My office hrs:

Monday immediately after lecture.

Linear Regression

UNIVERSITY *of* WASHINGTON

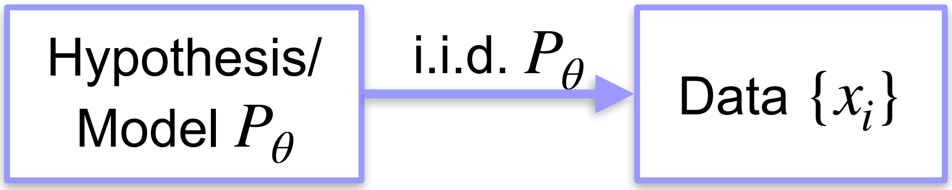
Recap

- 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



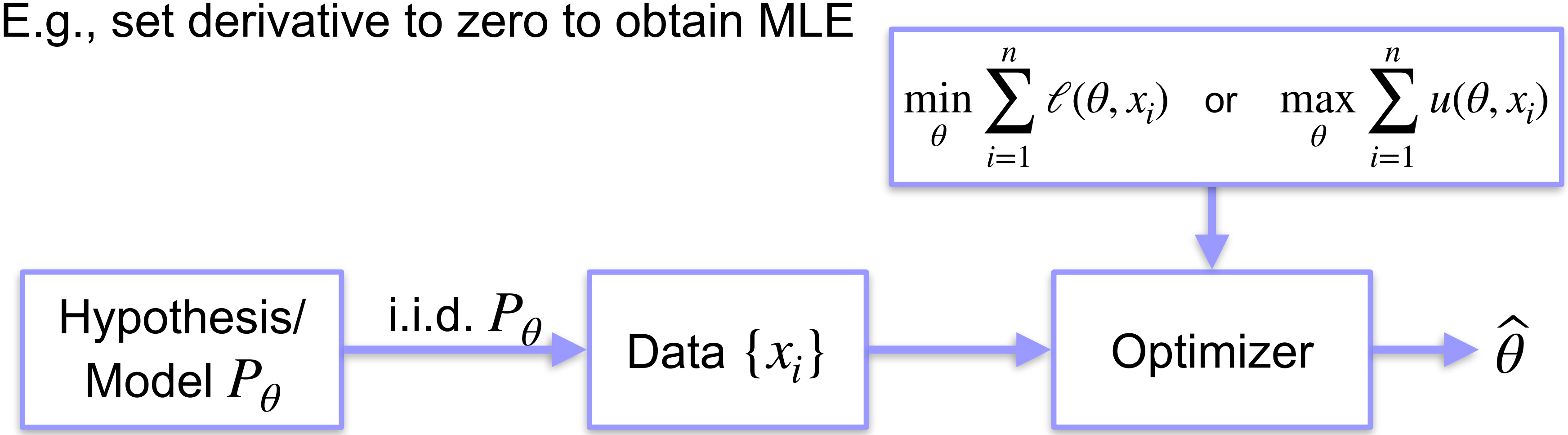
Recap

- 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

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

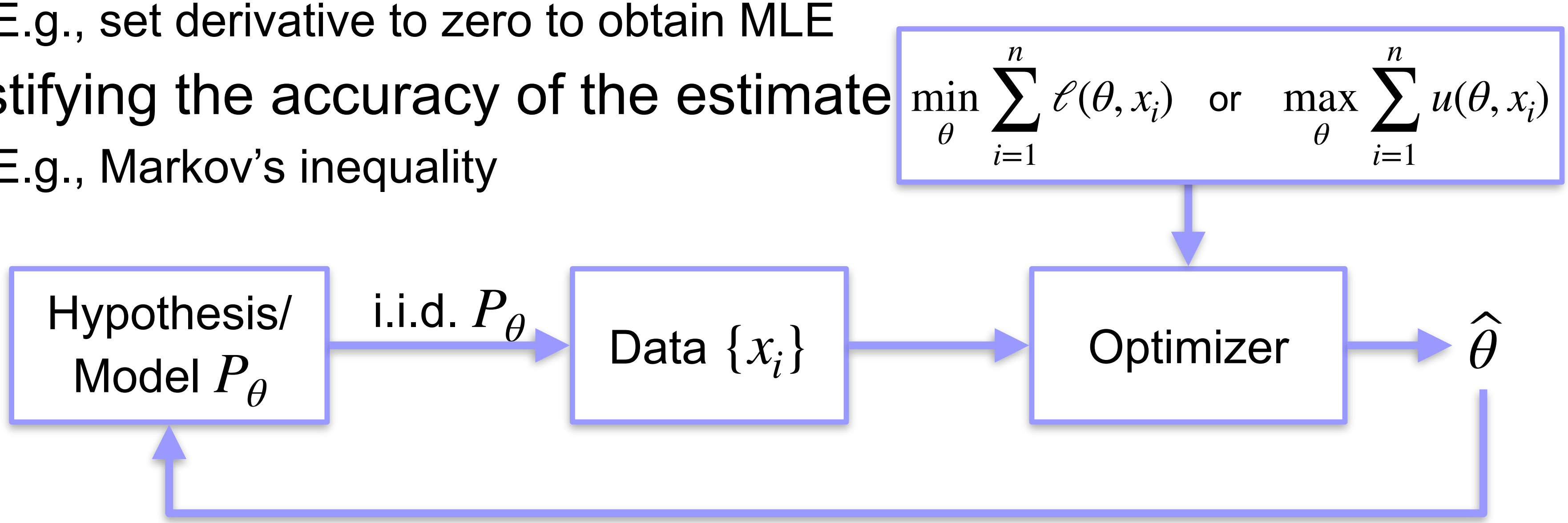
Recap

- 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

Hypothesis / Model

Real-valued predictions (supervised)

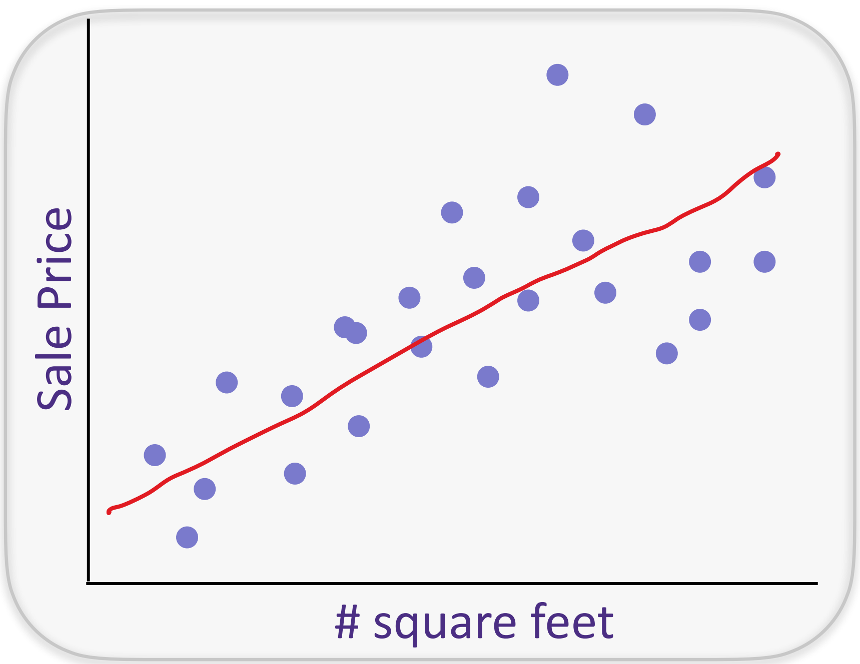
The regression problem, 1-dimensional

You want to sell your house that is 2,500 sq.ft.

Q. What is the right price?

Collect past sales data on zillow.com:

y = **House sale price** *and* x = **{# sq. ft.}**



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

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

label / sale price

square footage

Process

1. Decide on a **model/hypothesis class**

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

+ Noise/error

2. Find the function/model/hypothesis which explains/fits the data best

3. Use function to make prediction on new examples

How much should you put your house on the market?

Fit a function to our data, 1-dimension

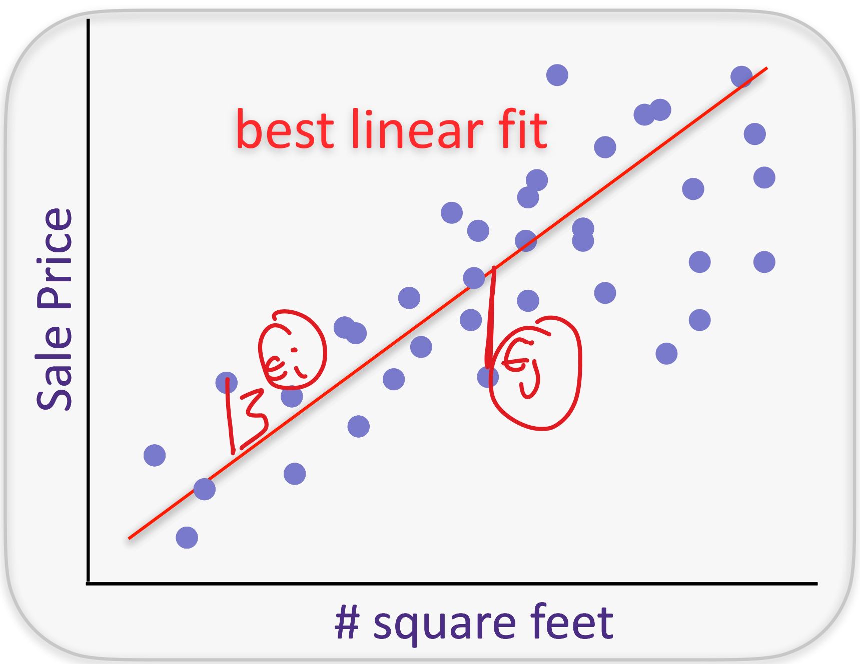
Given past sales data on zillow.com, predict:

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

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

Fixing $w=10$

$\varepsilon_i \quad 10 \cdot x_i - y_i$



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

2. Hypothesis/Model: linear

$$y_i = w \cdot x_i + \epsilon_i$$

Noise

$y_i - w x_i = \epsilon_i$

3. Measure of good fit: ℓ_2-loss

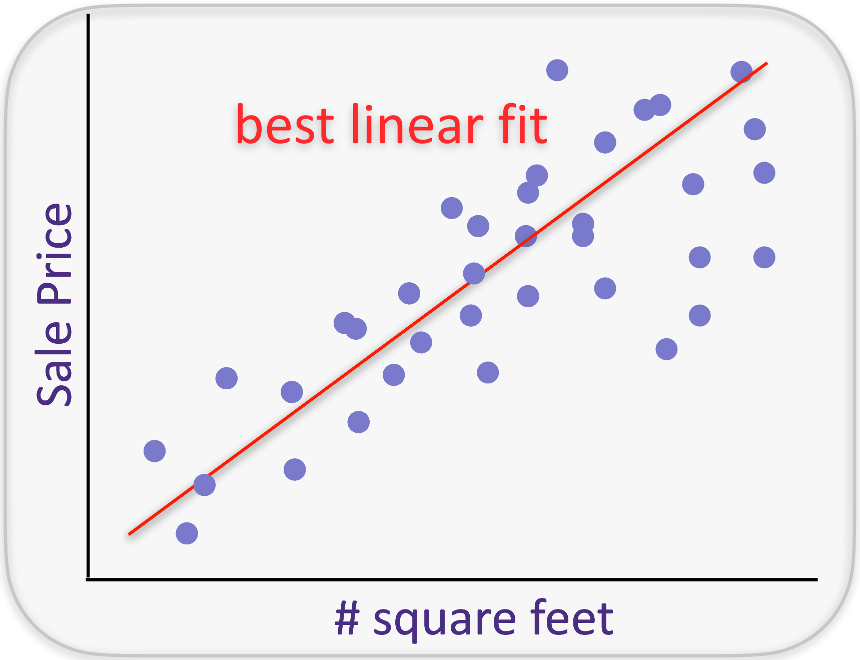
$$\min_{w \in \mathbb{R}} \sum_{i=1}^n (y_i - w x_i)^2 = \sum_{i=1}^n \varepsilon_i^2$$

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.}**



1. Training Data:

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

2. Hypothesis/Model: linear

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

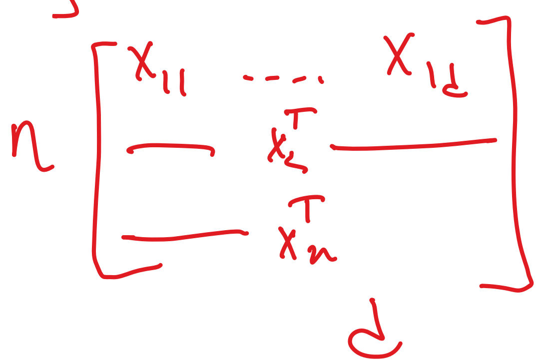
3. Measure of good fit: ℓ_2-loss

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^n (y_i - w^T x_i)^2 = \sum_{i=1}^n \varepsilon_i^2$$

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/size of the input

n : # of examples/datapoints

The regression problem in matrix notation

d : # of features/size of the input

n : # of examples/datapoints

Data:

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

Linear Model:

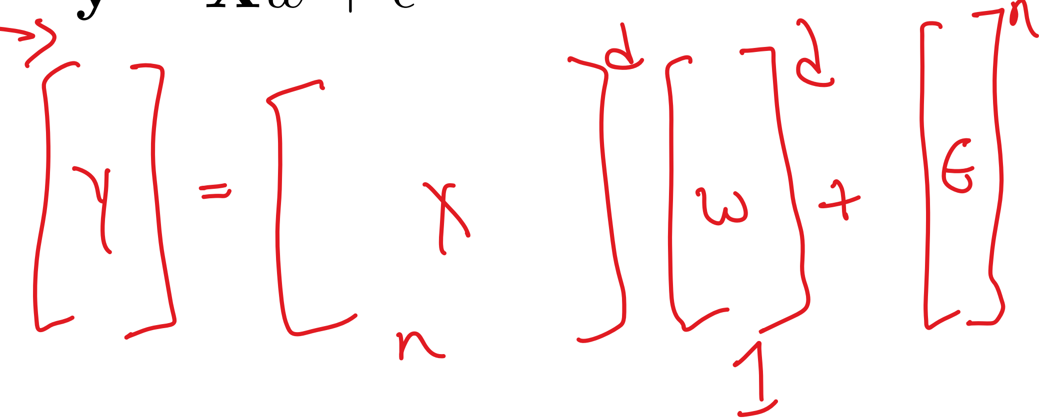
$$y_1 = x_1^T w + \epsilon_1$$

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

$$\vdots$$

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

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



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/size of the input

n : # of examples/datapoints

Linear Model:

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

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

ℓ_2-norm of a vector:
(also known as Euclidean norm)

$$\|\epsilon\|_2 = \sqrt{\epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_d^2}$$

it follows that

$$\sum_{i=1}^{d} \epsilon_i^2 = \|\epsilon\|_2^2 = \epsilon^T \epsilon$$

ℓ_2-Loss: $\widehat{w}_{LS} = \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - x_i^T w)^2$

this is also known as **Least Squares** solution

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/size of the input

n : # of examples/datapoints

Linear Model:

$$y_1 = x_1^T w + \epsilon_1 \qquad \mathbf{y} = \mathbf{X}w + \epsilon$$

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

$$\vdots$$

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

ℓ_2-norm of a vector:

$$\|\epsilon\|_2 = \sqrt{\epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_d^2}$$

it follows that

$$\sum_{i=1}^{d} \epsilon_i^2 = \|\epsilon\|_2^2 = \epsilon^T \epsilon$$

ℓ_2-Loss: $$\widehat{w}_{LS} = \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - x_i^T w)^2 = \arg\min_w \|\mathbf{y} - \mathbf{X}w\|_2^2$$

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

The regression problem in matrix notation

$$f\left(\widehat{w}_{LS}\right) = \arg\min_{w \in \mathbb{R}^d} (\mathbf{y} - \mathbf{X}w)^T(\mathbf{y} - \mathbf{X}w)$$

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

$\gamma = w$

$\Omega = -X$

$\beta = y$

$$\nabla_w f = 2\Omega^T(\Omega\gamma + \beta)$$

$$= -2X^T(-Xw + y)$$

$$= 2X^TXw - 2X^Ty$$

$$\Rightarrow X^TXw = X^Ty$$

$$\Rightarrow w = (X^TX)^{-1}X^Ty$$

A few reminders on vector calculus

- Gradient of a function:

$$\nabla_w f(w) = \begin{bmatrix} \frac{df(w)}{dw_1} \\ \frac{df(w)}{dw_2} \\ \vdots \\ \frac{df(w)}{dw_d} \end{bmatrix}$$

$f(\gamma) = \gamma^T\gamma$

$\nabla_\gamma f = 2\gamma$

$f(\gamma) = (\Omega\gamma)^T(\Omega\gamma)$

$\Rightarrow \nabla_\gamma f = 2\Omega\Omega^T\gamma$

$f(\gamma) = (\Omega\gamma + \beta)^T(\Omega\gamma + \beta)$

- Example:

$f(w) = w^Tw \implies \nabla_W f(w) = 2w$

$f(w) = (Aw)^T(Aw) \implies \nabla_W f(w) = 2AA^Tw$

$f(w) = (Aw + b)^T(Aw + b) \implies \nabla_W f(w) = 2A^T(Aw + b)$

$\Rightarrow \nabla_\gamma f = 2\Omega^T(\Omega\gamma + \beta)$

The regression problem in matrix notation

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

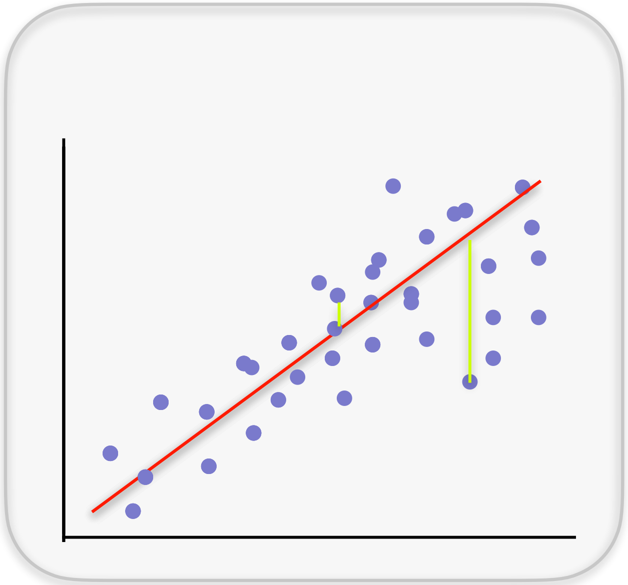
"Closed form" solution!

The regression problem in matrix notation

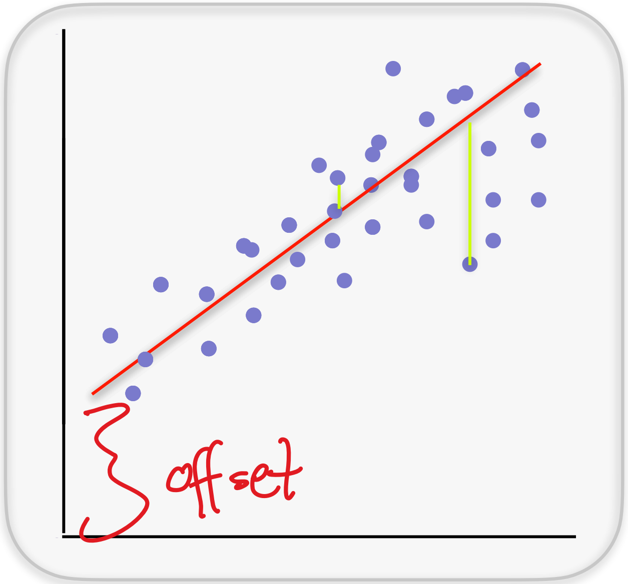
Linear model: $y_i = x_i^T w + \epsilon_i$

Least squares solution:

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

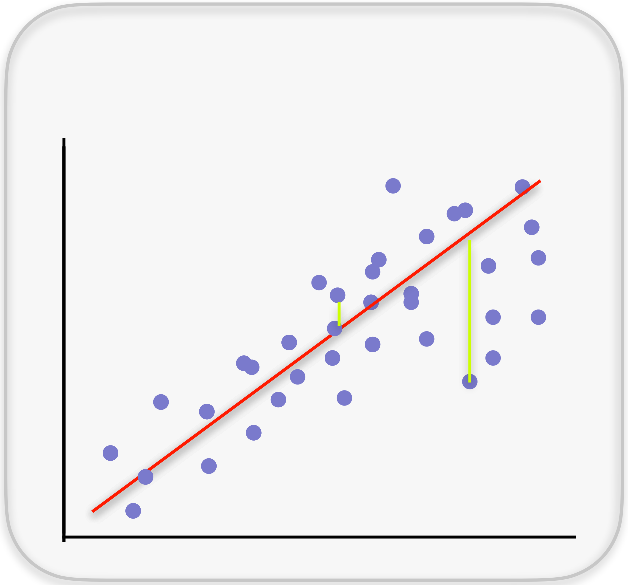


What about an offset (a.k.a intercept)?



The regression problem in matrix notation

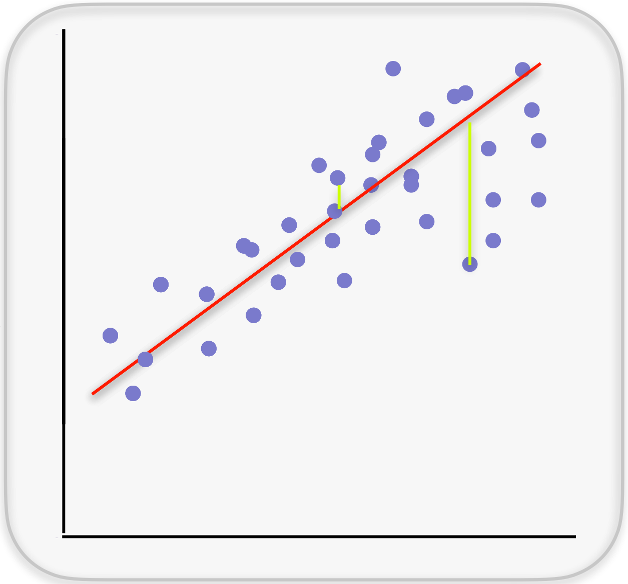
Linear model: $y_i = x_i^T w + \epsilon_i$



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:

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

$\mathbb{R}$

Dealing with an offset

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

$$= \arg \min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \underbrace{(\mathbf{y} - (\mathbf{X}w + \mathbf{1}b))^T(\mathbf{y} - (\mathbf{X}w + \mathbf{1}b))}_{\mathscr{L}(w,b)}$$

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

$\beta = y - 1b$

A reminder on vector calculus

$f(w) = (Aw + b)^T(Aw + b) \implies \nabla_W f(w) = 2A^T(Aw + b)$

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,

$$\widehat{w}_{\text{LS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$
$$\widehat{b}_{\text{LS}} = (\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

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 $\mathbf{X^T1} = 0$,

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

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

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

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 $\mathbf{X^T1} = 0$,

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

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

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

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

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

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

$$\widehat{b}_{\text{LS}} = \frac{1}{n}\sum_{i=1}^{n} y_i - \mu^T\widehat{w}_{\text{LS}}$$

Process for linear regression with intercept

Collect data: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$

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 x_{new}

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

Another way of dealing with an offset

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

reparametrize the problem as $\overline{\mathbf{X}} = [\mathbf{X}, \mathbf{1}]$ and $\overline{w} = \begin{bmatrix} w \\ b \end{bmatrix}$

$$\overline{\mathbf{X}}\,\overline{w} \;=$$

Why do we use least squares (i.e. ℓ_2-loss)?

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

$$\implies y_i \sim$$

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

Why do we use least squares (i.e. ℓ_2-loss)?

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

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

$\implies y_i \sim$

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

Why do we use least squares (i.e. ℓ_2-loss)?

Maximum Likelihood Estimator:

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

Why do we use least squares (i.e. ℓ_2-loss)?

Maximum Likelihood Estimator:

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

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

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

Recap of linear regression

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

Minimize the loss
(Empirical Risk Minimization)

Choose a loss
e.g., ℓ_2-loss: $(y_i - x_i^T w)^2$

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

Maximize the likelihood
(MLE)

Choose a Hypothesis class
e.g., $y_i = x_i^T w + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Maximize the likelihood,
$$\widehat{w}_{\text{MLE}} = \arg\max_w \left\{ -n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^n \frac{(y_i - x_i^T w)^2}{2\sigma^2} \right\}$$

Analysis of Error under additive Gaussian noise

Let's suppose $y_i = x_i^T w^* + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$, then this can be written as
$\mathbf{y} = \mathbf{X}w^* + \epsilon$

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

Maximum Likelihood Estimator is unbiased:

Analysis of Error under additive Gaussian noise

Let's suppose $y_i = x_i^T w^* + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$, then this can be written as
$\mathbf{y} = \mathbf{X}w^* + \epsilon$

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

Covariance is:

Analysis of Error under additive Gaussian noise

Let's suppose $y_i = x_i^T w^* + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$, then this can be written as $\mathbf{y} = \mathbf{X}w^* + \epsilon$, and the MLE is

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

This random estimate has the following distribution:

$$\mathbb{E}[\hat{w}_{\text{MLE}}] = w^*, \text{Cov}(\hat{w}_{\text{MLE}}) = \mathbb{E}[(\hat{w} - \mathbb{E}[\hat{w}])(\hat{w} - \mathbb{E}[\hat{w}])^T] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

$$\hat{w}_{\text{MLE}} \sim \mathcal{N}(w^*, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

Interpretation: consider an example with $\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ -1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}$

The covariance of the MLE, $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, captures how each sample gives information about the unknown w^*, but each sample gives information about for different (linear combination of) coordinates and of different quality/strength

Questions?