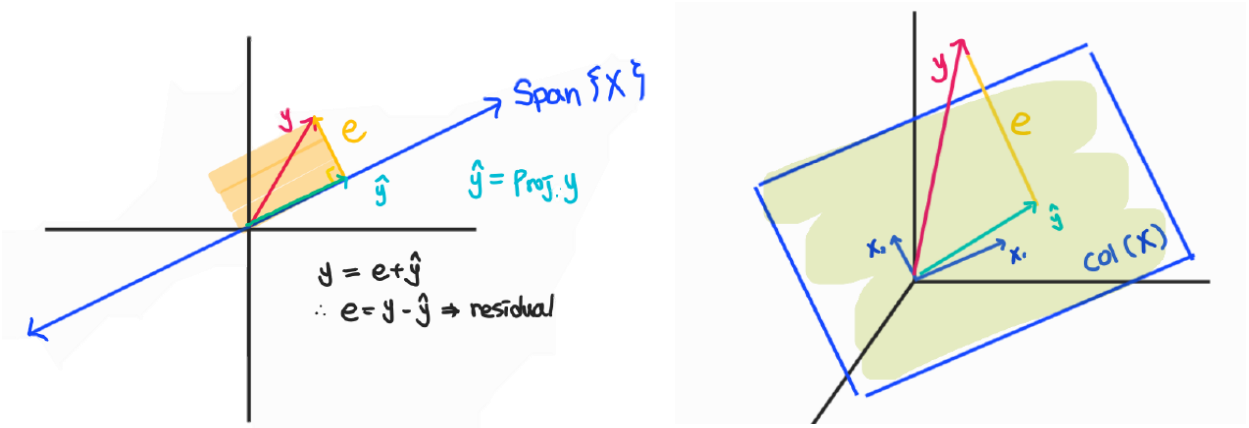
[내용추가 + 수정]

🌞 회귀와 투영벡터 개념 다시 정리할게요!

투영벡터 : $proj_x y$ = 벡터 y를 span{x}에 정사영함 = span{x} 위에 벡터 y와의 거리가 최소인 간접해$\hat{y}$ 을 찾음

= 오차e ($y-\hat{y}$)와 span{x}는 직교 = e• x가 0

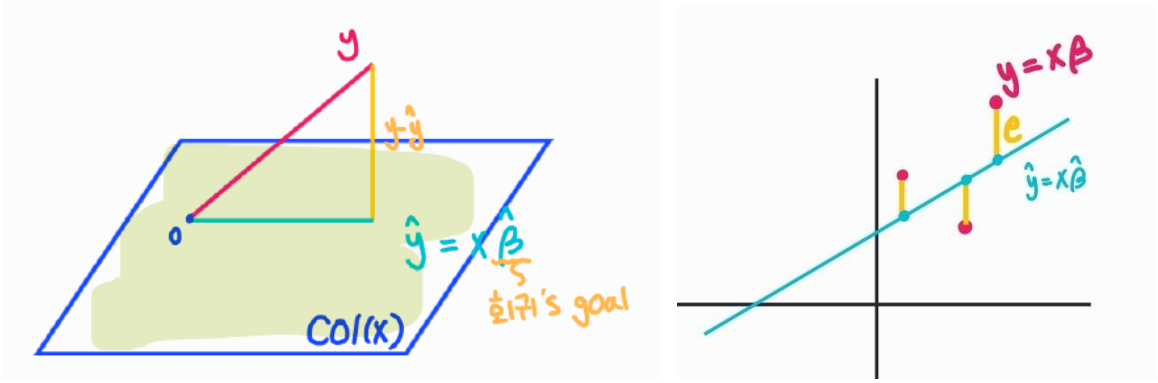


$x\beta$=y 의 해가 없다 ⇔ y가 span{x}위에 없다. ⇔ y가 col(x)위에 없다

LSM : $X\beta$=y 해가 없는 경우 $x\beta$ 와 y의 최소거리를 바탕으로 간접해 $\hat{y}$을 구하는 것

선형회귀분석 : $X\beta$=y -- Least-Square problem -- > $X\hat{\beta} = \hat{y}$

-- > 간접해 $\hat{\beta}$ 찾기 = 예측값과 실제값의 거리(residual)를 최소화하는 회귀선의 회귀계수 찾기!



회귀분석은 y를 projection ($\hat{y}$)시켜 최종적으로는 최적(잔차제곱의 최소)의 베타계수를 찾는 것!!

꿩 대신 닭 ⇔ proj ⇔ regression

🌼 R 의 내장데이터 Carseats 을 이용해 Age(x)로 Sales(y)를 예측하는 회귀모델링을 해보자!

Sales	Age
9.50	42
11.22	65
10.06	59
7.40	55
4.15	38

선형회귀모델

$$X = \begin{pmatrix} 1 & 42 \\ 1 & 65 \\ 1 & 59 \\ 1 & 55 \\ 1 & 38 \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad y = \begin{pmatrix} 9.50 \\ 11.22 \\ 10.06 \\ 7.40 \\ 4.15 \end{pmatrix}$$

$$X^T X\hat{\beta} = X^T y$$

$$\therefore \hat{\beta} = \begin{pmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \end{pmatrix} = \begin{pmatrix} -0.83 \\ 0.18 \end{pmatrix}$$

가중선형회귀모델'

나이가 50이상인 데이터에 2배 가중치를 둘까?

$$X = \begin{pmatrix} 1 & 42 \\ 1 & 65 \\ 1 & 59 \\ 1 & 55 \\ 1 & 38 \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad y = \begin{pmatrix} 9.50 \\ 11.22 \\ 10.06 \\ 7.40 \\ 4.15 \end{pmatrix} \quad W = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$(WX)^T WX\hat{\beta} = (WX)^T Wy$$

$$WX = \begin{pmatrix} 1 & 42 \\ 2 & 130 \\ 2 & 118 \\ 2 & 110 \\ 1 & 38 \end{pmatrix} \quad Wy = \begin{pmatrix} 9.50 \\ 22.44 \\ 20.12 \\ 14.80 \\ 4.15 \end{pmatrix} \quad \therefore \hat{\beta} = \begin{pmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \end{pmatrix} = \begin{pmatrix} -2.37 \\ 0.20 \end{pmatrix}$$

🌼 $X^T X\hat{\beta} = X^T y$ 에서 t(x)x 가 invertible 하지 않으면 어떻게 하지?

사실 non-invertible한 경우가 많지 않을 뿐만 아니라, 프로그램에서는 역행렬 존재의 유무와 관계없이 계산을 해줍니다! 하지만! 만약을 위해 설명합니다. $X^T X$ 가 invertible 하지 않은 이유는 크게 두가지로 나뉩니다. (1) 불필요한(중복) feature 을 가진 경우, (2) 과도하게 많은 feature 을 가진 경우. 각각 살펴봅시다

(1) 불필요한(중복) feature 을 가진 경우

x1 : 키, 단위 cm
x2 : 신장, 단위 m
x3 : 키, 단위 feet

위의 예시처럼 같은 의미를 갖는 데이터를 사용하는 경우, 중복된 features 문제가 발생합니다. 이것이 원인인 경우는 중복된 feature 를 찾아 제거하면 해결됩니다.

(2) 과도하게 많은 feature 을 가진 경우

주어진 data set 의 크기에 비해 너무 많은 feature 를 사용하는 경우 발생합니다. 예를 들어 주어진 데이터는 10 개인데, 각 데이터의 feature 은 100 개다? 이럴 때는 (a) 일부 features 를 삭제해 줄이는 방법 (b) regularization(정규화) 하는 방법 두가지의 방법이 있습니다.

이때 정규화란 회귀계수에 제약을 가하는 것으로, 과적합(overfitting) 문제를 막기 위해 주로 사용하는 방법입니다. 구체적으로 회귀계수에 어떻게 패널티를 주는지에 따라 Ridge, Rasso, ElasticNet 으로 구분하지만, 그것은 회귀 클린업에서 다뤄야하는 일!! 선대에서는 여기까지만 하겠습니다.

🌻 (1,2),(2,2),(3,2) 데이터 라더니 어쩌다 행렬이 저렇게 정리된거죠?

$$\underset{A}{\begin{bmatrix}1 & 1\\ 2 & 1\\ 3 & 1\end{bmatrix}}\underset{x}{\begin{bmatrix}c\\ d\end{bmatrix}}=\underset{b}{\begin{bmatrix}1\\ 2\\ 2\end{bmatrix}}$$

왼쪽 그림은 b=ct+d 라는 식으로 정리했을 때인데,
맥락설명 없이 캡쳐이미지를 그래도 쓴 제 잘못이 있으니 다시 정리하겠습니다.

"ax+b=y"
1a+b=1
2a+b=2
3a+b=2

$$= a\begin{bmatrix}1\\ 2\\ 3\end{bmatrix}+b\begin{bmatrix}1\\ 1\\ 1\end{bmatrix}=\begin{bmatrix}1\\ 2\\ 2\end{bmatrix} = \begin{bmatrix}1 & 1\\ 1 & 2\\ 1 & 3\end{bmatrix}\begin{bmatrix}a\\ b\end{bmatrix}=\begin{bmatrix}1\\ 2\\ 2\end{bmatrix} \Rightarrow \begin{bmatrix}1 & 1\\ 1 & 2\\ 1 & 3\end{bmatrix}\begin{bmatrix}\hat{a}\\ \hat{b}\end{bmatrix}=\begin{bmatrix}\widehat{y1}\\ \widehat{y2}\\ \widehat{y3}\end{bmatrix}$$

데이터를 Ax=y 형태로 정리한 뒤, 해를 구할 수 있는 $A\hat{x}=\hat{y}$으로 바꾸는 과정

🌻Ax=y 와 y=Xβ에 모두 x,y 가 들어가서 혼란스러워요
Ax=y : 해(x)가 존재하지 않는다 = y 가 Ax 위에 없다 → $A\hat{x}=\hat{y}$; 이때 A 는 선형변환 행렬
Xβ=y : y 가 Xβ위에 없다 → $X\hat{\beta}=\hat{y}$; 이때 x 는 [1 데이터의 독립변수 데이터 모음]
y 예측(투영,근접해)은 공통이고, 위의 x 와 아래의 β(구하고싶은거) / 위의 A 와 아래의 X(주어진거) 가 같은 맥락이다.

🌻등분산성가정이 만족하지 않을 때는?

(1) log(Y), 1/Y, sqrt(Y)등의 변수변환을 하거나 (2)가중치행렬로 등분산성 가정을 처리하는데, 실제로 가중치행렬을 이용할 때는 분산의 역수를 사용하는 것이 일반적입니다~ 다만! 분산(σ^2)을 알기 어렵기 때문에 경험적으로 정해야 하는데, 자세한 내용은 회귀 클린업~