회귀분석의 기하학

회귀 벡터공간

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

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

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.

실제 종속변수 데이터 y와 예측값 $\hat{y}$의 차이가 잔차 벡터 e이다. 따라서 잔차 벡터 e의 크기를 가장 작게 만드는 최적의 예측값 $\hat{y}$는 벡터공간내에 존재하면서 y와 가장 가까운 벡터이다. 이 때 잔차 벡터 e는 벡터 공간에 직교한다. 따라서 예측값 벡터 $\hat{y}$는 y를 X의 각 열 $c_1, \cdots, c_M$을 기저벡터로 하는 벡터공간에 투영(projection)한 벡터이고 잔차 벡터 e는 투영하고 남은 직교 벡터이다.

그림 2.2.1 : 회귀 벡터공간

잔차행렬과 투영행렬

벡터 a에서 다른 벡터 b를 변형하는 과정은 변형행렬(transforma matrix) T를 곱하는 연산으로 나타낼 수 있다.

$$b = Ta$$

종속값 벡터 y를 잔차 벡터 e로 변형하는 변환 행렬 M를 정의하자. 이 행렬을 **잔차행렬(residual matrix)**이라고 한다.

$$e = My$$

종속값 벡터 y를 예측값 벡터 $\hat{y}$로 변형하는 변환 행렬 H를 정의하자.. 이 행렬을 **투영행렬(projection matrix)**이라고 한다.

$$\hat{y} = Hy$$

잔차행렬은 다음과 같이 구한다.

$$
\begin{aligned}
e &= y - \hat{y} \\
&= y - Xw \\
&= y - X(X^T X)^{-1} X^T y \\
&= (I - X(X^T X)^{-1} X^T) y \\
&= My
\end{aligned}
$$

투영행렬은 다음과 같이 구한다.

$$\begin{aligned}\hat{y} &= y - e \\ &= y - My \\ &= (I - M)y \\ &= X(X^TX)^{-1}X^Ty \\ &= Hy\end{aligned}$$

따라서 M, H는 각각 다음과 같다.

$$H = X(X^TX)^{-1}X^T$$

$$M = I - X(X^TX)^{-1}X^T$$

투영 행렬은 y로부터 ^기호가 붙은 $\hat{y}$를 계산한다고 해서 **햇(hat) 행렬** 또는 **영향도 행렬(influence matrix)**이라고 부르기도 한다. 영향도 행렬이라는 명칭의 의미는 아웃라이어 분석에서 다시 다룬다.

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.

$$M^2 = M$$

$$H^2 = H$$

(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.

(1) 대칭행렬의 증명

$$\begin{aligned}M^T &= (I - X(X^TX)^{-1}X^T)^T \\ &= I - X(X^TX)^{-T}X^T \\ &= I - X((X^TX)^T)^{-1}X^T \\ &= I - X(X^TX)^{-1}X^T \\ &= M\end{aligned}$$

$$
\begin{aligned}
H^T &= (X(X^TX)^{-1}X^T)^T \\
&= X((X^TX)^T)^{-1}X^T \\
&= X(X^TX)^{-1}X^T \\
&= H
\end{aligned}
$$

(2) 멱등성 증명

$$
\begin{aligned}
M^2 &= (I - X(X^TX)^{-1}X^T)(I - X(X^TX)^{-1}X^T) \\
&= I - 2X(X^TX)^{-T}X^T + X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T \\
&= I - X((X^TX)^T)^{-1}X^T \\
&= M
\end{aligned}
$$

$$
\begin{aligned}
H^2 &= (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) \\
&= X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T \\
&= X(X^TX)^{-1}X^T \\
&= H
\end{aligned}
$$

(3) M과 H의 직교 증명

$$
\begin{aligned}
MH &= (I - X(X^TX)^{-1}X^T)X(X^TX)^{-1}X^T \\
&= X(X^TX)^{-T}X^T - X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T \\
&= X(X^TX)^{-T}X^T - X(X^TX)^{-1}X^T \\
&= 0
\end{aligned}
$$

(4) M과 X의 직교 증명

$$
\begin{aligned}
MX &= (I - X(X^TX)^{-1}X^T)X \\
&= X - X(X^TX)^{-1}X^TX \\
&= X - X \\
&= 0
\end{aligned}
$$

(5) H과 X의 곱에 대한 증명

$$
\begin{aligned}
HX &= (X(X^TX)^{-1}X^T)X \\
&= X(X^TX)^{-1}X^TX \\
&= X
\end{aligned}
$$

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

$$
y = \hat{y} + e = Hy + My = (H + M)y
$$

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

이 관계식은 나중에 분산 분석(ANOVA)에 사용된다.