# Linear Regression



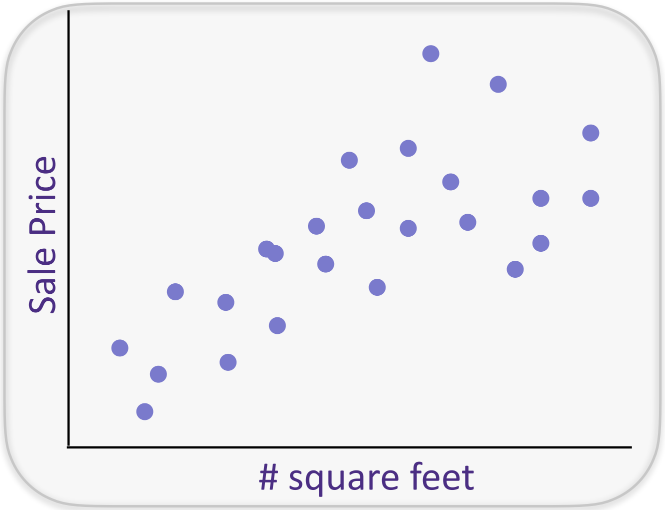
UNIVERSITY of WASHINGTON

# The regression problem, 1-dimensional

Given past sales data on zillow.com, predict:

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

$x$ = **{# sq. ft.}** → input




Training Data:

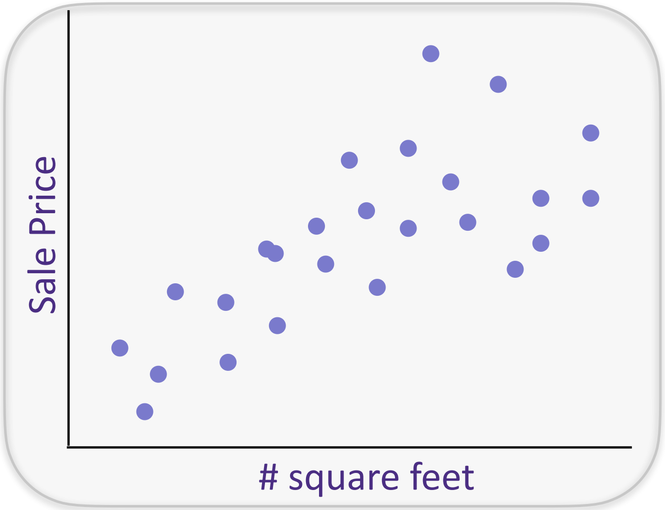
$\{(x_i, y_i)\}_{i=1}^n$ $\quad y_i \in \mathbb{R}$

# The regression problem, 1-dimensional

Given past sales data on zillow.com, predict:

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

$x$ = **{# sq. ft.}**




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

$x_i \in \mathbb{R}$

$y_i \in \mathbb{R}$

# Process

Decide on a **model**

Find the function which fits the data best

Use function to make prediction on new examples

# The Model

We ***assume*** house sale price is a linear function of square feet.

# Process

Decide on a **model**

Find the function which fits the data best

**Choose a loss function** Choice

**Pick the function which minimizes loss on data**

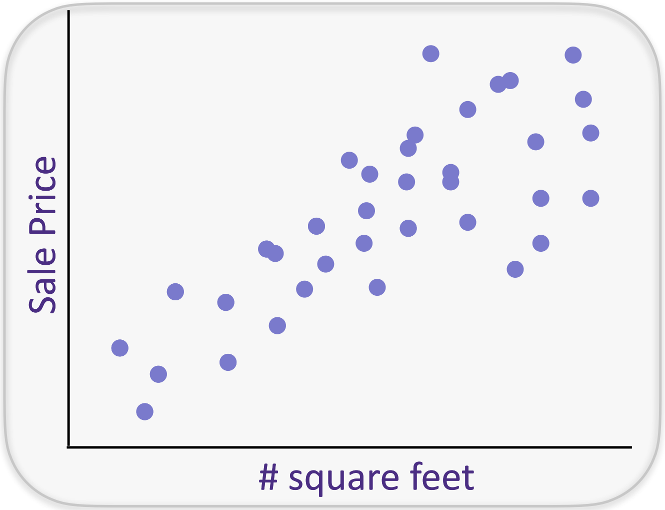
Use function to make prediction on new examples

# Fit a function to our data, 1-d

Given past sales data on zillow.com, predict:

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

*x* = **{# sq. ft.}**




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

Hypothesis/Model: linear

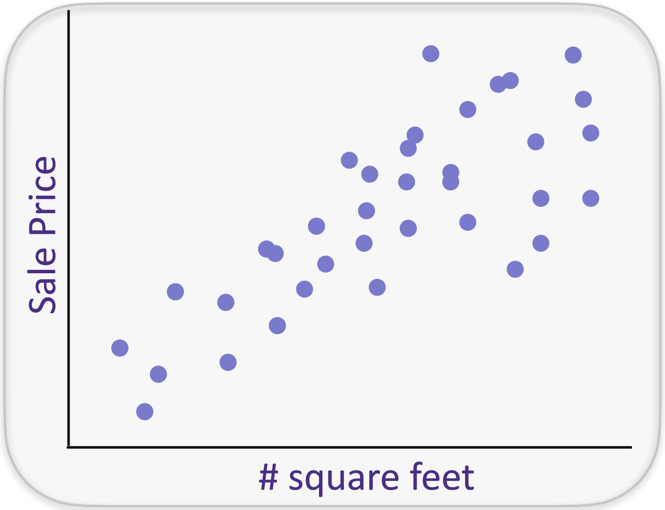
$$y_i \approx x_i w$$

# Fit a function to our data, 1-d

Given past sales data on zillow.com, predict:

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

***x* = {# sq. ft.}**




Training Data:

$\{(x_i, y_i)\}_{i=1}^n$ $\quad x_i \in \mathbb{R}$, $y_i \in \mathbb{R}$

Hypothesis/Model: linear

$$y_i \approx x_i w$$

Loss: $\boxed{\text{least squares}}$

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

Find function: ?

$$0 = \frac{d}{dw} \sum_{i=1}^n (y_i - x_i w)^2$$

$$0 = \sum 2(y_i - x_i w) \cdot$$

# Fit a function to our data, 1-d

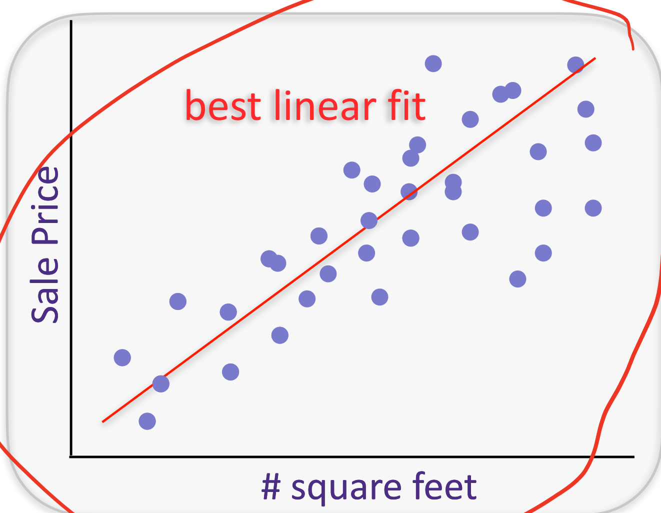
$- x_i$

$0 = \sum (y_i - x_i w) \cdot x_i$

Given past sales data on zillow.com, predict:

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

***x* = {# sq. ft.}**




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

Hypothesis/Model: linear

$$y_i \approx x_i w$$

Loss: least squares

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

Find function: ?

# Fit a function to our data, 1-d

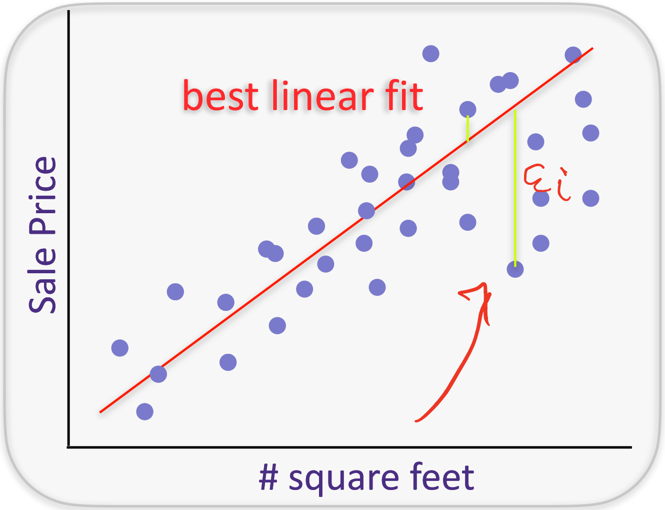
Given past sales data on zillow.com, predict:

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

***x* = {# sq. ft.}**

Error:

$$y_i = x_i w + \epsilon_i$$




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

Hypothesis/Model: linear

$$y_i \approx x_i w$$

Loss: least squares

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

Find function: ?

# Process

Decide on a **model**  Linear

Find the function which fits the data best

**Choose a loss function** → Least squares

**Pick the function which minimizes loss on data**

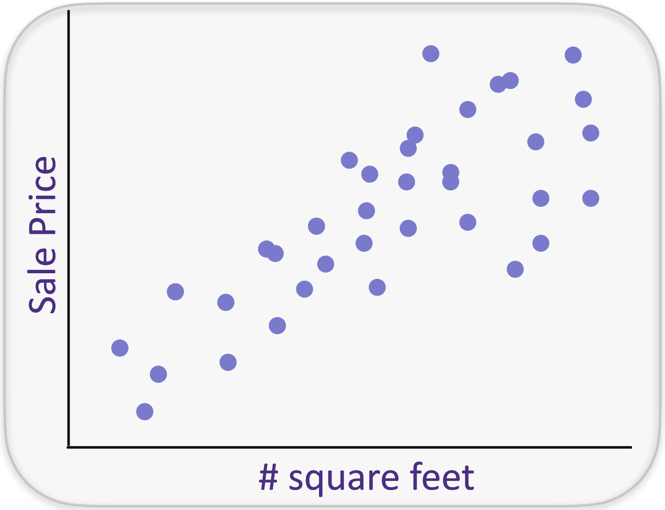
Use function to make prediction on new examples

# Make a Prediction

Given past sales data on zillow.com, predict:

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

$x$ = **{# sq. ft.}**




$$y_i \approx \hat{w}_{LS} x_i$$

$x_{new} \rightarrow$ sq ft.

$y_{new} \approx \hat{w}_{LS} \cdot x_{new}$

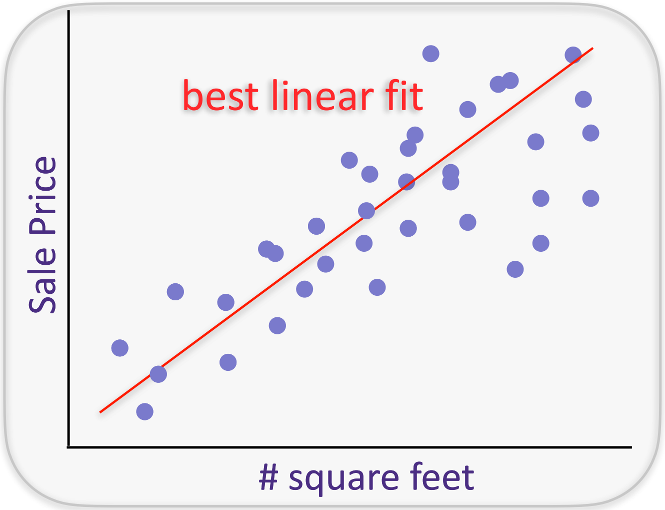
# Make a Prediction

Given past sales data on zillow.com, predict:

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

$x$ = **{# sq. ft.}**

$$y_i \approx \hat{w}_{LS} x_i$$



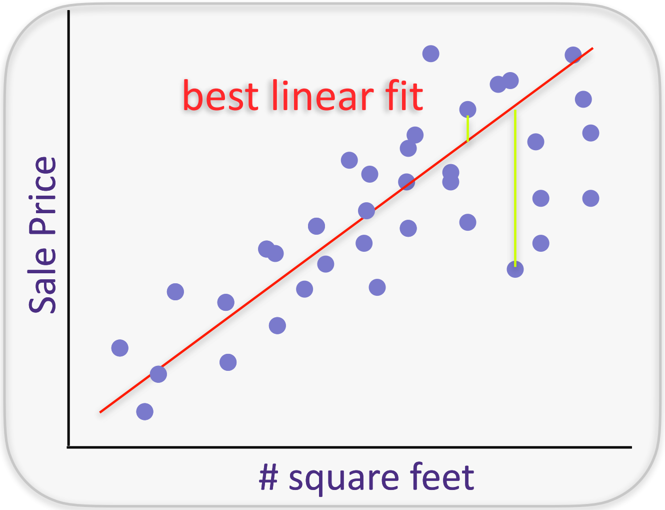
# Make a Prediction

Given past sales data on zillow.com, predict:

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

$x$ = **{# sq. ft.}**

$$y_i \approx \hat{w}_{LS} x_i$$



# Process

Decide on a **model**

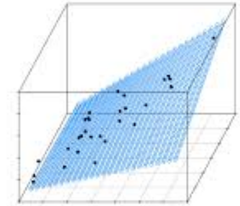
Find the function which fits the data best

**Choose a loss function**

**Pick the function which minimizes loss on data**

Use function to make prediction on new examples

# The regression problem, d-dim

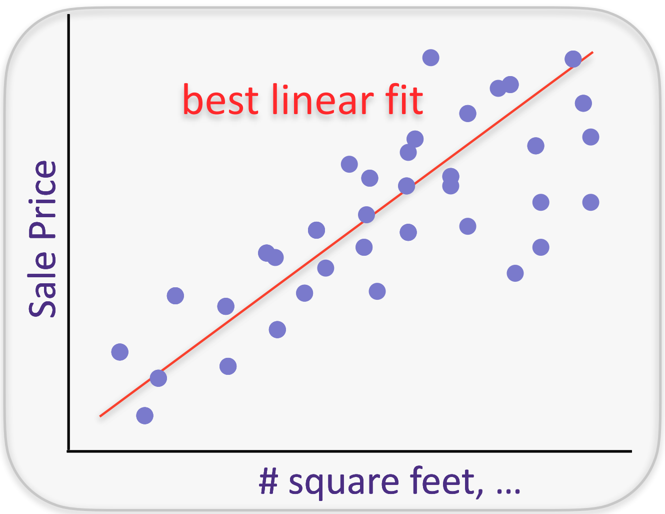


Given past sales data on zillow.com, predict:

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

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

$x_i = ( \quad )$




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

Hypothesis: linear

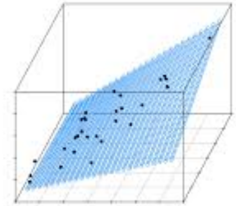
$$y_i \approx x_i^T w$$

$w = ( \quad )$

Loss: least squares

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

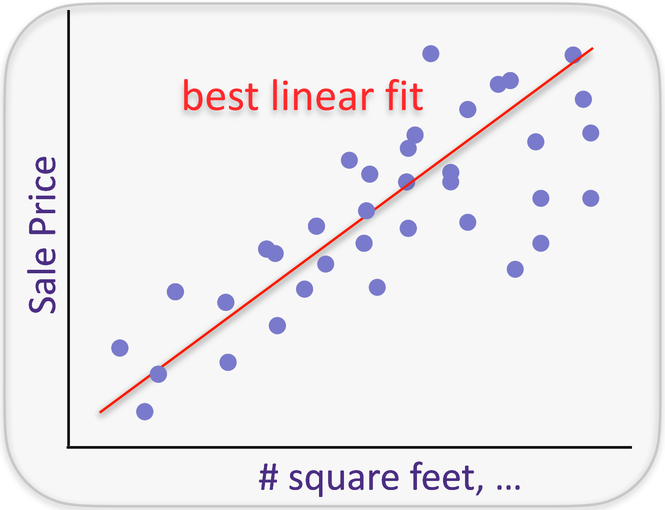
# The regression problem, d-dim



Given past sales data on zillow.com, predict:

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

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




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

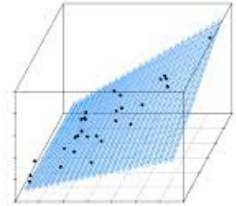
Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

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

# The regression problem, d-dim



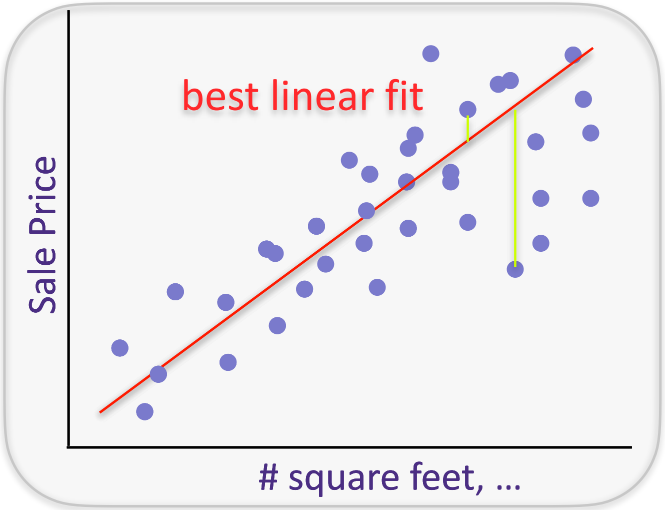
Given past sales data on zillow.com, predict:

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

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

Error:

$$y_i = x_i^T w + \epsilon_i$$




Training Data:

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

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

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

# The regression problem in matrix notation

Sale Prices

single vector in d dim.

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

d : # of features

n : # of examples/datapoints

$$( \text{---} )$$

$$X = \begin{pmatrix} \text{---} & x_1^T & \text{---} \\ \text{---} & x_n^T & \text{---} \end{pmatrix}$$

# The regression problem in matrix notation

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

d : # of features
n : # of examples/datapoints

$x_i^T \in \mathbb{R}^d \quad i \in [n]$

$$y_i \approx x_i^T w$$

$$y_i = x_i^T w + \epsilon_i$$

# The regression problem in matrix notation

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

d : # of features

n : # of examples/datapoints

$$\begin{pmatrix} x_i^T \\ \vdots \\ x_n^T \end{pmatrix}$$

$$y_i \approx x_i^T w$$

$$y_i = x_i^T w + \epsilon_i$$

$$\underline{\mathbf{y} = \mathbf{X} w + \epsilon}$$

# Process

particular set

Decide on a **model** Linear in d dimensions

Find the function which fits the data best

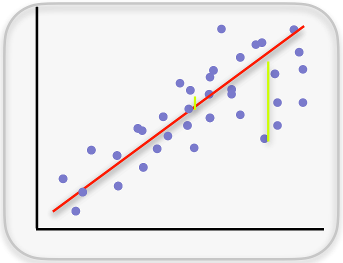
**Choose a loss function- least squares**

**Pick the function which minimizes loss on data**

Use function to make prediction on new examples

# Loss function: least squares in matrix notation

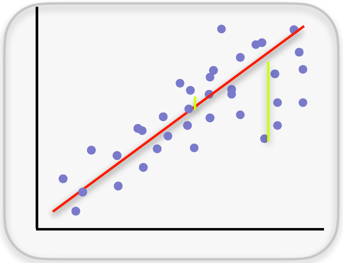
$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$



Error:
$y_i = x_i^T w + \epsilon_i$

# Loss function: least squares in matrix notation

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



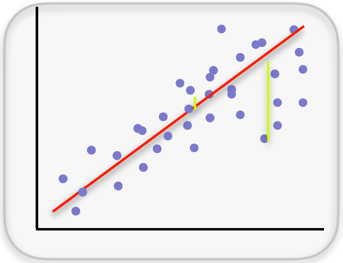
Error:
$y_i = x_i^T w + \epsilon_i$

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

$$\frac{d}{dw} = \sum_{i=1}^n 2(y_i - x_i^T w) \cdot -x_i^T$$

# Loss function: least squares in matrix notation

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



Error: $y_i = x_i^T w + \epsilon_i$

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

$$= \arg\min_w (\mathbf{y} - \mathbf{X}w)^T(\mathbf{y} - \mathbf{X}w)$$

$$\frac{d}{dw} = \sum_i \cancel{2}(y_i - x_i^T w)(-x_i^T)$$

$$0 = -\sum_i y_i x_i^T - x_i^{T2} w$$

0 = - E yixi^T - xi^T^2 * w

0 = -X^T(y-Xw)

0 = -X^Ty + X^TXw

X^Ty = X^TXw

w = (X^TX)^-1 * X^Ty

# The regression problem in matrix notation

$0 = -X^T(y - Xw)$

$X^T y = X^T X w$

$w = (X^T X)^{-1} X^T y$

$$\widehat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= \arg\min_w (\mathbf{y} - \mathbf{X}w)^T(\mathbf{y} - \mathbf{X}w)$$

# The regression problem in matrix notation

$$\widehat{w}_{MLE} = \widehat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= \arg\min_w (\mathbf{y} - \mathbf{X}w)^T(\mathbf{y} - \mathbf{X}w)$$
$$= \underline{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}$$

at least d linearly independent rows

X

n

d

# The regression problem in matrix notation

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

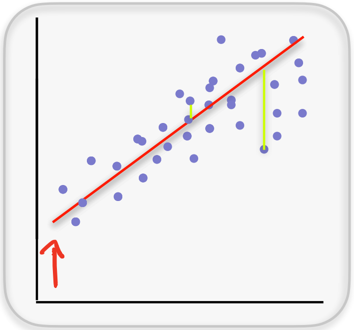
"Closed form" solution!

# The regression problem: an offset

$$\widehat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

# The regression problem: an offset

$$\widehat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$



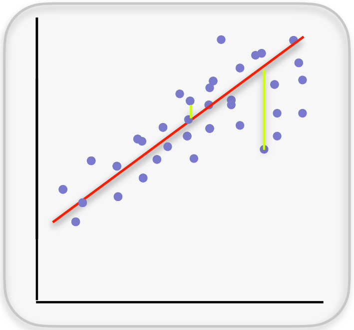
Error:

$\cancel{y_i = x_i^T w + \epsilon_i}$

$y_i = x_i^T w + b + \epsilon_i$

# The regression problem: an offset

$$\widehat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$



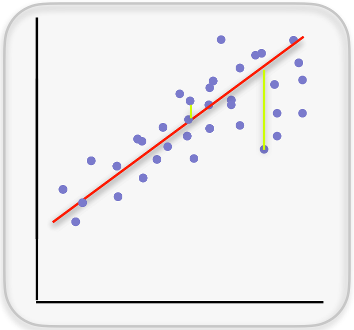
Error:

$$y_i = x_i^T w + \epsilon_i$$

What about an offset?

# The regression problem: an offset

$$\widehat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$



Error:

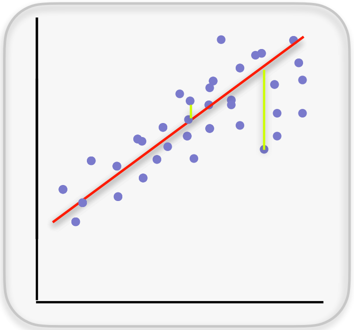
$$y_i = x_i^T w + \epsilon_i$$

What about an offset?

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} \sum_{i=1}^n \left(y_i - (x_i^T w + b)\right)^2$$

# The regression problem: an offset

$$\widehat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$



Error:

$y_i = x_i^T w + \epsilon_i$ $+b$

What about an offset?

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} \sum_{i=1}^n \left(y_i - (x_i^T w + b)\right)^2$$
$$= \arg\min_{w,b} \boxed{||\mathbf{y} - (\underline{\mathbf{X}w + \mathbf{1}b})||_2^2}$$

# Dealing with an offset

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

# Dealing with an offset

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

# Dealing with an offset

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

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

# Dealing with an offset

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

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

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

If $X^T$ is invertible

# Dealing with an offset

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

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

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

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

# Dealing with an offset

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

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

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

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

# Dealing with an offset

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

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

$$\underbrace{\cancel{\mathbf{1}^T\mathbf{X}\widehat{w}_{LS}}}_{0} + \underline{\widehat{b}_{LS}\mathbf{1}^T\mathbf{1} = \mathbf{1}^T\mathbf{y}}$$

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

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

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

# Process

Decide on a **model**

Find the function which fits the data best

- **Choose a loss function- least squares**
- **Pick the function which minimizes loss on data**

Use function to make prediction on new examples

# Make Predictions

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

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

A new house is about to be listed. What should it sell for?

# Make Predictions

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

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

A new house is about to be listed. What should it sell for?

$$\hat{y}_{\text{new}} = x_{\text{new}}^T \hat{w}_{LS} + \boxed{\hat{b}_{LS}}$$

new house

$\hat{w}_{LS} = ( \quad )$

$x_{new} = ( \quad )$

# Process

Decide on a **model**

Find the function which fits the data best

- **Choose a loss function- least squares**
- **Pick the function which minimizes loss on data**

Use function to make prediction on new examples

# Process

Decide on a **model**

Find the function which fits the data best

**Choose a loss function- least squares**

**Pick the function which minimizes loss on data**

Use function to make prediction on new examples

**Why did we choose this loss function?**

# Process

Decide on a **model**

Find the function which fits the data best

**Choose a loss function- least squares**

**Pick the function which minimizes loss on data**

Use function to make prediction on new examples

**Why did we choose this loss function?**

# Why is least squares a good loss function?

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

# Why is least squares a good loss function?

$$\widehat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Noise, error

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

$y_i \sim N(x_i^T w, \sigma^2)$

$N(x, y)$

$y = a + R$

$y \sim N(x + a, y)$

# What is the probability of training data I w?

## Maximize Log Likelihood:

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

$= n \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \sum_i \frac{(y_i - x_i^T w)^2}{2\sigma^2}$

Density fn of Normal dist w. mean $x_i^T w$

$\max_w \log P(D|w,\sigma)$

$= \max_w \; n\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \sum \frac{(y_i - x_i^T w)^2}{2\sigma^2}$

$= \max_w \; -\sum (y_i - x_i^T w)^2$

maximizing a negative means minimizing a positive

$= \min_w \sum (y_i - x_i^T w)^2$

# MLE is LS under linear model

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

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

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

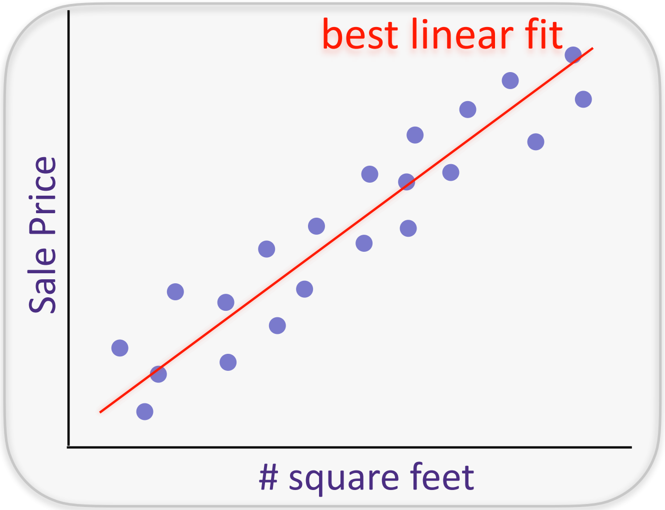
$$\boxed{\widehat{w}_{LS} = \widehat{w}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}}$$

# The regression problem

Given past sales data on zillow.com, predict:

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

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




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

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

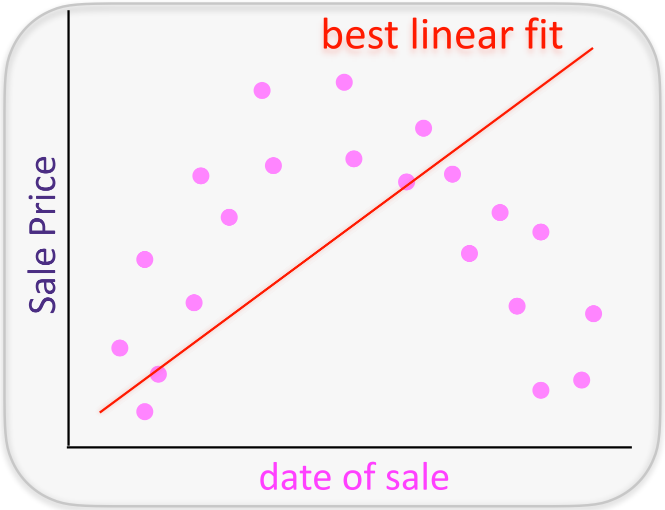
$$\min_w \sum_{i=1}^n \left(y_i - x_i^T w\right)^2$$

# The regression problem

Given past sales data on zillow.com, predict:

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

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




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

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

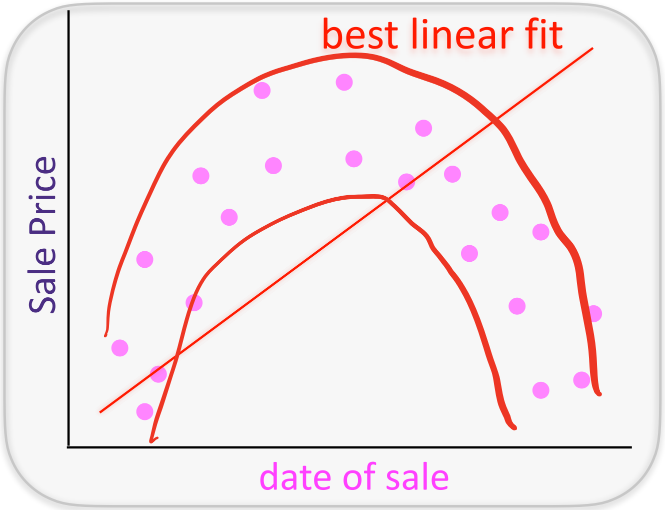
$$\min_w \sum_{i=1}^n \left(y_i - x_i^T w\right)^2$$

# The regression problem

Given past sales data on zillow.com, predict:

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

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




Best linear model of data of sale is a very poor fit!

Either because date of sale doesn't predict price well, or…

… because the relationship isn't linear.

# Process

Decide on a **model** Polynomial (quadratic)

Find the function which fits the data best

**Choose a loss function**

**Pick the function which minimizes loss on data**

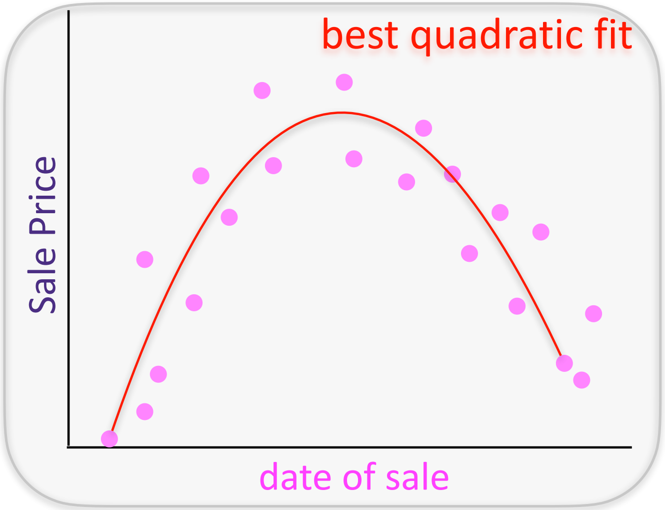
Use function to make prediction on new examples

# Quadratic Regression

Given past sales data on zillow.com, predict:

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

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




Training Data:

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

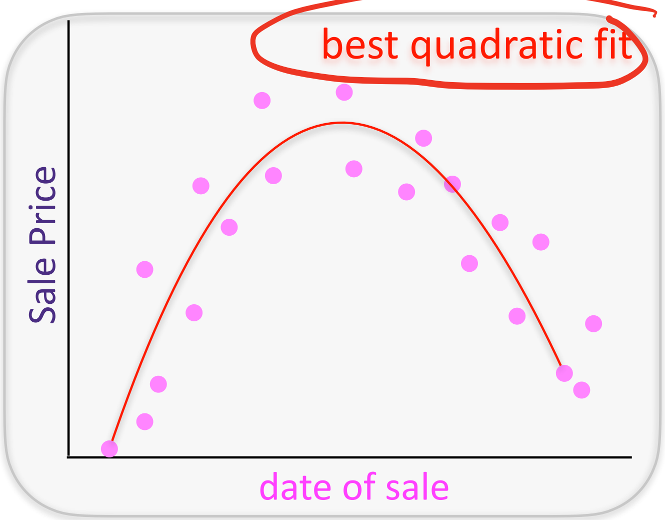
$$y_i \approx \sum_{j=1}^{d} x_{i,j} w_{j,1} + x_{i,j}^2 w_{j,2}$$

# Quadratic Regression

Given past sales data on zillow.com, predict:

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

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




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

Hypothesis: quadratic

$$y_i \approx \sum_{j=1}^d x_{i,j} w_{j,1} + x_{i,j}^2 w_{j,2}$$

Loss fn??

$$\min_w \sum \Big( y_i - \sum_{j=1}^d x_{ij} w_{j1} + x_{ij}^2 \cdot w_{j2} \Big)^2$$

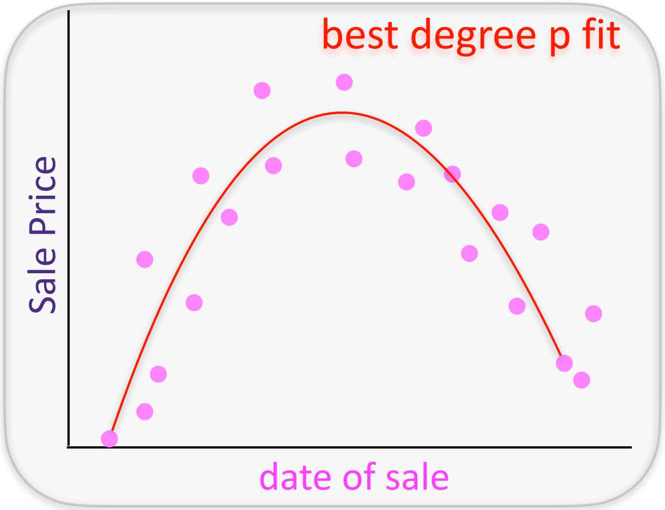
Least Squares

# Polynomial regression

Given past sales data on zillow.com, predict:

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

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




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

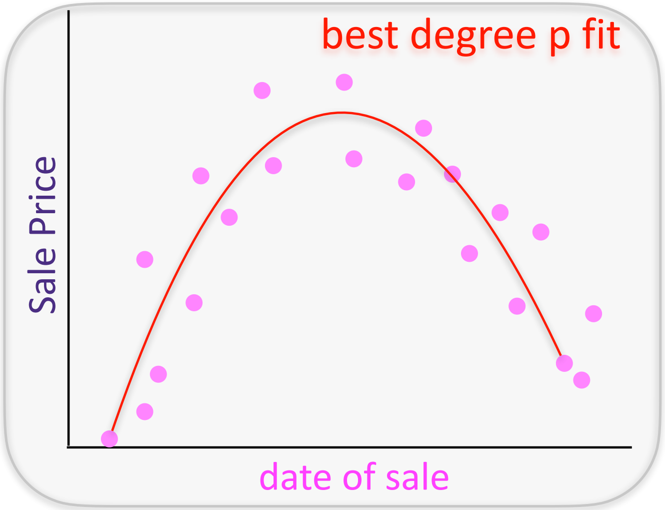
$$y_i \approx \sum_{j=1}^{d} \sum_{\ell=1}^{p} x_{i,j}^{\ell} w_{j,\ell}$$

# Polynomial regression

Given past sales data on zillow.com, predict:

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

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




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

Hypothesis:

degree p polynomial

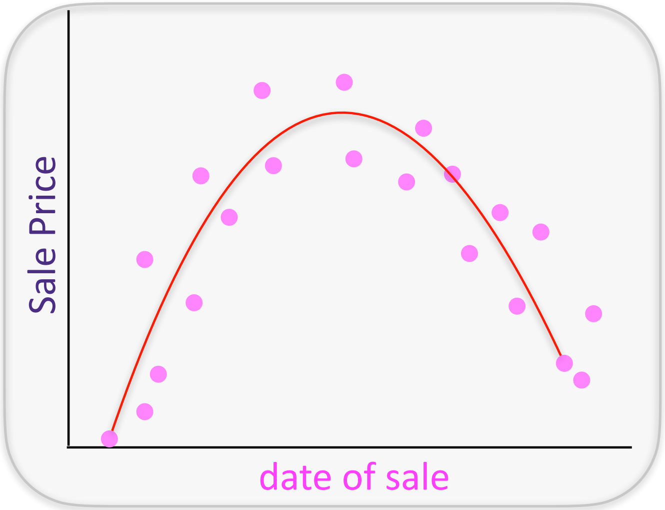
$$y_i \approx \sum_{j=1}^{d} \sum_{\ell=1}^{p} x_{i,j}^{\ell} w_{j,\ell}$$

# Generalized linear regression

Given past sales data on zillow.com, predict:

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

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




Training Data:

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

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

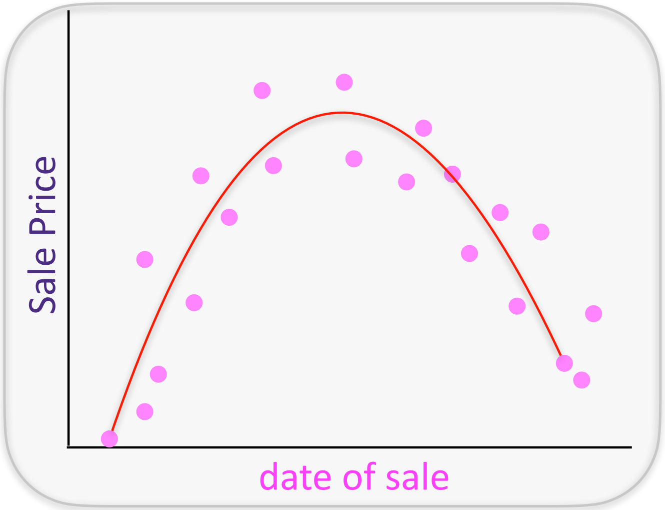
$$y_i \approx \sum_{\ell=1}^{p} h_\ell(x_i)^T w_\ell$$

# Generalized linear regression

Given past sales data on zillow.com, predict:

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

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




Training Data: $x_i \in \mathbb{R}^d$ $y_i \in \mathbb{R}$

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

Hypothesis:

generalized linear fn of x

$$y_i \approx \sum_{\ell=1}^{p} h_\ell(x_i)^T w_\ell$$

# Process

Decide on a **model**

Find the function which fits the data best

- **Choose a loss function**
- **Pick the function which minimizes loss on data**

Use function to make prediction on new examples

# The regression problem

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

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

Transformed data:

# The regression problem

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

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

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

Transformed data:

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

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

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

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

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

# The regression problem

Training Data:

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

~~Hypothesis: linear~~

$$\cancel{y_i \approx x_i^T w}$$

~~Loss: least squares~~

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

Transformed data:

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

# The regression problem

Training Data:

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

~~Hypothesis: linear~~

$$\cancel{y_i \approx x_i^T w}$$

~~Loss: least squares~~

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

Transformed data:

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

Hypothesis: linear in h

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

# The regression problem

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

~~Hypothesis: linear~~

$$\cancel{y_i \approx x_i^T w}$$

~~Loss: least squares~~

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

Transformed data:

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

Hypothesis: linear in h

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

Loss: least squares

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

# 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 in h

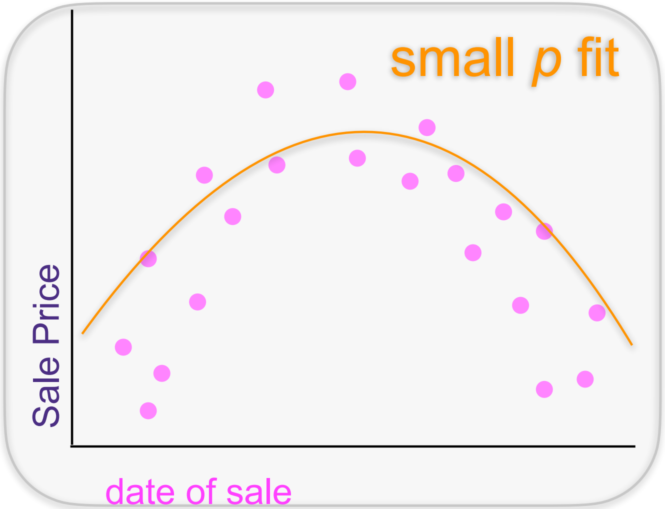
$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

Loss: least squares

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

# The regression problem

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




Transformed data:

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

Hypothesis: linear in h

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

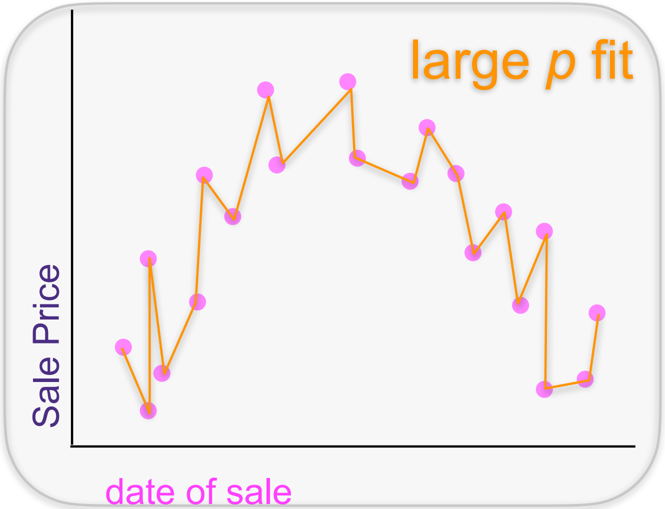
Loss: least squares

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

# The regression problem

Training Data:

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




Transformed data:

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

Hypothesis: linear in h

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

Loss: least squares

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

# Bias-Variance Tradeoff

# Statistical Learning

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

**Goal: Predict Y given X**

Find function $\eta$ that minimizes

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

# Statistical Learning

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

**Goal: Predict Y given X**

Find function $\eta$ that minimizes

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

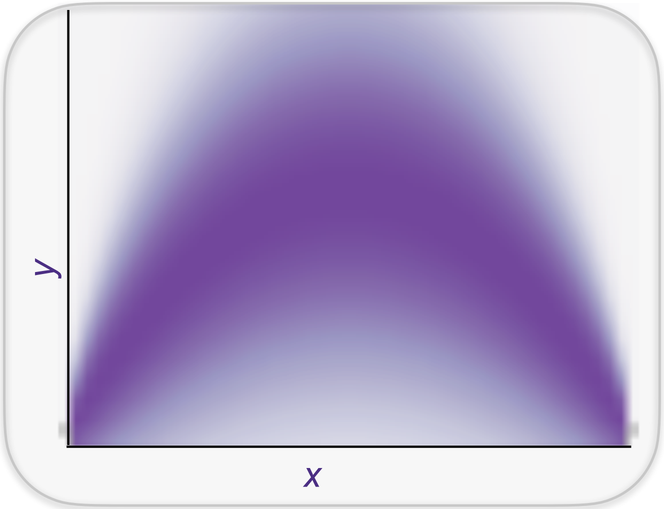
$$\eta(x) = \arg\min_c \mathbb{E}_{Y|X}[(Y - c)^2 | X = x] = \mathbb{E}_{Y|X}[Y | X = x]$$

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

# Statistical Learning

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

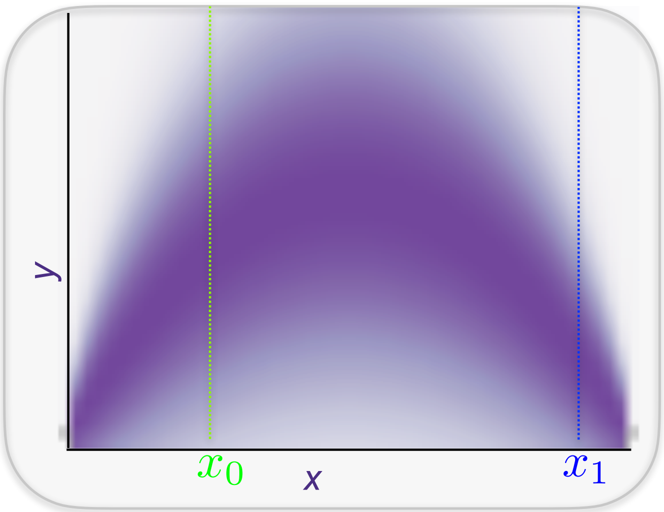
$$P_{XY}(X = x, Y = y)$$



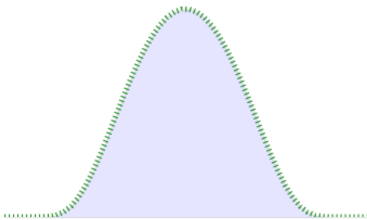
# Statistical Learning

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

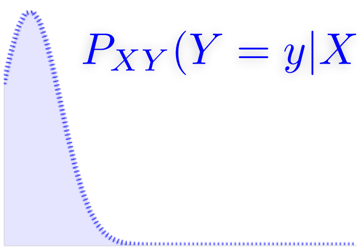
$$P_{XY}(X = x, Y = y)$$




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



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



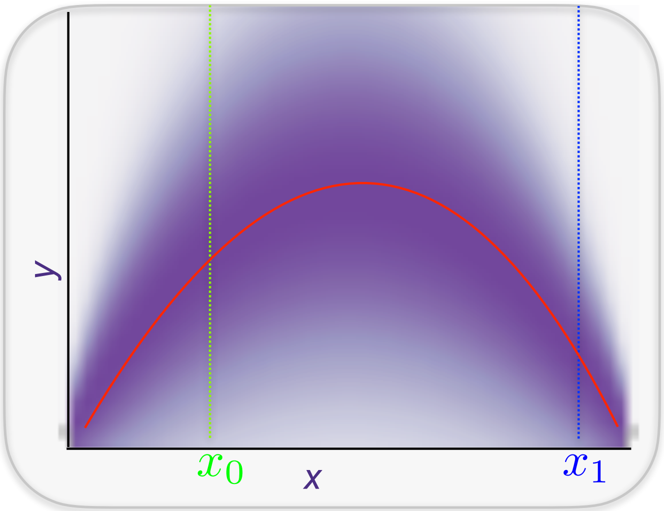
# Statistical Learning

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

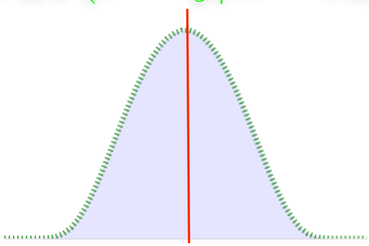
Ideally, we want to find:

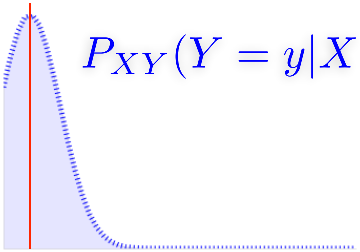
$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

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




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

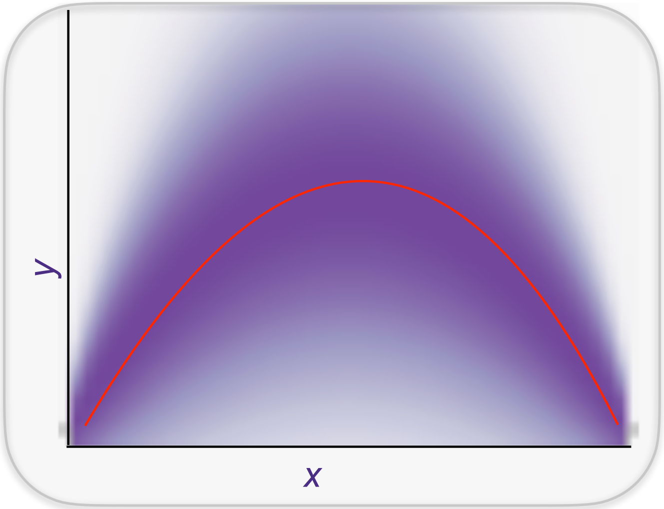




$P_{XY}(Y = y|X = x_1)$

# Statistical Learning

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

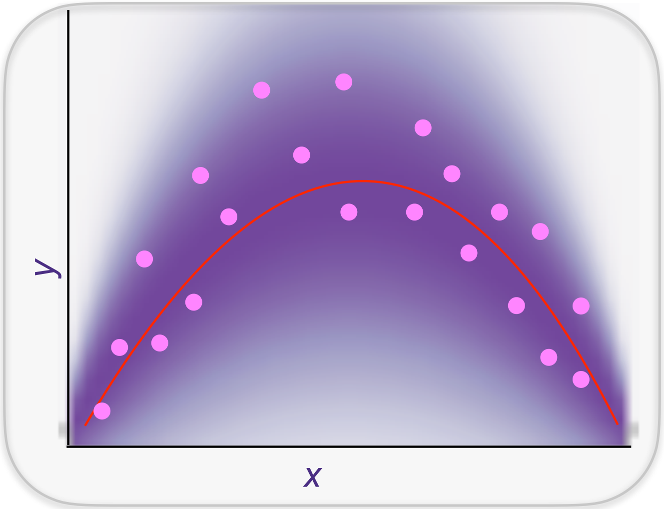



Ideally, we want to find:

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

# Statistical Learning

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




Ideally, we want to find:

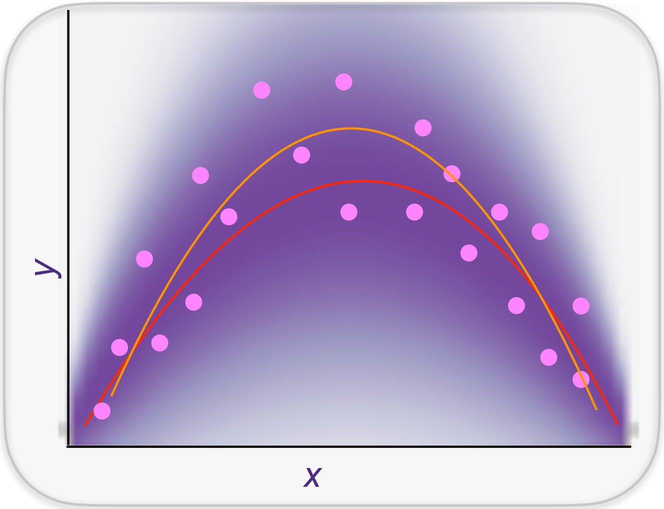
$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

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

# 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:

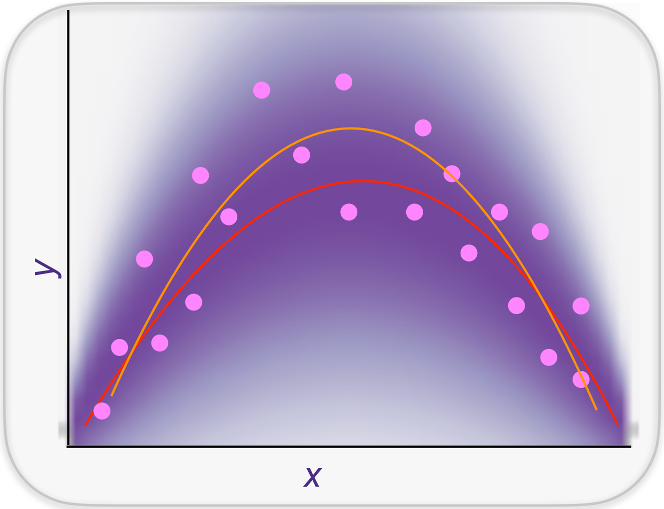
$$(x_i, y_i) \overset{i.i.d.}{\sim} P_{XY} \quad \text{for } i = 1, \dots, n$$

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

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

# 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:

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

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

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

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

# 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:

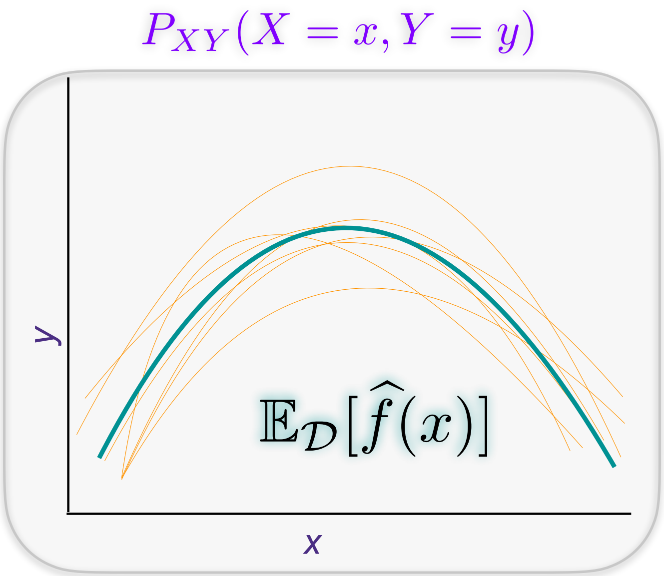
$$(x_i, y_i) \overset{i.i.d.}{\sim} P_{XY} \quad \text{for } i = 1, \dots, n$$

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

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

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

# Statistical Learning




Ideally, we want to find:

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

But we only have samples:

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

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

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

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

# Bias-Variance Tradeoff

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

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

# Bias-Variance Tradeoff

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

$$\begin{aligned}
\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y-\widehat{f}_{\mathcal{D}}(x))^2]\big|X=x] &= \mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y-\eta(x)+\eta(x)-\widehat{f}_{\mathcal{D}}(x))^2]\big|X=x] \\
&= \mathbb{E}_{Y|X}\Big[\mathbb{E}_{\mathcal{D}}[(Y-\eta(x))^2 + 2(Y-\eta(x))(\eta(x)-\widehat{f}_{\mathcal{D}}(x)) \\
&\quad + (\eta(x)-\widehat{f}_{\mathcal{D}}(x))^2]\big|X=x\Big] \\
&= \underbrace{\mathbb{E}_{Y|X}[(Y-\eta(x))^2\big|X=x]} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\eta(x)-\widehat{f}_{\mathcal{D}}(x))^2]}
\end{aligned}$$

**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

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

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

# Bias-Variance Tradeoff

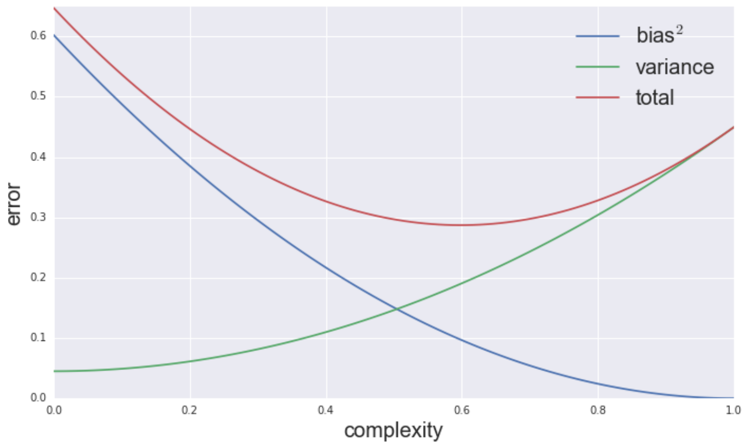
$$\eta(x) = \mathbb{E}_{Y|X}[Y|X=x] \qquad \widehat{f} = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

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

# Bias-Variance Tradeoff

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

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



# Example: Linear LS

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

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

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

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

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

# Example: Linear LS

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

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

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

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

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

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

# Example: Linear LS

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

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

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

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

$$\underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]}_{\text{variance}} =$$

# Example: Linear LS

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

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

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

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

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

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

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

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

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

# Example: Linear LS

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

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

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

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

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

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

**irreducible error**

$$\underline{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])^2} = 0$$

**biased squared**

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

**variance**