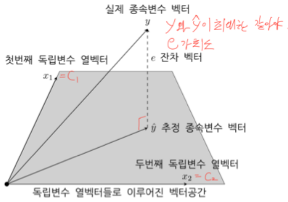
In [ ]:

05.02 회귀분석의 기하학

1) 회귀 벡터공간
- y_hat = Xw
- y_hat 은 X의 각 열을 기저벡터로 하는 벡터공간 내 존재(span of X)
- y_hat = y를 X가 이루는 벡터공간에 투영한 벡터 = Hy (H = 투영행렬)
- e = My (M = 잔차행렬) (1page)

선형 회귀분석으로 예측한 값 $\hat{y}$는 X의 각 열 $c_1, \cdots, c_M$의 선형조합으로 표현된다.

$$\hat{y} = Xw = \begin{bmatrix} \vec{c_1} & \cdots & \vec{c_M} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_M \end{bmatrix} = w_1 \vec{c_1} + \cdots + w_M \vec{c_M}$$



모든 열이 선형독립이면 예측값 $\hat{y}$는 X의 각 열 $c_1, \cdots, c_M$을 기저벡터(basis vector)로 하는 벡터공간(vector space)위에 존재한다는 것을 알 수 있다.

2) 잔차행렬과 투영행렬(=햇행렬, 영향도행렬)
- y를 각각 잔차와 예측값으로 변환하는 행렬
- 잔차행렬과 투영행렬의 성질(4page) (꼭 암기)

잔차 행렬과 투영 행렬은 다음과 같은 성질이 있다.

(1) 대칭행렬이다.

$$M^T = M$$
$$H^T = H$$

(2) 곱해도 자기 자신이 되는 행렬이다. 이러한 행렬을 **멱등(idempotent)행렬**이라고 한다. 멱등행렬은 몇번을 곱해도 자기 자신이 된다.

$$M^k = M^3 = M^2 = M$$
$$H^k = H^3 = H^2 = H$$

(3) M과 H는 서로 직교한다.

$$MH = HM = 0$$

(4) M은 X와 직교한다.

$$MX = 0$$

(5) X에 H를 곱해도 변하지 않는다.

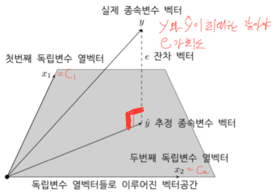
$$HX = X$$

위 성질은 다음과 같이 증명한다.

- y벡터의 제곱합 = 잔차 벡터 e제곱합 + 추정치 벡터 y_hat제곱합 (6page)
  (증명해보기)

위 성질들을 이용하면 y 벡터의 제곱합은 잔차 벡터 e의 제곱합과 추정치 벡터 $\hat{y}$의 제곱합의 합이라는 것을 알 수 있다.

$$\begin{aligned} y &= \hat{y} + e = Hy + My = (H + M)y \\ y^T y &= ((H + M)y)^T((H + M)y) \\ &= y^T(H + M)^T(H + M)y \\ &= y^T(H + M)(H + M)y \\ &= y^T(H^2 + MH + HM + M^2)y \\ &= y^T(H + M)y \\ &= y^T Hy + y^T My \\ &= y^T H^2 y + y^T M^2 y \\ &= y^T H^T Hy + y^T M^T My \\ &= (Hy)^T(Hy) + (My)^T(My) \\ &= \hat{y}^T \hat{y} + e^T e \end{aligned}$$



이 관계식은 나중에 분산 분석(ANOVA)에 사용된다.

05.01 확률론적 선형 회귀모형

4) 잔차의 분포
- 확률론적 선형회귀모형 : "잔차 = e = y-w.Tx 도 정규분포따름"
- 확률론적 선형회귀모형에서는 잔차와 잡음이 다른 개념이다.
- 잡음이 정규분포이면(확률론적 선형 회귀모형 하), 잔차도 정규분포를 따른다. (7page 증명!)
- 잔차는 잡음의 선형변환. 따라서, 잡음의 가정들이 잔차에도 적용됨
  ex) 잔차의 기대값 = 0

확률론적 선형 회귀모형에 따르면 회귀분석에서 생기는 **잔차** $e = y - \hat{w}^T x$ **도 정규 분포를 따른다.** 다음과 같이 증명할 수 있다.

<확률론적 선형회귀 에서는 4번만>
$W^T x + \varepsilon$ ⇐ 1. $w \neq \hat{w}$ (정답은 있지만 예측값과 다르다)
실제 실제!
2. $\varepsilon \neq e$
disturbance (입실론) residual
$e = y - \hat{y}$
↳ 예측 기반
* $\varepsilon \Rightarrow W^T x + \varepsilon$
실제기반!

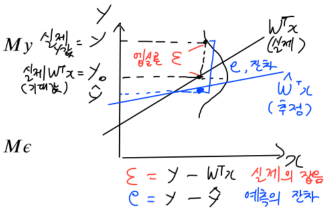
확률론적 선형 회귀모형의 잡음 ϵ와 잔차 e는 다음과 같은 관계를 가진다.

$$\hat{y} = X\hat{w} = X(X^TX)^{-1}X^Ty = Hy$$

이 행렬 H은 **Hat 행렬** 혹은 **프로젝션(projection) 행렬** 또는 **영향도(influence) 행렬**이라고 부르는 대칭 행렬이다.

Hat 행렬을 이용하면 잔차는 다음처럼 표현된다.

$$e = y - \hat{y} = y - Hy = (I - H)y = My$$



이 행렬 M은 **잔차(residual) 행렬**이라고 부른다.

확률적 선형 회귀 모형의 가정을 적용하면,

$$e = My = M(Xw + \epsilon) = MXw + M\epsilon$$

그런데

$$MX = 0$$

에서

$$e = M\epsilon$$

즉, **잔차 e는 잡음 ϵ의 선형 변환(linear transform)**이다. 정규 분포의 선형 변환은 마찬가지로 정규 분포이므로 잔차도 정규 분포를 따른다.

5) 회귀계수의 표준오차
- 다시 처음으로, 확률론적 선형회귀모형 쓰는 이유는? = "우리가 예측한 것의 오차는?" 에 답하기 위해
- "우리가 예측한 것의 오차" = 회귀계수의 표준오차(se, Standard Error)
- 증명은 나중에 시간 되면..

[se] (9page 필기)
정규화된 모수(w) 오차 ==>> 표준스튜던트t분포를 따름 (자유도 N-K)
*N = 표본 데이터 수, K = 가중치 갯수(0~K-1 까지 가중치)

0. "$E[\hat{w}] = w$", 확률론적 선형회귀모델 가정 하 MLE 추정결과, $\hat{w}$는, 추정은, 비편향 추정치이다. (실제로 본래 값 w 주변에서 추정치가 $\hat{w}$ 분포된다) 치우치지 않고!

그리고, 추정치는 w(실제) 근처이지만, 결국 오차를 갖는다.

회귀계수 표준오차 ($\hat{w}_i$) ⇔ 1. $\hat{w}_i$ 추정이 얼마나 흔들리는지 (분산이 어느정도인지) 알아보자!

2. $Cov[\hat{w}_i]$ ⇒ 공분산행렬화
⇓
대각성분 $Cov[\hat{w}_i]_{ii}$
$= Var[\hat{w}_i]$

* se_i = 표준오차 = $\sqrt{Var[\hat{w}_i]}$

3. $\hat{w}_i$의 분산 알았으니, 추정치가 ($\hat{w}_i$) 실제와 (w_i) 얼마나 다른지 (오차) 살펴보자 (추정치와 실제가 너무 다르면 추정치 쓰면안됨)
↳ 오차 정규화해 $H_0 : w = \hat{w}$ 검증

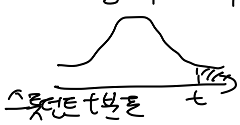
4. 각 오차들의 표준적인 비교를 위해 정규화 진행! (그래야 상대적 비교 용이)

5. 정규화한 모수오차 ⇒ 알고보니 스튜던트 t분포 따른다.

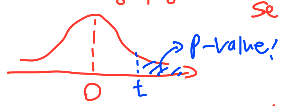
$$\frac{\hat{w}_i - w_i}{se_i} \sim t_{N-K}$$
(자유도: N - K)
표본갯수 가중치갯수

모수(추정치)오차를 정규화한 값 ≒ "t 통계량" (자유도 N-K인 스튜던트 t분포 따르는)

6. t 통계량 ↑ = 모수 오차가 크다!



* 특히, $H_0 : w = 0$ 가설검정시 유용. t통계량 = $\frac{\hat{w}_i - 0}{se_i}$



⇒ t 통계량 ↑라면, H_0 가설 하 t 통계량의 분포 하, $\hat{w}_i$는 굉장히 드문숫자라 오차가 크게 나타남!
∴ H_0는 기각!