Maximum Likelihood Estimation

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Likelihood function $L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$

Log-Likelihood function $l_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^{n} \log(f(X_i;\theta))$

Maximum Likelihood Estimator (MLE) $\widehat{\theta}_{MLE} = \arg\max_{\theta} L_n(\theta)$

What about continuous variables?

- *Client*: What if I am measuring a **continuous variable**?
- ***You*: Let me tell you about Gaussians…**

$$P(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Some properties of Gaussians

- affine transformation (multiplying by scalar and adding a constant)
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b \quad \rightarrow \quad Y \sim N(a\mu+b,a^2\sigma^2)$

- Sum of Gaussians
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y \quad \rightarrow \quad Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

MLE for Gaussian

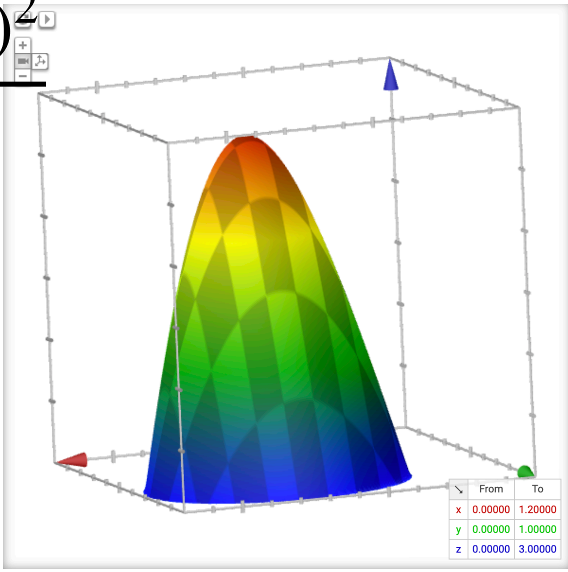
- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_n\}$ (e.g., temperature):

$$P(\mathcal{D};\mu,\sigma) = P(x_1,\ldots,x_n;\mu,\sigma)$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- Log-likelihood of data:

$$\log P(\mathcal{D};\mu,\sigma) = -n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

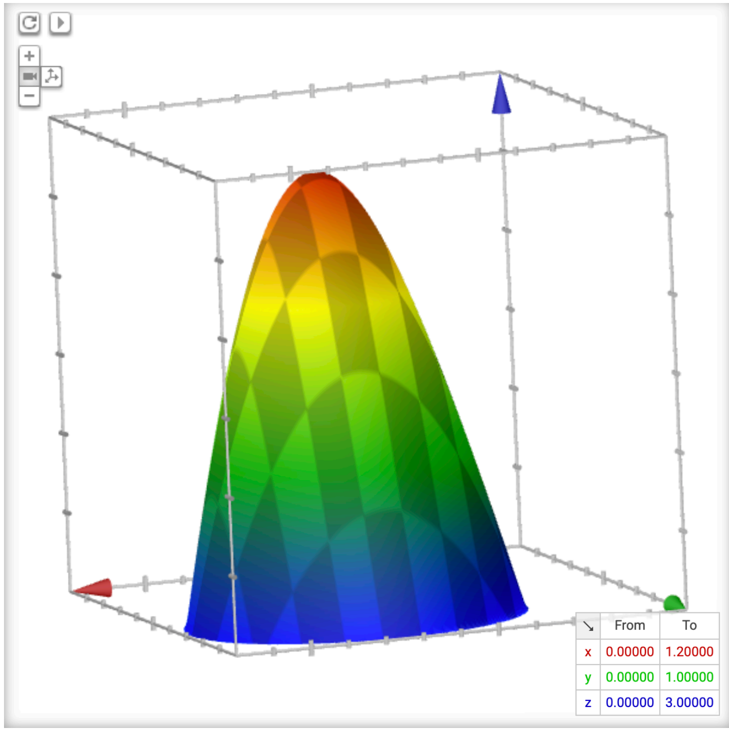


- What is $\widehat{\theta}_{MLE}$ for $\theta = (\mu,\sigma^2)$?

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

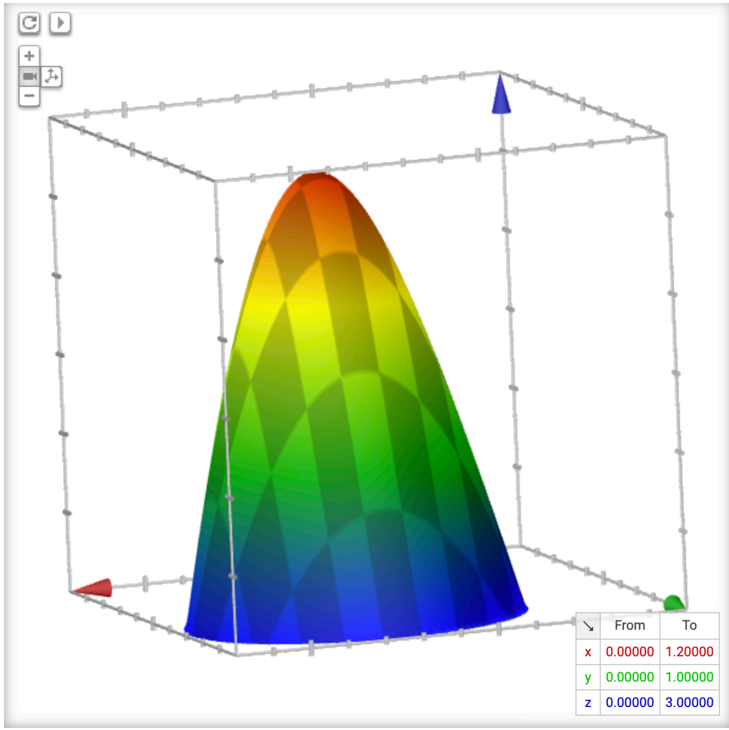
$$\frac{d}{d\mu}\log P(\mathcal{D};\mu,\sigma) = \frac{d}{d\mu}\Big[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\Big]$$



MLE for variance

- Again, set derivative to zero:

$$\frac{d}{d\sigma}\log P(\mathscr{D};\mu,\sigma) = \frac{d}{d\sigma}\left[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$



What can we say about the MLE?

- MLE:

$$\widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\sigma^2}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu}_{MLE})^2$$

- MLE for the variance of a Gaussian is **biased**

$$\mathbb{E}[\widehat{\sigma^2}_{MLE}] \neq \sigma^2$$

- Unbiased variance estimator:

$$\widehat{\sigma^2}_{unbiased} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu}_{MLE})^2$$

Maximum Likelihood Estimation

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Likelihood function $L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$

Log-Likelihood function $l_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^{n} \log(f(X_i;\theta))$

Maximum Likelihood Estimator (MLE) $\widehat{\theta}_{MLE} = \arg\max_{\theta} L_n(\theta)$

Properties (under benign regularity conditions—smoothness, identifiability, etc.):

- Asymptotically consistent and normal: $\frac{\widehat{\theta}_{MLE} - \theta_*}{\widehat{se}} \sim \mathcal{N}(0,1)$
- Asymptotic Optimality, minimum variance (see Cramer-Rao lower bound)

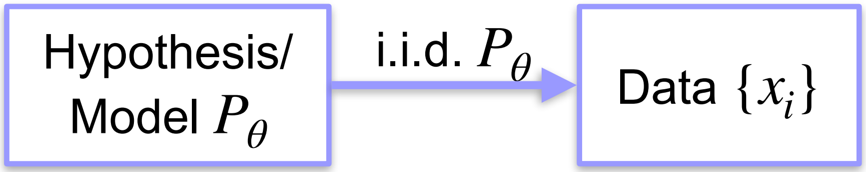
Recap

- Learning is…
  - Collect some data
    - E.g., coin flips

Data $\{x_i\}$

Recap

- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., binomial



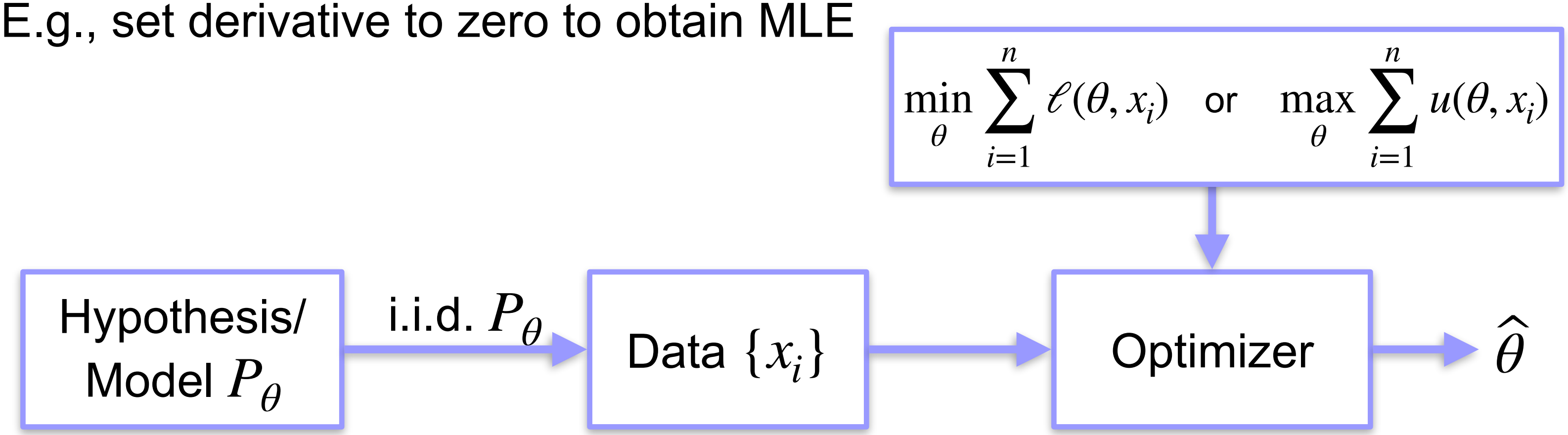
Recap

- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., binomial
  - Choose a loss function
    - E.g., data likelihood

$$\min_{\theta} \sum_{i=1}^{n} \ell(\theta, x_i) \quad \text{or} \quad \max_{\theta} \sum_{i=1}^{n} u(\theta, x_i)$$

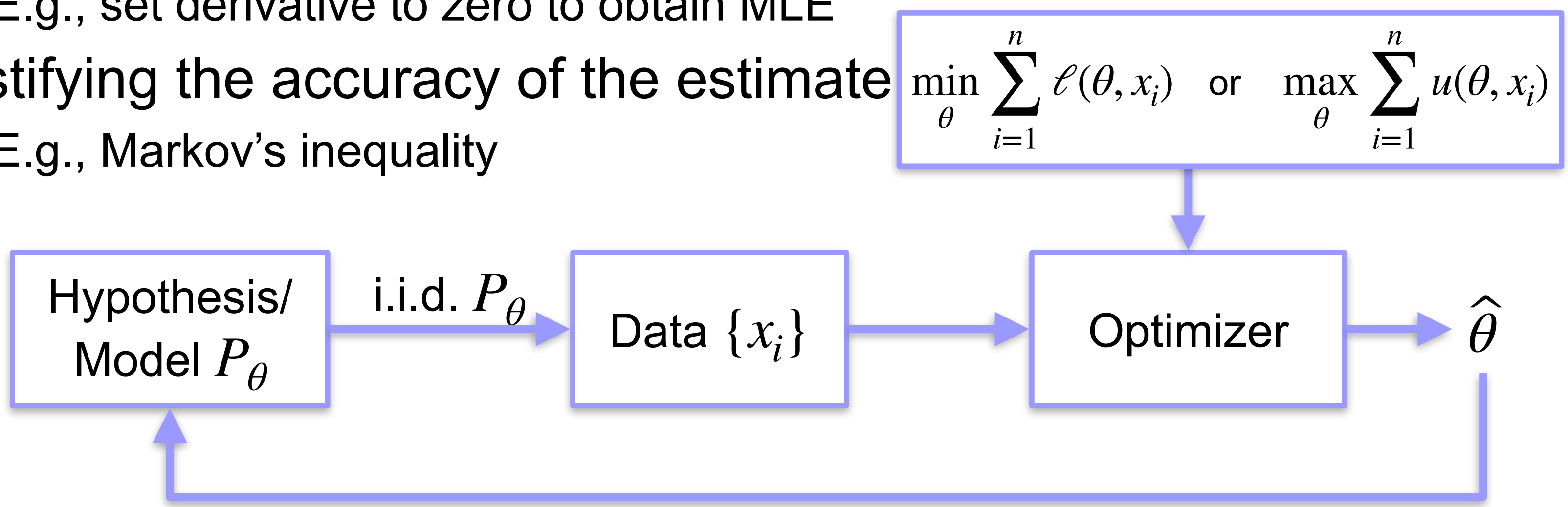
Recap

- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., binomial
  - Choose a loss function
    - E.g., data likelihood
  - Choose an optimization procedure
    - E.g., set derivative to zero to obtain MLE



Recap

- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., binomial
  - Choose a loss function
    - E.g., data likelihood
  - Choose an optimization procedure
    - E.g., set derivative to zero to obtain MLE
  - Justifying the accuracy of the estimate
    - E.g., Markov's inequality



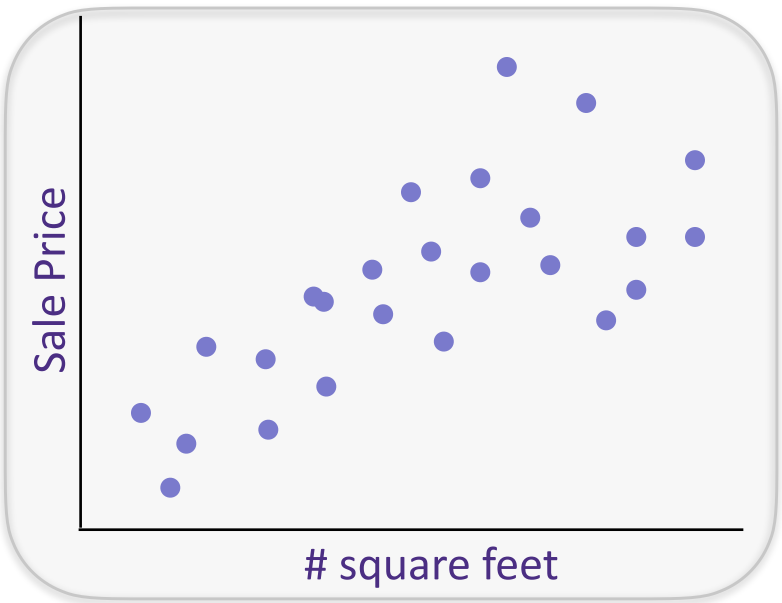
Linear Regression

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**

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

Find the function which fits the data best

Use function to make prediction on new examples

Fit a function to our data, 1-dimension

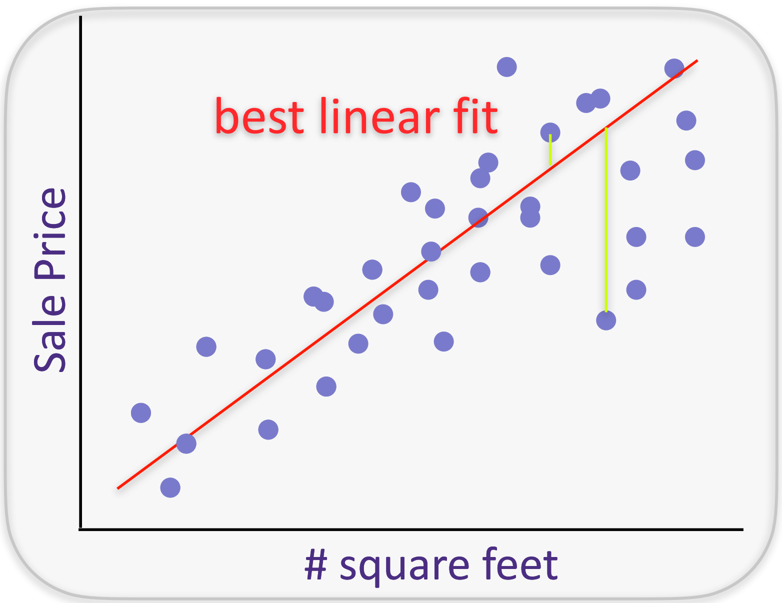
Given past sales data on zillow.com, predict:

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

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

Error

$$y_i = x_i w + \epsilon_i$$



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

Hypothesis/Model: linear

$$y_i \approx x_i w$$

Loss: least squares solution

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

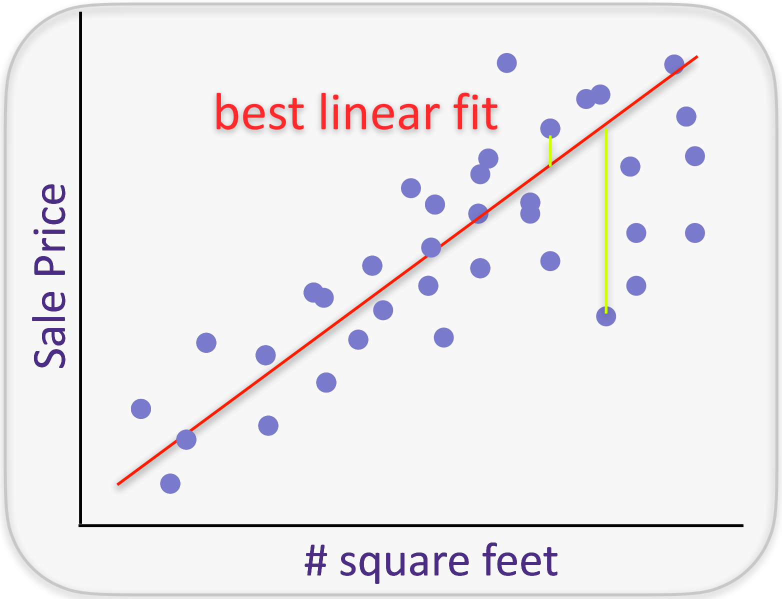
The regression problem, d-dimensions

Given past sales data on zillow.com, predict:

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

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

Error:
$y_i = x_i w + \epsilon_i$



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

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

Hypothesis/Model: linear

$$y_i \approx x_i^T w$$

Loss: least squares solution

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

The regression problem in matrix notation

Data:

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

d : # of features

n : # of examples/datapoints

The regression problem in matrix notation

Data:

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

d : # of features

n : # of examples/datapoints

Model:

$$y_1 = x_1^T w + \epsilon_1$$

$$y_2 = x_2^T w + \epsilon_2$$

$$\vdots$$

$$y_n = x_n^T w + \epsilon_n$$

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

The regression problem in matrix notation

Data:

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

d : # of features

n : # of examples/datapoints

Model:

$$\begin{aligned} y_1 &= x_1^T w + \epsilon_1 \\ y_2 &= x_2^T w + \epsilon_2 \\ &\vdots \\ y_n &= x_n^T w + \epsilon_n \end{aligned}$$

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

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

The regression problem in matrix notation

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

Data:

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

Model:

$$y_1 = x_1^T w + \epsilon_1$$
$$y_2 = x_2^T w + \epsilon_2$$
$$\vdots$$
$$y_n = x_n^T w + \epsilon_n$$

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

Loss: $\widehat{w}_{LS} = \arg\min_w \sum_{i=1}^n \left(y_i - x_i^T w\right)^2 = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$

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

The regression problem in matrix notation

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

Set gradient w.r.t. w to zero to find the minima:

The regression problem in matrix notation

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

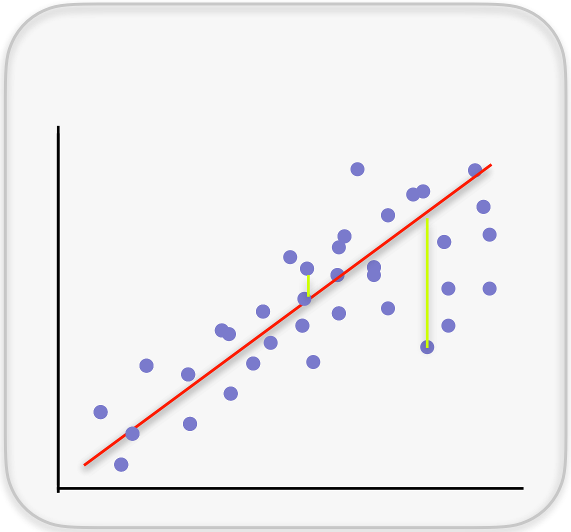
"Closed form" solution!

The regression problem in matrix notation

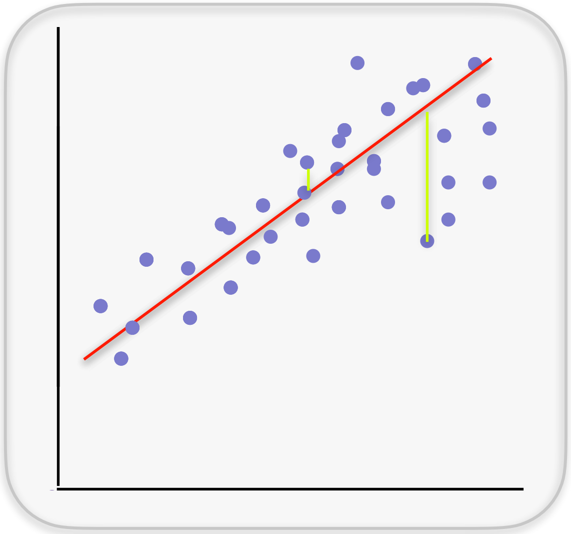
Linear model: $y_i = x_i^T w + \epsilon_i$

Least squares solution:

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

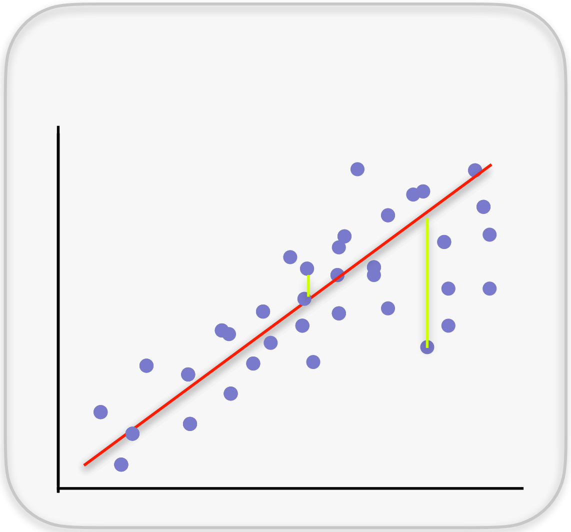


What about an offset (a.k.a intercept)?



The regression problem in matrix notation

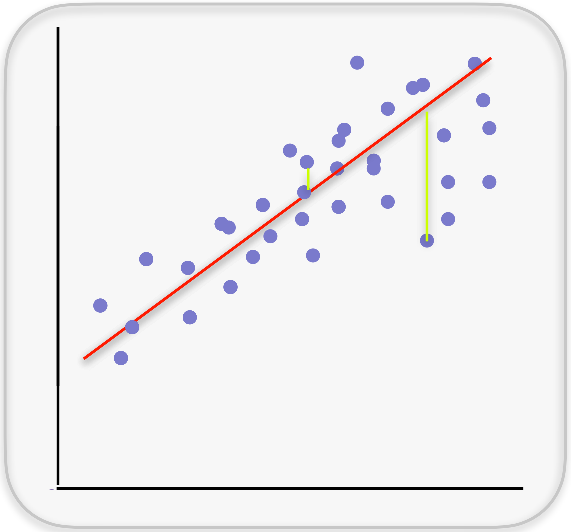
Linear model: $y_i = x_i^T w + \epsilon_i$



Least squares solution:

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

Affine model: $y_i = x_i^T w + b + \epsilon_i$



Least squares solution:

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

Dealing with an offset

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

Set gradient w.r.t. w and b to zero to find the minima:

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 $\mathbf{X}^T\mathbf{1} = 0$, if the features have zero mean,

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

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

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 $\mathbf{X}^T\mathbf{1} = 0$,

In general, when $\mathbf{X}^T\mathbf{1} \neq 0$,

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

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

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 $\mathbf{X}^T\mathbf{1} = 0$,

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

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

In general, when $\mathbf{X}^T\mathbf{1} \neq 0$,

$$\mu = \frac{1}{n}\mathbf{X}^T\mathbf{1}$$

$$\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\mu^T$$

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

$$\widehat{b}_{\text{LS}} = \frac{1}{n}\mathbf{1}^T\mathbf{y} - \mu^T\widehat{w}_{\text{LS}}$$

Process

Decide on a **model:** $y_i = x_i^T w + b + \epsilon_i$

Choose a loss function - least squares

Pick the function which minimizes loss on data

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

Use function to make prediction on new examples

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

Dealing with an offset - Revisted

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

$$\widetilde{\mathbf{X}} = (\mathbf{X}, \mathbf{1})$$

$$\widetilde{\mathbf{X}}\widetilde{w} = \mathbf{X}w + b\mathbf{1} \quad \text{with} \quad \widetilde{w} = (w, b)$$

Why is least squares a good loss function?

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

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

$$P(y_i; x_i, w, \sigma) \; =$$

Why is least squares a good loss function?

Maximum Likelihood Estimator:

$$\widehat{w}_{\text{MLE}} = \arg\max_w \log P(\{y_i\}_{i=1}^n; \{x_i\}_{i=1}^n, w, \sigma)$$

$$= \arg\max_w -n\log(\sigma\sqrt{2\pi}) + \sum_{i=1}^n -\frac{(y_i - x_i^T w)^2}{2\sigma^2}$$

Why is least squares a good loss function?

Maximum Likelihood Estimator:

$$\begin{aligned}\widehat{w}_{\text{MLE}} &= \arg\max_w \log P(\{y_i\}_{i=1}^n; \{x_i\}_{i=1}^n, w, \sigma) \\ &= \arg\max_w -n\log(\sigma\sqrt{2\pi}) + \sum_{i=1}^n -\frac{(y_i - x_i^T w)^2}{2\sigma^2} \\ &= \arg\min_w \sum_{i=1}^n (y_i - x_i^T w)^2\end{aligned}$$

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

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

Analysis of Error

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

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

Maximum Likelihood Estimator is unbiased:

Analysis of Error

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

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

Covariance is:

Analysis of Error

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

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

$$\mathbb{E}[\widehat{w}_{MLE}] = w$$

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

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