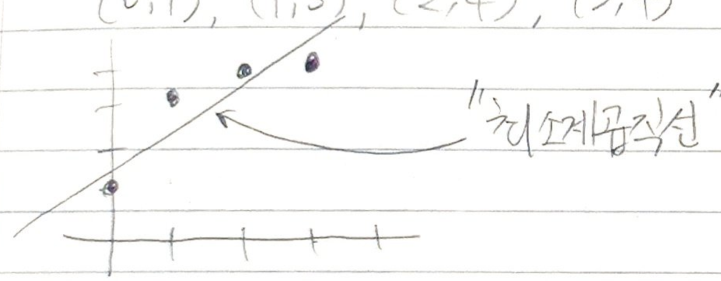
Week 5. 정사영 (projection) 과 최소제곱문제 (least square problem)

- 주어진 데이터에 대한 적절한 모델을 찾는 문제는 선형연립방정식으로 귀결된다.
- 측정 오차를 줄이기 위해 대개 미지수 개수보다 많은 데이터를 사용하여 방정식의 수가 미지수보다 많은 선형연립방정식이 생기게 된다
- 이런 경우 일반적으로 해가 존재하지 않는다
  - → 정사영 개념을 이용하여 최소제곱문제로 변환
  - → 최소제곱해를 최적해로서 구한다
  - → 데이터에 적합한 곡선을 찾는다

5.1 최소제곱문제

x, y에 대한 오차된 데이터

(0,1), (1,3), (2,4), (3,4)



[이상적 상황] 모든 데이터 (x_i, y_i)에 대해서 $y_i = a + bx_i$ 가 만족되는 y절편 a 와 기울기 b를 구하는 것

데이터 (x,y)	→	일차함수 $y=a+bx$	선형연립방정식	행렬표현
(0,1)		$1=a+b\cdot 0$	$a = 1$	
(1,3)		$3=a+b\cdot 1$	$a+b=3$	
(2,4)		$4=a+b\cdot 2$	$a+2b=4$	
(3,4)		$4=a+b\cdot 3$	$a+3b=4$	

$$\left\{\begin{aligned} a &= 1 \\ a+b &= 3 \\ a+2b &= 4 \\ a+3b &= 4 \end{aligned}\right.
\qquad
\underbrace{\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}}_{A}
\underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{u}
=
\underbrace{\begin{bmatrix} 1 \\ 3 \\ 4 \\ 4 \end{bmatrix}}_{y}$$

$$\boxed{Au = y}$$

선형연립방정식을 만족하는 해 ($Au=y$인 u)를 기대할 수 없음

→ Au 와 y 사이의 거리 ($dist(Au, y) = \|Au-y\|$)가 최소가 되는 근사해, 즉 다음을 만족하는 $\hat{u}$을 찾는 문제 = $\boxed{\text{최소제곱문제}}$

$$\boxed{\min \|Au-y\|}$$

5.2 최소제곱문제의 의미

- 각 데이터 (x_i, y_i)에 대하여 x_i를 일차함수 $y=a+bx$에 대입하여 얻은 값을 $\hat{y}_i$라 하자 (즉 $\hat{y}_i = a+bx_i$)
- 선형연립방정식의 해가 존재하지 않는 경우는 y_i와 $\hat{y}_i$가 같지 않아서 발생하므로, 차선책으로 $(y_i-\hat{y}_i)^2$이 최소가 되는 a, b를 구한다.
- 주어진 모든 데이터에 대하여 오차(error)를 합하면 다음을 얻는다.

$$\boxed{E(u) = E(a,b) = (a-1)^2+(a+b-3)^2+(a+2b-4)^2+(a+3b-4)^2 = \|Au-y\|^2}$$

따라서 최소제곱문제는 아래와 같이 오차를 최소화하는 최적화(optimization) 문제로 이해할 수 있다.

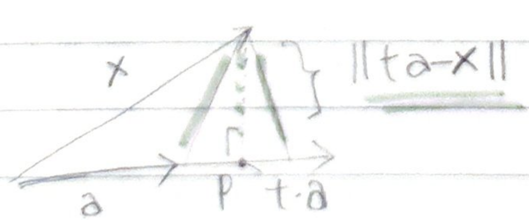
$$\langle \min E(u) \rangle$$

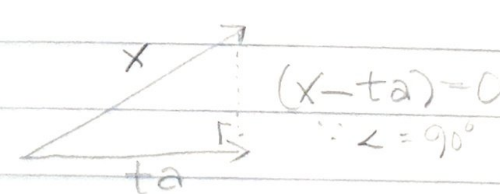
5.3 정사영과 최소제곱해 (least square solution)

최소제곱문제를 풀기 위해 정사영(projection)에 관하여 알아야 한다.
먼저 다음 식을 만족하는 해 t를 찾는 문제를 보자.

$$\min \|ta - x\| \quad (\text{여기서 } t\text{는 실수})$$

이 문제는 시작점이 같은 두 벡터 a와 x에 대하여,
a를 포함하는 직선과 x 사이의 거리가 최소가 되게 하는 t를 찾으려고 하는 것이다.





ta는 a와 평행이므로, a를 포함하는 직선 위에 놓이게 되고, $\|ta-x\|$는 실수 t에 의해 실선으로 표시된 선분의 길이를 나타낸다. 직관적으로 $\|ta-x\|$ 최소는 x의 끝점에서 a를 포함하는 직선 위에 수선을 내려 생기는 벡터 $p=ta$임을 알 수 있다.
이 때의 t가 $\min\|ta-x\|$의 해가 되며, 벡터 p를 x 위로의 a의 정사영이라고 한다.

$(x-ta) \perp a$ 이므로, $a\cdot(x-ta)=0 \Rightarrow t(a\cdot a)=a\cdot x \Rightarrow \boxed{t = \dfrac{a\cdot x}{a\cdot a} = (a^Ta)^{-1}a^Tx}$

5.4 데이터에 적합한 곡선찾기 (Curve Fitting)

이상의 방법으로 (x, y)에 대한 best fit 이차 근사식도 찾을 수 있다.
즉 다음과 같이 미지수가 a, b, c인 선형연립방정식과 행렬표현을 얻게된다.

(x, y)	$\longrightarrow$ $y = a + bx + cx^2$	선형연립방정식
$(0,1)$	$1 = a + b\cdot 0 + c\cdot 0$	$a = 1$
$(1,3)$	$3 = a + b\cdot 1 + c\cdot 1$	$a + b + c = 3$
$(2,4)$	$4 = a + b\cdot 2 + c\cdot 4$	$a + 2b + 4c = 4$
$(3,4)$	$4 = a + b\cdot 3 + c\cdot 9$	$a + 3b + 9c = 4$

행렬표현

$$\underset{A}{\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}} \underset{u}{\begin{bmatrix} a \\ b \\ c \end{bmatrix}} = \underset{y}{\begin{bmatrix} 1 \\ 3 \\ 4 \\ 4 \end{bmatrix}}$$

$$Au = y$$

$$\min \| Au - y \|$$

$$E(u) = E(a,b,c) = (a-1)^2 + (a+b+c-3)^2 + (a+2b+4c-4)^2 + (a+3b+9c-4)^2$$
$$= \| Au - y \|^2$$

$$\hat{u} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} = (A^TA)^{-1}Ay = \begin{bmatrix} 1 \\ 5/2 \\ -1/2 \end{bmatrix}$$

$\therefore$ $(0,1), (1,3), (2,4), (3,4)$에 대한 최소제곱곡선 (least square curve)

$$\Rightarrow \boxed{y = 1 + \frac{5}{2}x - \frac{1}{2}x^2}$$