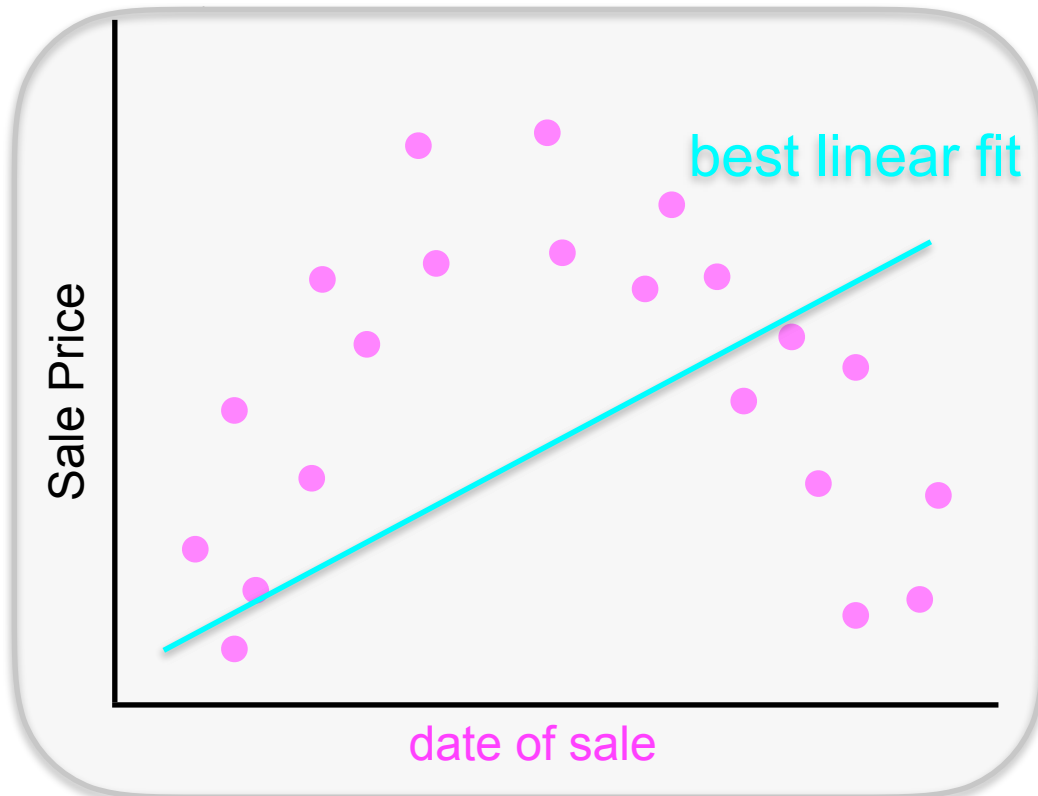
$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

The regression problem

Given past sales data on [zillow.com](https://www.zillow.com), predict:

$y =$ House sale price *from*

$x = \{\# \text{ sq. ft.}, \text{ zip code}, \text{ date of sale}, \text{ etc.}\}$



Training Data:

$$\{(x_i, y_i)\}_{i=1}^n$$

$$x_i \in \mathbb{R}^d$$
$$y_i \in \mathbb{R}$$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$

Transformed data:

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

Transformed data:

$h : \mathbb{R}^d \rightarrow \mathbb{R}^p$ maps original features to a rich, possibly high-dimensional space

$$\text{in } d=1: h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix} = \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^p \end{bmatrix}$$

for $d > 1$, generate $\{u_j\}_{j=1}^p \subset \mathbb{R}^d$

$$h_j(x) = \frac{1}{1 + \exp(u_j^T x)}$$

$$h_j(x) = (u_j^T x)^2$$

$$h_j(x) = \cos(u_j^T x)$$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear

$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - h(x_i)^T w)^2$$

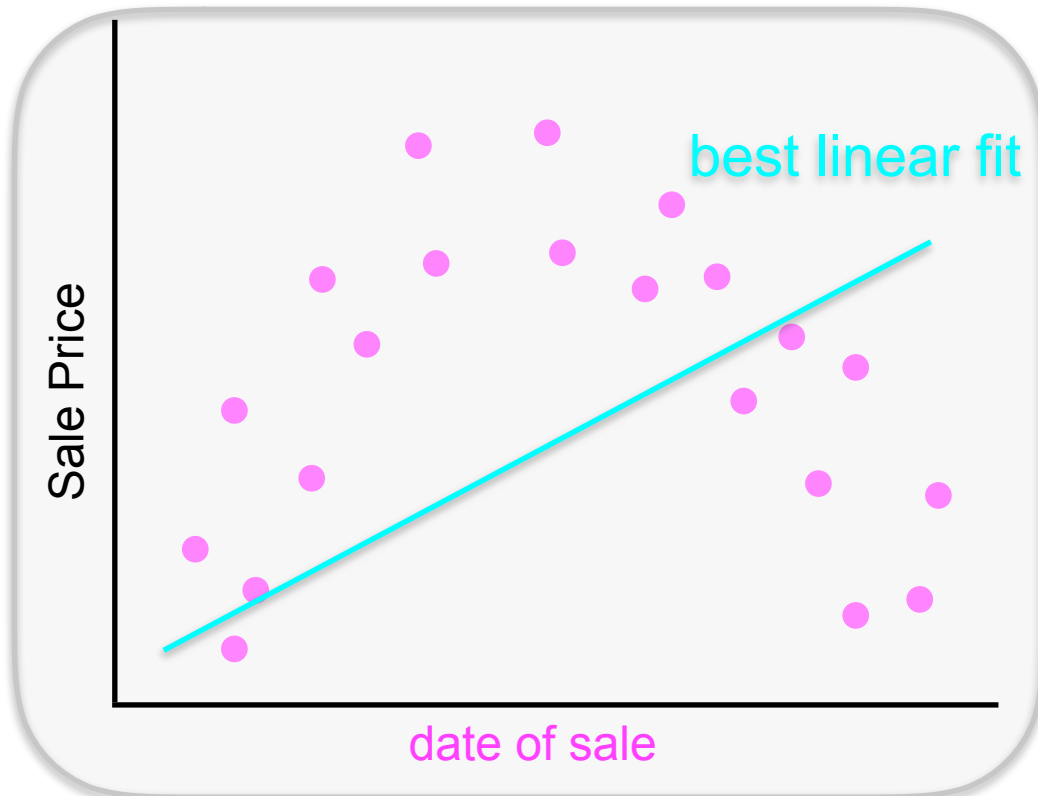
The regression problem

Training Data:

$$\{(x_i, y_i)\}_{i=1}^n \quad \begin{array}{l} x_i \in \mathbb{R}^d \\ y_i \in \mathbb{R} \end{array}$$

Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$



Hypothesis: linear

$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

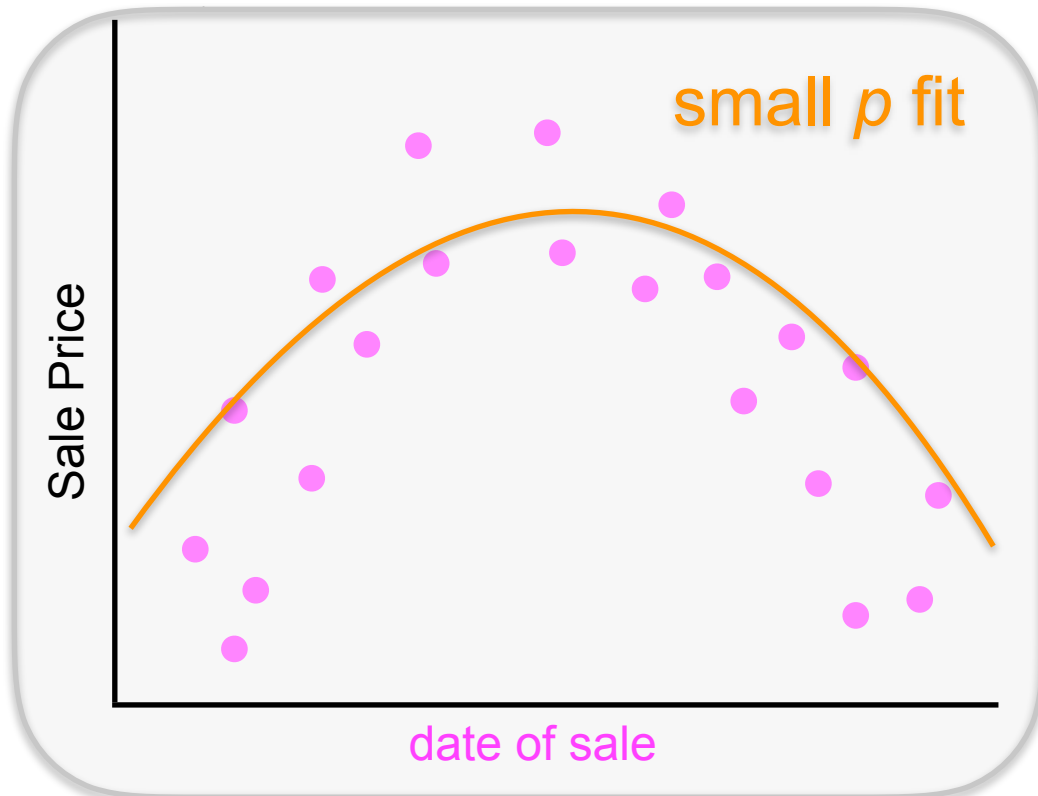
Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - h(x_i)^T w)^2$$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$
 $\{(x_i, y_i)\}_{i=1}^n$

Transformed data: $h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$



Hypothesis: linear

$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - h(x_i)^T w)^2$$

The regression problem

Training Data:

$$\{(x_i, y_i)\}_{i=1}^n$$

$x_i \in \mathbb{R}^d$
 $y_i \in \mathbb{R}$

Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$



Hypothesis: linear

$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

Loss: least squares

$$\min_w \sum_{i=1}^n (y_i - h(x_i)^T w)^2$$

What's going on here?



Bias-Variance Tradeoff

Machine Learning – CSE546

Kevin Jamieson

University of Washington

Oct 5, 2018

Statistical Learning

$$P_{XY}(X = x, Y = y)$$

Goal: Predict Y given X

Find function η that minimizes

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

Statistical Learning

$$P_{XY}(X = x, Y = y)$$

Goal: Predict Y given X

Find function η that minimizes

$$\mathbb{E}_{XY}[(Y - \eta(X))^2] = \mathbb{E}_X \left[\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x] \right]$$

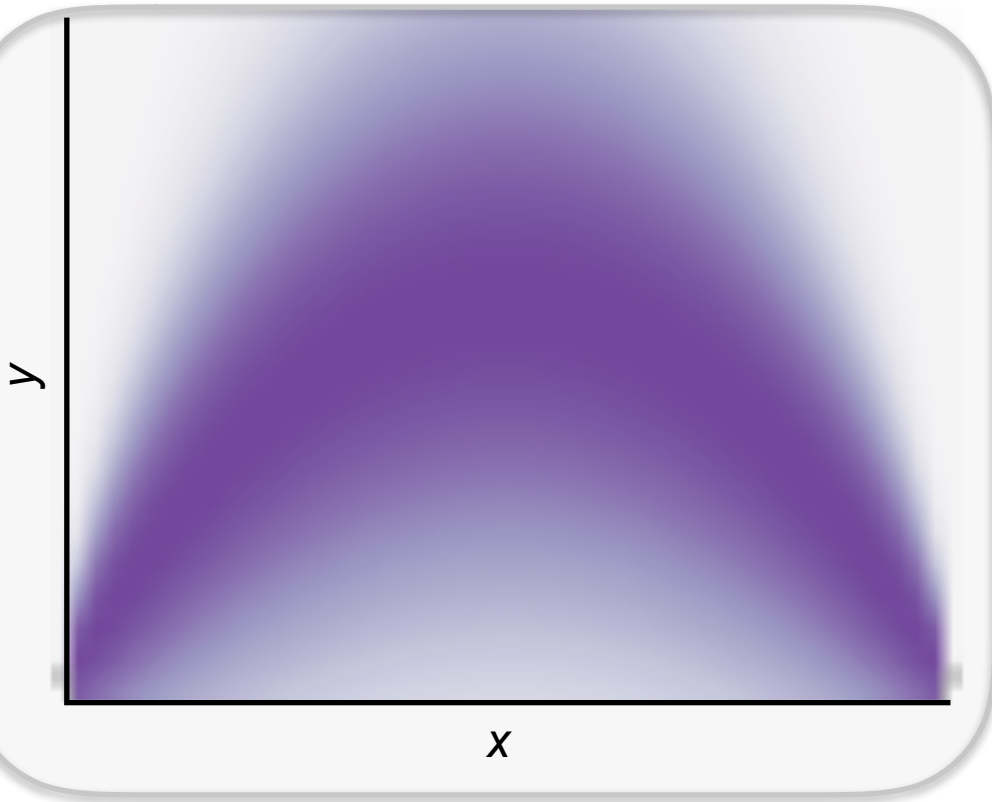
$$\eta(x) = \arg \min_c \mathbb{E}_{Y|X}[(Y - c)^2 | X = x] = \mathbb{E}_{Y|X}[Y | X = x]$$

Under LS loss, optimal predictor: $\eta(x) = \mathbb{E}_{Y|X}[Y | X = x]$

Statistical Learning

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

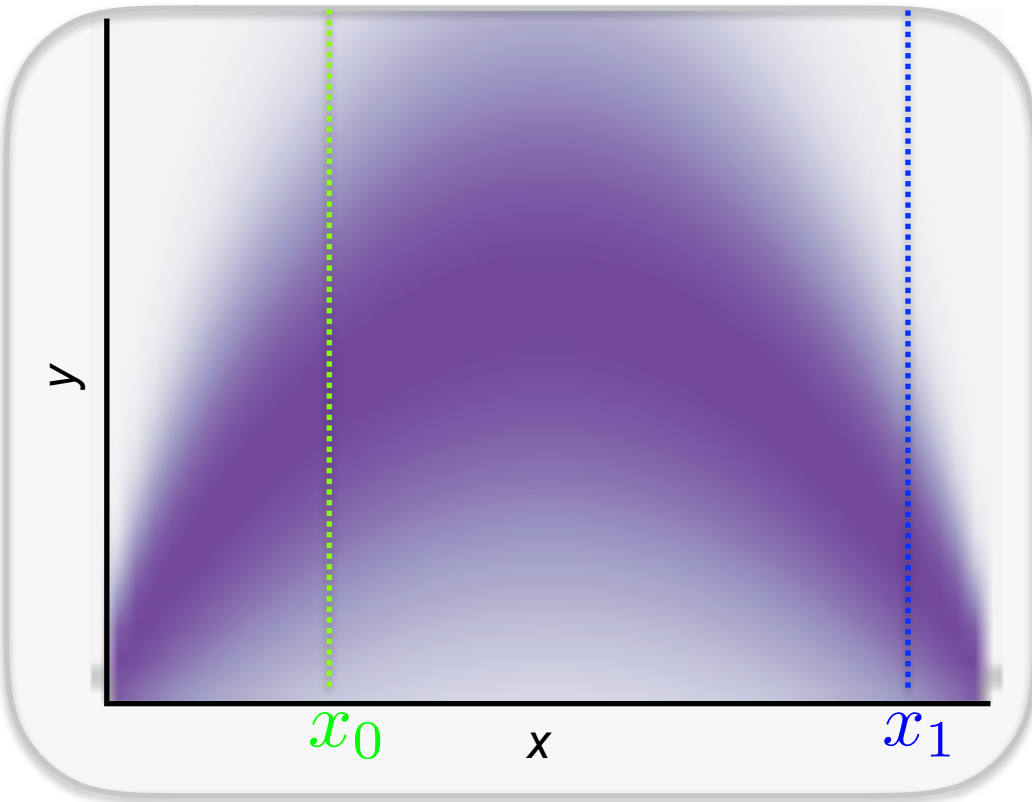
$$P_{XY}(X = x, Y = y)$$



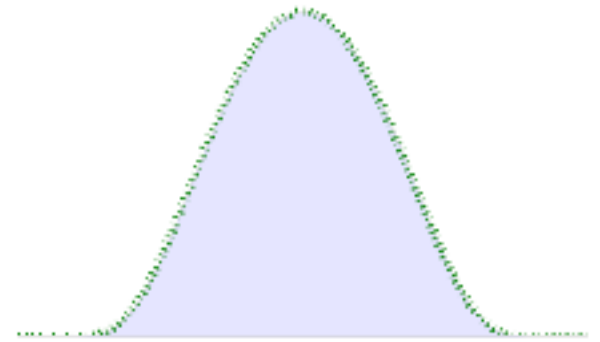
Statistical Learning

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

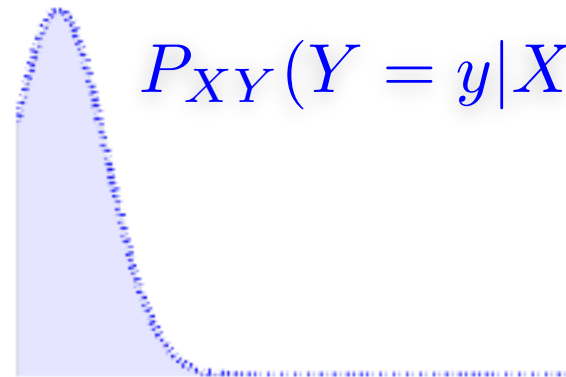
$$P_{XY}(X = x, Y = y)$$



$$P_{XY}(Y = y | X = x_0)$$



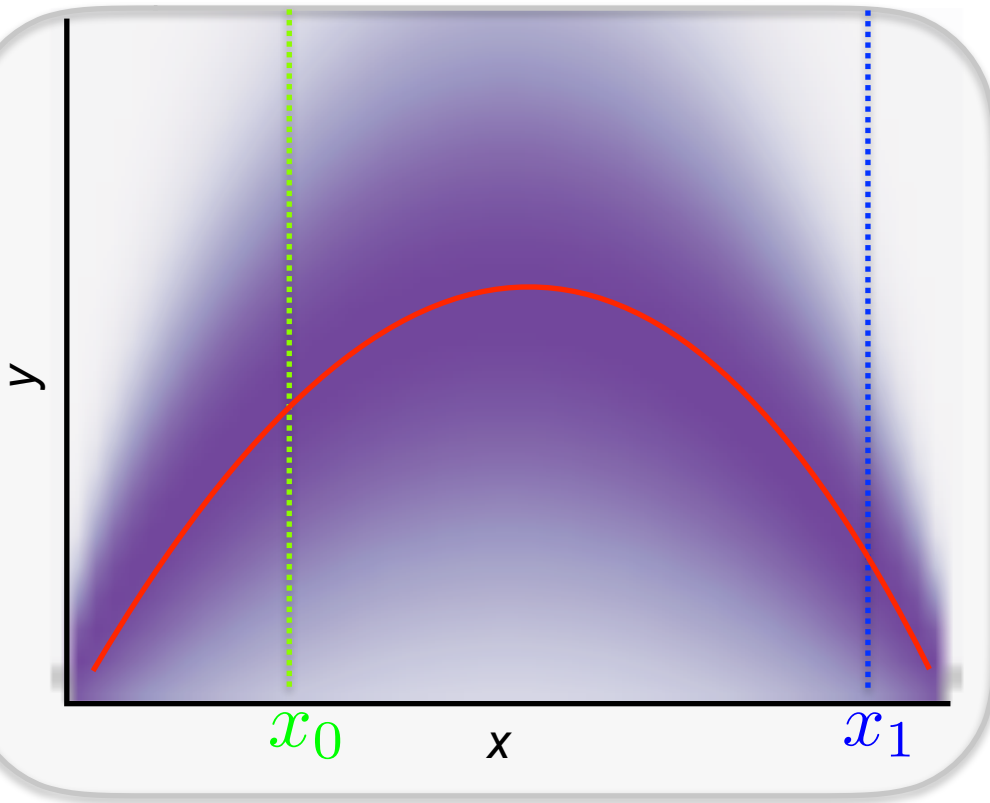
$$P_{XY}(Y = y | X = x_1)$$



Statistical Learning

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

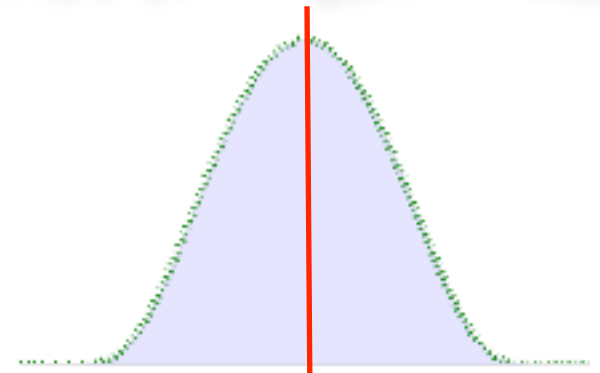
$$P_{XY}(X = x, Y = y)$$



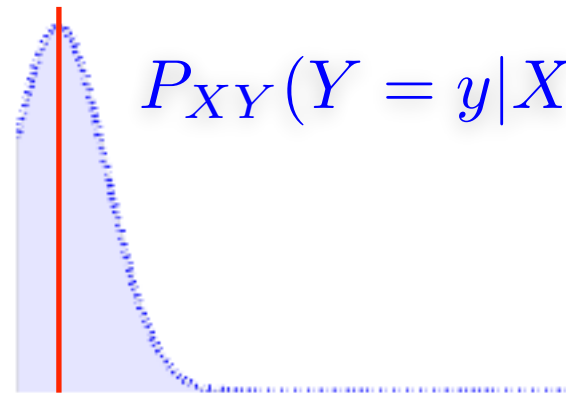
Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$P_{XY}(Y = y|X = x_0)$$



$$P_{XY}(Y = y|X = x_1)$$

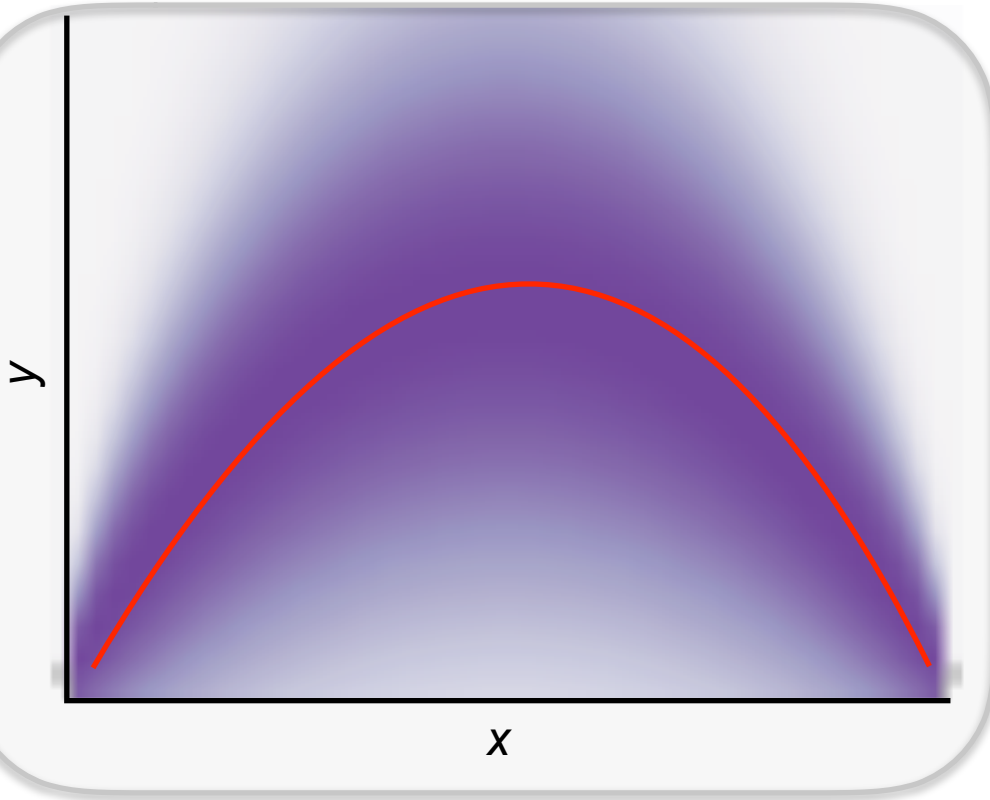


Statistical Learning

$$P_{XY}(X = x, Y = y)$$

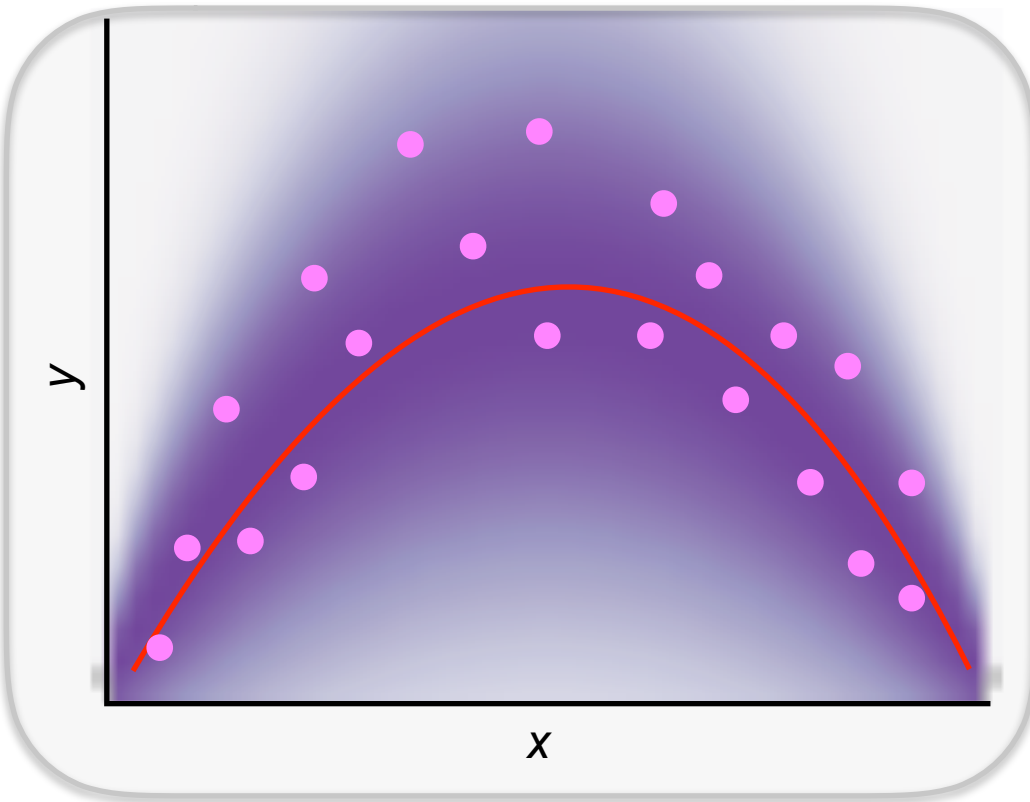
Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$



Statistical Learning

$$P_{XY}(X = x, Y = y)$$



Ideally, we want to find:

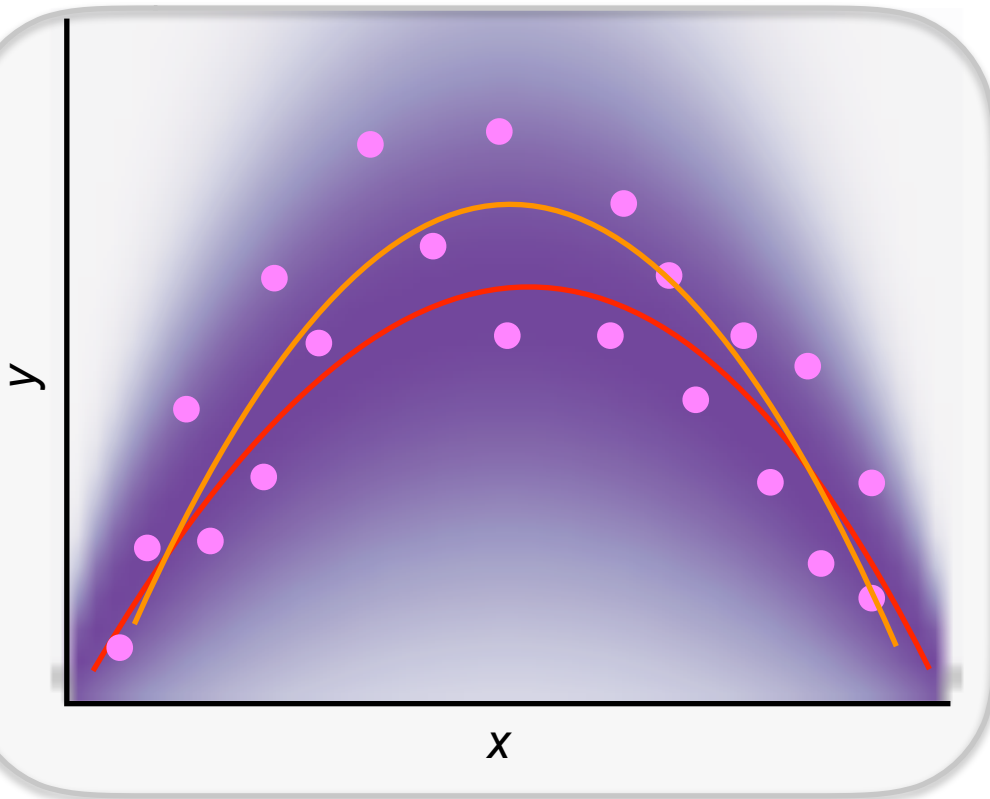
$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

$$(x_i, y_i) \stackrel{i.i.d.}{\sim} P_{XY} \quad \text{for } i = 1, \dots, n$$

Statistical Learning

$$P_{XY}(X = x, Y = y)$$



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

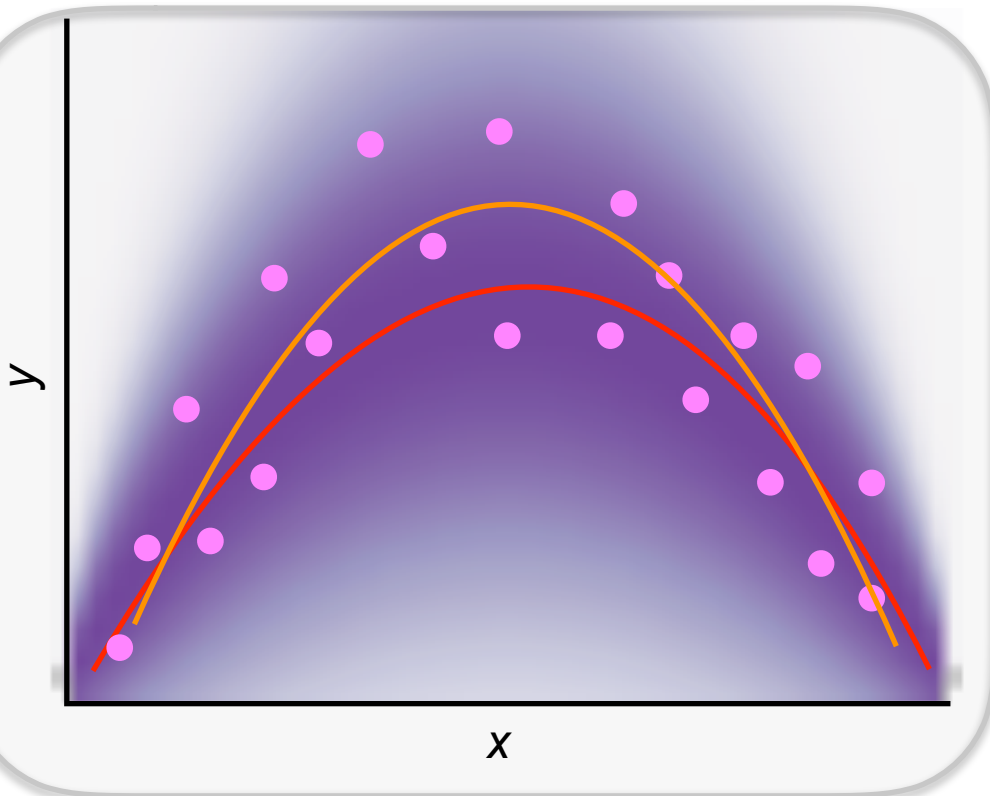
But we only have samples:
 $(x_i, y_i) \stackrel{i.i.d.}{\sim} P_{XY}$ for $i = 1, \dots, n$

and are restricted to a
function class (e.g., linear)
so we compute:

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

Statistical Learning

$$P_{XY}(X = x, Y = y)$$



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:
 $(x_i, y_i) \stackrel{i.i.d.}{\sim} P_{XY}$ for $i = 1, \dots, n$

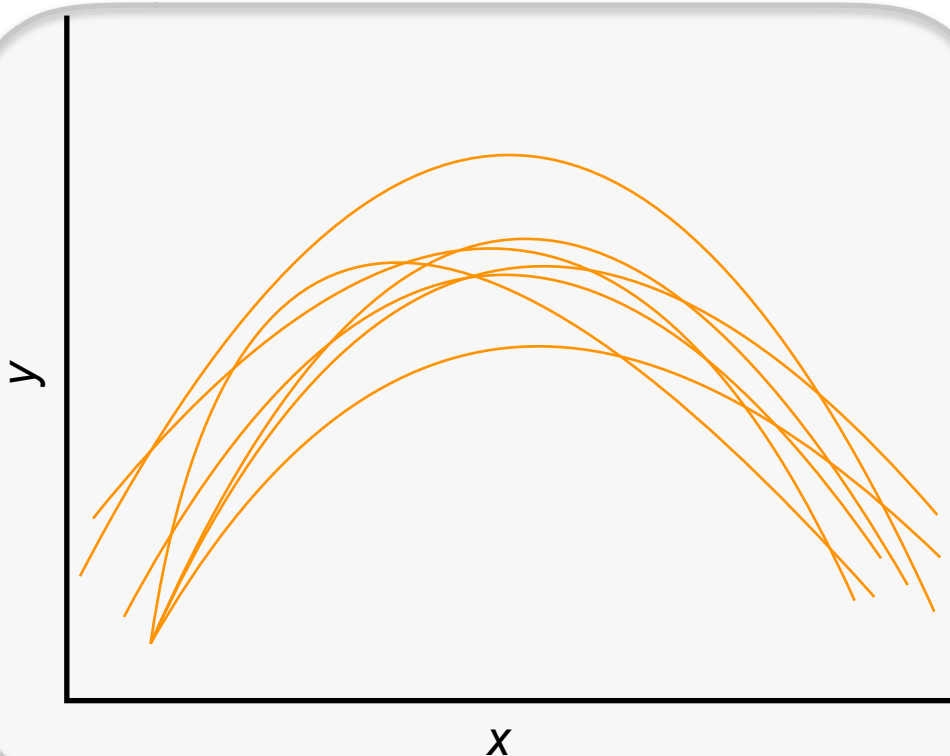
and are restricted to a function class (e.g., linear) so we compute:

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

We care about future predictions: $\mathbb{E}_{XY}[(Y - \hat{f}(X))^2]$

Statistical Learning

$$P_{XY}(X = x, Y = y)$$



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:
 $(x_i, y_i) \stackrel{i.i.d.}{\sim} P_{XY}$ for $i = 1, \dots, n$

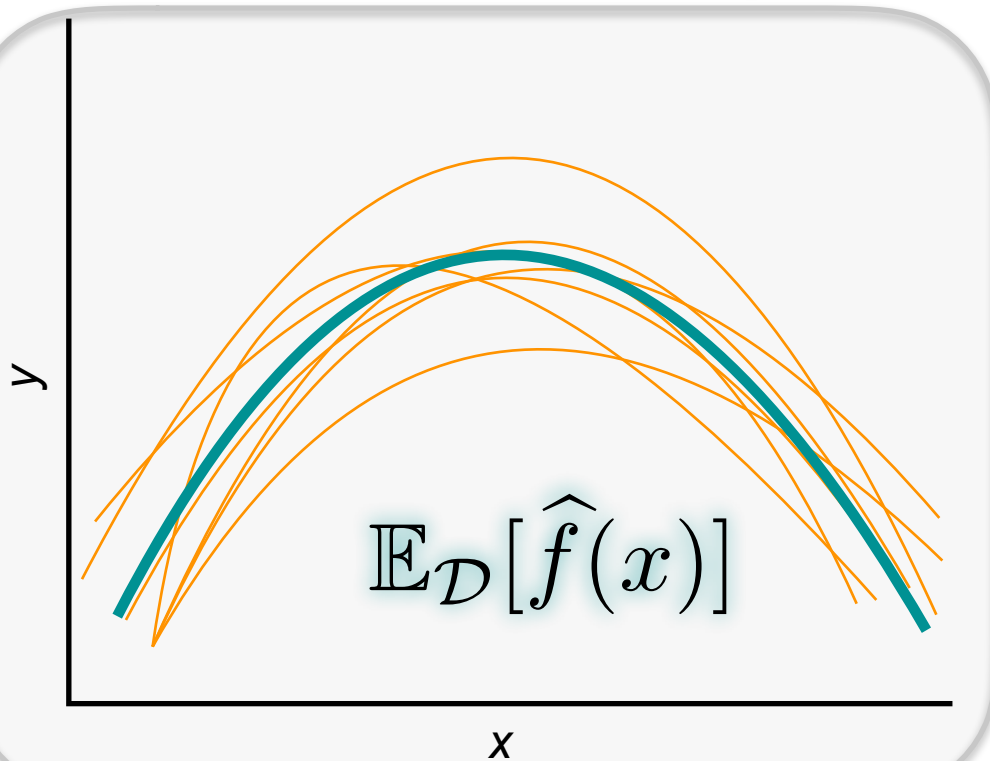
and are restricted to a function class (e.g., linear) so we compute:

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

Each draw $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ results in different \hat{f}

Statistical Learning

$$P_{XY}(X = x, Y = y)$$



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:
 $(x_i, y_i) \stackrel{i.i.d.}{\sim} P_{XY}$ for $i = 1, \dots, n$

and are restricted to a function class (e.g., linear) so we compute:

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

Each draw $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ results in different \hat{f}

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \qquad \hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \hat{f}_{\mathcal{D}}(x))^2]|X = x] = \mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \eta(x) + \eta(x) - \hat{f}_{\mathcal{D}}(x))^2]|X = x]$$

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \quad \hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

$$\begin{aligned} \mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \hat{f}_{\mathcal{D}}(x))^2]|X = x] &= \mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \eta(x) + \eta(x) - \hat{f}_{\mathcal{D}}(x))^2]|X = x] \\ &= \mathbb{E}_{Y|X} \left[\mathbb{E}_{\mathcal{D}}[(Y - \eta(x))^2 + 2(Y - \eta(x))(\eta(x) - \hat{f}_{\mathcal{D}}(x)) \right. \\ &\quad \left. + (\eta(x) - \hat{f}_{\mathcal{D}}(x))^2] | X = x \right] \\ &= \underbrace{\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x]}_{\text{irreducible error}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2]}_{\text{learning error}} \end{aligned}$$

irreducible error

Caused by stochastic
label noise

learning error

Caused by either using too “simple”
of a model or not enough
data to learn the model accurately

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

$$\underline{\mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2]} = \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] + \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]$$

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

$$\begin{aligned} \underline{\mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2]} &= \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] + \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] \\ &= \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 + 2(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)) \\ &\quad + (\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] \\ &= \underline{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2} + \underline{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]} \\ &\quad \text{biased squared} \qquad \qquad \qquad \text{variance} \end{aligned}$$

Bias-Variance Tradeoff

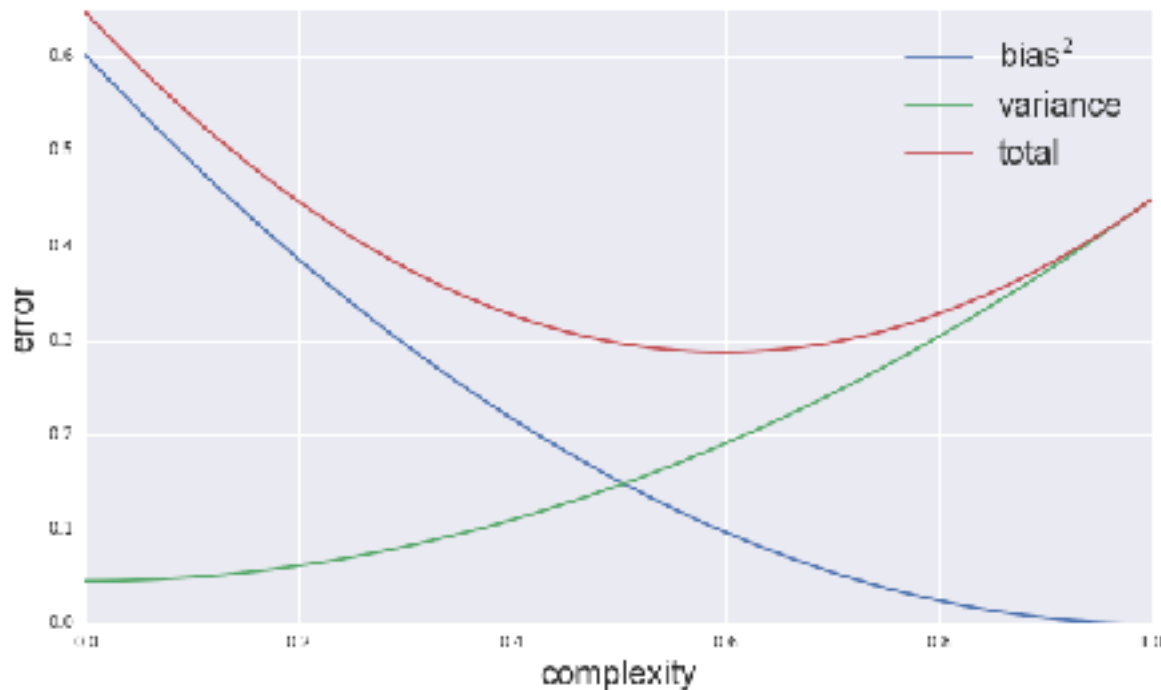
$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \hat{f}_{\mathcal{D}}(x))^2] | X = x] = \underbrace{\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x]}_{\text{irreducible error}}$$

irreducible error

$$+ \underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2}_{\text{biased squared}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]}_{\text{variance}}$$

biased squared

variance



Example: Linear LS $Y = \mathbf{X}w + \epsilon$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\eta(x) = \mathbb{E}_{Y|X} [Y | X = x]$$

$$\hat{f}_{\mathcal{D}}(x) =$$

Example: Linear LS $\mathbf{Y} = \mathbf{X}w + \epsilon$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\underbrace{\mathbb{E}_{XY}[(Y - \eta(x))^2 | X = x]}_{\text{irreducible error}} = \sigma^2 \qquad \underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2}_{\text{biased squared}} = 0$$

Example: Linear LS $\mathbf{Y} = \mathbf{X}w + \epsilon$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\underline{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]} =$$

variance

Example: Linear LS $\mathbf{Y} = \mathbf{X}w + \epsilon$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]}_{\text{variance}} = \mathbb{E}_{\mathcal{D}}[x^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x]$$

variance

$$= \sigma^2 x^T (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$= \sigma^2 \text{Trace}((\mathbf{X}^T \mathbf{X})^{-1} x x^T)$$

$$\mathbf{X}^T \mathbf{X} = \sum_{i=1}^n x_i x_i^T \xrightarrow{n \text{ large}} n \Sigma \quad \Sigma = \mathbb{E}[X X^T], \quad X \sim P_X$$

$$\mathbb{E}_{X=x}[\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]] = \frac{\sigma^2}{n} \mathbb{E}_X[\text{Trace}(\Sigma^{-1} X X^T)] = \frac{d\sigma^2}{n}$$

Example: Linear LS $Y = Xw + \epsilon$

if $y_i = x_i^T w + \epsilon_i$ and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = w + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon$$

$$\eta(x) = \mathbb{E}_{Y|X} [Y | X = x]$$

$$\hat{f}_{\mathcal{D}}(x) = \hat{w}^T x = w^T x + \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x$$

$$\underbrace{\mathbb{E}_{XY} [(Y - \eta(x))^2 | X = x]}_{\text{irreducible error}} = \sigma^2 \qquad \underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2}_{\text{biased squared}} = 0$$

irreducible error

biased squared

$$\mathbb{E}_{X=x} \left[\underbrace{\mathbb{E}_{\mathcal{D}} [(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]}_{\text{variance}} \right] = \frac{d\sigma^2}{n}$$

variance