Lecture 3: Linear regression (continued)

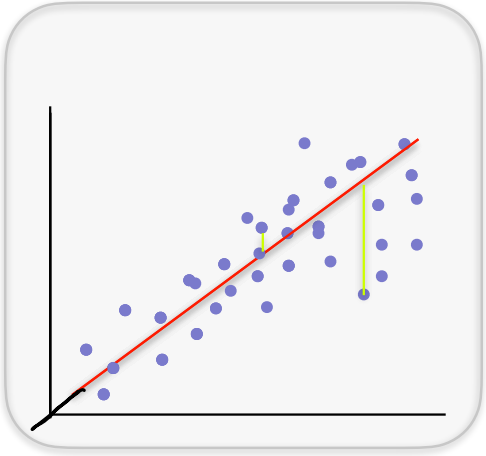
The regression problem in matrix notation

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

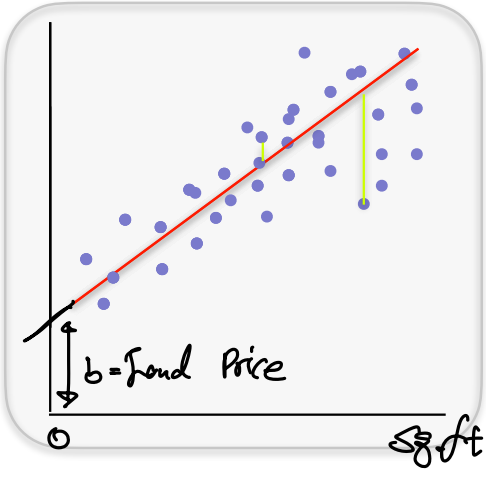
$x_i, w \in \mathbb{R}^d$

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)?



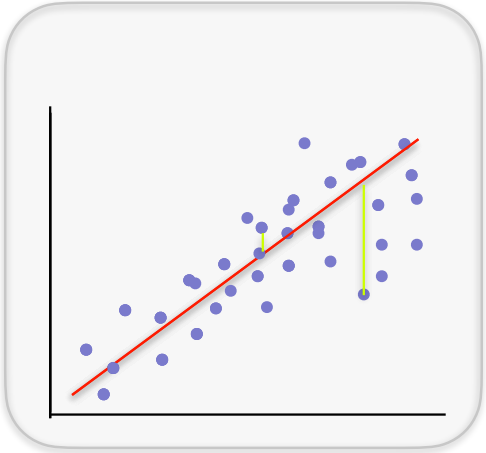
The regression problem in matrix notation

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

$x_i, w \in \mathbb{R}^d$

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

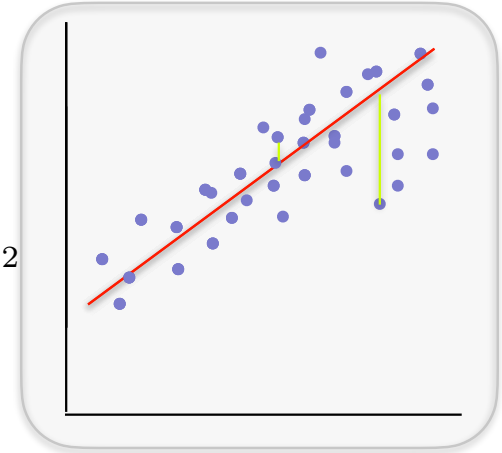
Intercept

$b \in \mathbb{R}$

Least squares solution:

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

$\varepsilon = y - Xw + 1 \cdot b$



Dealing with an offset

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

=

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

$\nabla_w \mathcal{L}(w,b) \stackrel{!}{=} 2X^T(y - Xw - 1 \cdot b) \quad \in \mathbb{R}^d$

$\longrightarrow X^T y = X^T X \hat{w} + X^T 1 \cdot \hat{b} \quad \in \mathbb{R}^d$

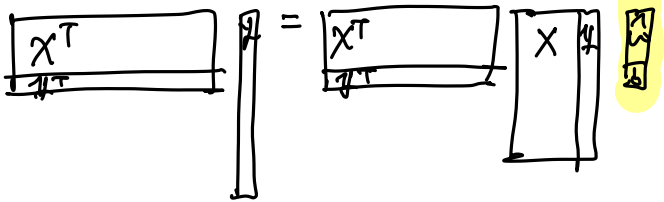
$\nabla_b \mathcal{L}(w,b) \stackrel{!}{=} 2 1^T (y - Xw - 1 \cdot b) \quad \in \mathbb{R}$

$\longrightarrow 1^T y = 1^T X \cdot \hat{w} + \underbrace{1^T 1}_{n} \cdot \hat{b} \quad \in \mathbb{R}$

$$\begin{bmatrix} X^T \\ 1^T \end{bmatrix} \cdot y = \begin{bmatrix} X^T \\ 1^T \end{bmatrix} \begin{bmatrix} X & 1 \end{bmatrix} \begin{bmatrix} \hat{w} \\ \hat{b} \end{bmatrix}$$

* Series of linear equations with
  - $d+1$-dim variables $(\hat{w}, \hat{b})$
  - $d+1$ equations.

- How do you solve this?



Schur Complement

$$\left(\begin{bmatrix} X^T \\ 1^T \end{bmatrix} \begin{bmatrix} X & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} X^T \\ 1^T \end{bmatrix} \cdot y = \begin{bmatrix} \hat{w} \\ \hat{b} \end{bmatrix}$$

Dealing with an offset

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

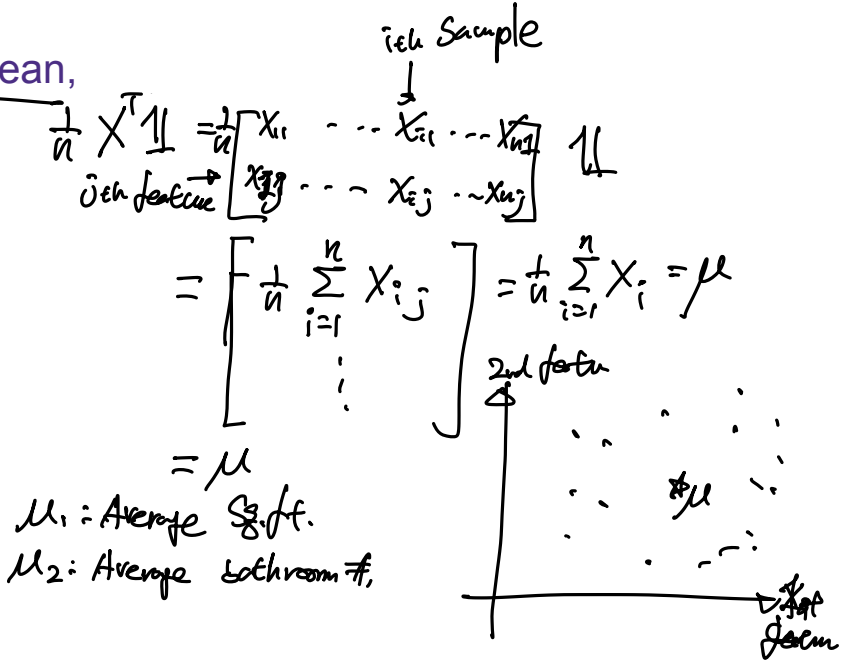
$$\mathbf{X}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{X}^T\mathbf{1} = \mathbf{X}^T\mathbf{y}$$

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

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

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

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

$$c = b + \mu^T w$$

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

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

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

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

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

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

$$\mathcal{L}(w,b) = ||y - Xw + \mathbf{1}\mu^T w - \mathbf{1}\mu^T w - \mathbf{1}b||_2^2$$

$$\mathcal{L}(w,c) = ||y - (X - \mathbf{1}\mu^T)w - \mathbf{1}\cdot c||_2^2$$

$(X - \mathbf{1}\mu^T)$: Zero-Mean $\widetilde{X}$

$\widetilde{X}, c \Rightarrow$

$$\hat{w}_{LS} = (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T y \quad \text{, where } \widetilde{X} = X - \mathbf{1}\mu^T$$

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

$$\hat{b}_{LS} = \hat{c}_{LS} - \mu^T\hat{w}_{LS} = \frac{1}{n}\sum_{i=1}^{n} y_i - \mu^T\cdot\hat{w}_{LS}$$

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

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

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

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

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

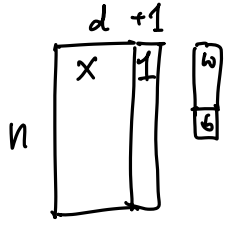
$$\hat{y}_{\text{new}} = x_{\text{new}}^T \hat{w}_{LS} + \hat{b}_{LS}$$

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 $\overline{w} = \begin{bmatrix} w \\ b \end{bmatrix}$

$\overline{\mathbf{X}}\,\overline{w} = [X \;\; 1] \cdot \begin{bmatrix} w \\ b \end{bmatrix} = X \cdot w + 1 \cdot b$



$\hat{\overline{w}}_{LS} = \arg\min \; ||y - \overline{x} \cdot \overline{w}||_2^2 = (\overline{x}^T \overline{x})^{-1} \cdot \overline{x}^T \cdot y$

$\begin{bmatrix} \hat{w} \\ \hat{b} \end{bmatrix} = \left(\begin{bmatrix} X^T \\ 1^T \end{bmatrix} [X \cdot 1] \right)^{-1} \cdot [X \cdot 1] \cdot y$

Why is least squares a good loss function?

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

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

$\implies y_i \sim \mathcal{N}(x_i^T w, \sigma^2)$

$\implies P(y_i; x_i, w, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(y_i - x_i^T w)^2}{2\sigma^2}}$

Why is least squares a good loss function?

Maximum Likelihood Estimator:

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

$= \arg\min \sum_{i=1}^{n} (y_i - x_i^T w)^2$ ← least squares problem.

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

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

* Gaussian noise $\Longleftrightarrow$ l_2-loss minimization.

Recap of linear regression

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

Minimize the loss (Empirical Risk Minimization)

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

there are other choices

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

Maximize the likelihood (MLE)

there are other hypotheses

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

matrix form

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

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

↑ Random

Maximum Likelihood Estimator is unbiased:

$\mathbb{E}[\hat{w}_{MLE}] = w^* + \mathbb{E}[(X^TX)^{-1}X^T\cdot\varepsilon]$ ← Linearity of Expectation

$= w^* + (X^TX)^{-1}\cdot X^T\cdot \underbrace{\mathbb{E}[\varepsilon]}_{0}$ ← $\mathbb{E}[a\cdot\varepsilon + b] = a\,\mathbb{E}[\varepsilon] + b$

$= w^*$

Analysis of Error under additive Gaussian noise

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

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

Covariance is: $\mathbb{E}[(\hat{w} - w^*)(\hat{w} - w^*)^T]$

$= \mathbb{E}[(X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1}]$

linearity of expectation

$= (X^TX)^{-1}X^T\,\mathbb{E}[\varepsilon\varepsilon^T]\,X(X^TX)^{-1}$

$= (X^TX)^{-1}X^T\sigma^2\mathbb{I}\cdot X(X^TX)^{-1}$

$= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1}$

$= \sigma^2(X^TX)^{-1}$

$\mathbb{E}[\varepsilon\varepsilon^T] = \Sigma$

$\Sigma_{ij} = \mathbb{E}[\varepsilon_i\varepsilon_j] = \begin{cases} \sigma^2 & \text{if } i=j \\ 0 & \text{else} \end{cases}$

$\Sigma = \begin{bmatrix} \sigma^2 & & 0 \\ & \sigma^2 & \\ 0 & & \ddots \end{bmatrix} = \sigma^2\mathbb{I}.$

Analysis of Error under additive Gaussian noise

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

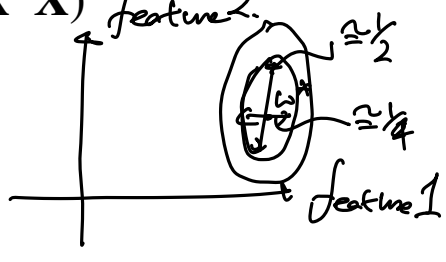
$$\begin{aligned}
\widehat{w}_{MLE} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}w + \epsilon) \\
&= w + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon
\end{aligned}$$

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

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

$\hat{w}_{\text{MLE}} \sim \mathcal{N}(w, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) = \mathcal{N}(w^*, \sigma^2 \begin{bmatrix} 1/4 & 0 \\ 0 & 1/2 \end{bmatrix})$

If $X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ -1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}$ then $X^TX = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}$ $(X^TX)^{-1} = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/2 \end{bmatrix}$



Questions?