*Announcement — Gradescope submission.
└ Code submission. .py

		A	B
HW0_A	← .pdf	✓	✓
HW0_B	← .pdf		✓
HW0_code	← .py	✓	✓

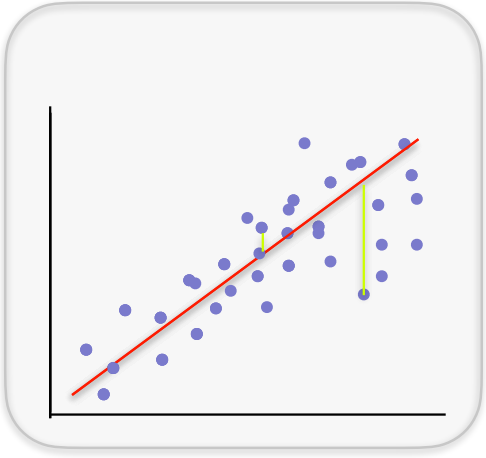
Lecture 3

The regression problem in matrix notation

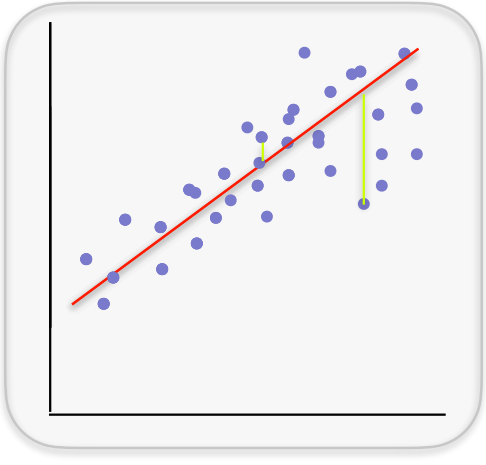
Linear model: $y_i = x_i^T w + \epsilon_i$

Least squares solution:

$$\mathbb{R}^d \ni \hat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$



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

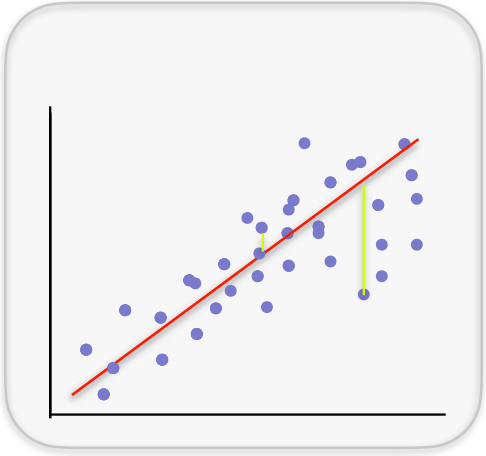


The regression problem in matrix notation

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

Least squares solution:

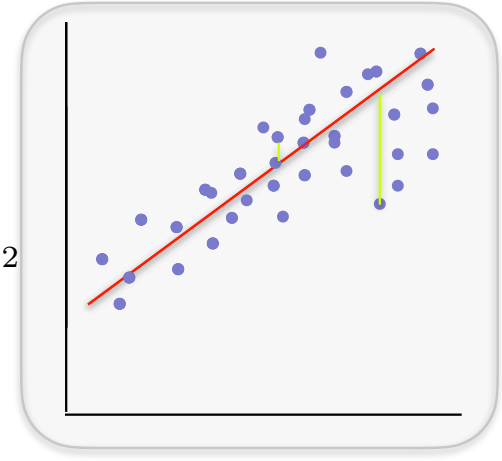
$$\begin{aligned}\widehat{w}_{LS} &= \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2 \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\end{aligned}$$



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

Least squares solution:

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



Dealing with an offset

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

vector $\begin{bmatrix} \\ \\ \\ \end{bmatrix}\}n$

$\mathbf{1} \in \mathbb{R}^n$ ← Quad fun. w, b.

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

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

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

$$\nabla_w \mathcal{L}(w,b) = 2 \cdot X^T (y - (Xw + \mathbf{1}b)) \Big|_{\hat{w},\hat{b}} = 0$$

$\nabla_w \mathcal{L}(w,b) \in \mathbb{R}^d$

$d \begin{bmatrix} X^T \\ n \end{bmatrix} \left(\begin{bmatrix} y \end{bmatrix} - \left(\begin{bmatrix} X \end{bmatrix} \begin{bmatrix} \hat{w} \end{bmatrix} + \begin{bmatrix} \mathbf{1}\hat{b} \end{bmatrix} \right) \right)$

$$\nabla_b \mathcal{L}(w,b) = 2 \cdot \mathbf{1}^T (y - (Xw + \mathbf{1}b)) \Big|_{\hat{w},\hat{b}} = 0$$

$\nabla_b \mathcal{L}(w,b) \in \mathbb{R}$

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

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

linear in $\hat{w}, \hat{b}$

$\hat{w} \in \mathbb{R}^d$, $\hat{b} \in \mathbb{R}$

$d+1$ variables

$d+1$ equations

Fact: $\nabla_w \left((Aw+b)^T (Aw+b) \right) = 2A^T (Aw+b)$

$(aw+b)^2 = 2a(aw+b)$

$\nabla_w (w^T w) = 2w$,

$\nabla_w (w^T A^T A w) =$ →

Dealing with an offset

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

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

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

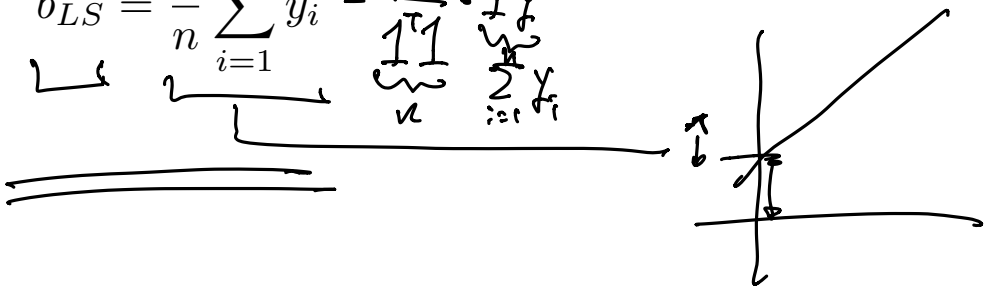
$$\widehat{b}\cdot n = \sum_{i=1}^{n} y_i$$

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

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

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

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



Dealing with an offset

$$\mu_j \triangleq \frac{1}{n}\sum_{i=1}^{n} x_{i,j}$$

μ_1: average sq.ft.
μ_2: avg # bathrooms

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

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

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

$c = \mu^T w + b$

$\mu = \frac{1}{n}X^T\mathbb{1} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i,1} \\ \vdots \end{bmatrix}$

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

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

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

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

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

$$= \| y - \underbrace{\tilde{X}}_{\text{centered}} \cdot w - \mathbb{1}\cdot c \|_2^2,$$

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

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

$\hat{c}_{LS} = \hat{b}_{LS} + \mu^T\hat{w}_{LS}$

$\tilde{X}$, c : new

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

$$\overline{\mathbf{X}}\,\overline{w} = \begin{bmatrix} & \overline{X} & \\ & \vdots & \\ & \vdots & \end{bmatrix} \cdot \begin{bmatrix} \bar{w} \\ \vdots \\ b \end{bmatrix} \begin{matrix} \} d \\ \} 1 \end{matrix} = X \cdot w + \mathbf{1} \cdot b$$

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

$d+1$ $\left\{ \begin{bmatrix} \\ \\ \end{bmatrix} \begin{matrix} \} \bar{w} \\ \} b \end{matrix} \right.$

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 $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ ← Generative Model

probabilistic

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

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

Why is least squares a good loss function?

Maximum Likelihood Estimator:

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

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

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

Recap of linear regression

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

Minimize the loss
(Empirical Risk Minimization)

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

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

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

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

$\epsilon \sim N(0, \sigma^2)$, n; ϵ → Random

Maximum Likelihood Estimator is unbiased:

$$\text{Bias}(\hat{W}_{MLE}) = \mathbb{E}[\hat{W}_{MLE}] - W$$

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

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

$(X^TX^{-1}) \to 0$, $\mathbb{E}[\varepsilon] = 0$

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

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

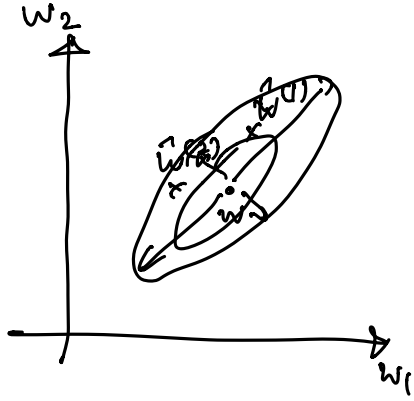
$= (X^TX)^{-1}X^T \underbrace{\mathbb{E}[\varepsilon\varepsilon^T]}_{\sigma^2 \cdot I} \cdot X(X^TX)^{-1} = \sigma^2 \cdot (X^TX)^{-1}\underbrace{X^TX(X^TX)^{-1}}_{\square \cdot \square^{-1} = I}$

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

Analysis of Error under additive Gaussian noise

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

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



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

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

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

Questions?

① What if $X^T X$ has rank $< d$.

$$v^T (X^T X) \cdot v = 0.$$

$$\hat{w}_{MLE} = (X^T X)^{\dagger} \cdot X^T \cdot y + c \cdot v$$

$\hat{w}_{MLE}^{(0)}$ → all in this line achieve min loss.