⟨**권이태, 0823**⟩

행렬계산

Block Matrix

아래의 블럭행렬

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

를 생각하자. 블럭행렬의 행렬식은

$$\det \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \det(A_{11})\det(A_{22} - A_{21}A_{11}^{-1}A_{12})$$

으로 주어진다. 블럭행렬의 역행렬은

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \\ -(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} & (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \end{pmatrix}$$

으로 주어진다.

블럭매트릭스 연산이 자주 사용되는 통계학적 상황은 다변량 정규분포의 조건부 분포를 구할 때이다.

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0_{n_1} \\ 0_{n_2} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

이때 $Y_2|Y_1$ 등의 조건부 분포를 구하고 싶을 수 있다. 그렇다면

$$\begin{aligned}
\mathrm{pdf}_{Y_2|Y_1}(y_2) &= \frac{\mathrm{pdf}_{Y_1,Y_2}(y_1,y_2)}{\mathrm{pdf}_{Y_1}(y_1)} \\
&= \frac{|\Sigma_{11}|}{(2\pi)^{n_2/2}|\Sigma|} \exp\left(-\frac{1}{2}\left(y^T\Sigma^{-1}y - y_1^T\Sigma_{11}^{-1}y_1 \right) \right) \\
&= \frac{|\Sigma_{11}|}{(2\pi)^{n_2/2}|\Sigma_{11}||\Sigma_{22\cdot 1}|} \\
&\times \exp\left(-\frac{1}{2}\left(y_1^T\Sigma_{11}^{-1}y_1 + y_1^T\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22\cdot 1}^{-1}\Sigma_{21}\Sigma_{11}^{-1}y_1 - 2y_2^T\Sigma_{22\cdot 1}^{-1}\Sigma_{21}\Sigma_{11}^{-1}y_1 + y_2^T\Sigma_{22\cdot 1}^{-1}y_2 - y_1^T\Sigma_{11}^{-1}y_1 \right) \right) \\
&= \frac{1}{(2\pi)^{n_2/2}|\Sigma_{22\cdot 1}|} \exp\left(-\frac{1}{2}(y_2^T\Sigma_{22\cdot 1}^{-1}y_2 - 2y_2^T\Sigma_{22\cdot 1}^{-1}\Sigma_{21}\Sigma_{11}^{-1}y_1 + y_1^T\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22\cdot 1}^{-1}\Sigma_{21}\Sigma_{11}^{-1}y_1) \right) \\
&= \frac{1}{(2\pi)^{n_2/2}|\Sigma_{22\cdot 1}|} \exp\left(-\frac{1}{2}(y_2 - \Sigma_{21}\Sigma_{11}^{-1}y_1)^T\Sigma_{22\cdot 1}^{-1}(y_2 - \Sigma_{21}\Sigma_{11}^{-1}y_1) \right)
\end{aligned}$$

이때 $\Sigma_{22\cdot 1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$이다. 따라서 위로부터 우리는

$$Y_2|Y_1 = y_1 \sim N(\Sigma_{21}\Sigma_{11}^{-1}y_1, \Sigma_{22\cdot 1})$$

을 얻는다.

혹은 회귀분석에서도 이를 사용할 수 있다. Frisch-Waugh-Lovell 정리의 증명 과정에서가 대표적이다. 회귀분석에서 $Y = X\beta + \epsilon$ 모형을 생각하자. 변수가 p개고 상수항이 있는 모형을 적합할 때, 설명변수의 행렬 X는 $n \times (p+1)$ 행렬이었다. 만약 우리가 이 설명변수들을 앞의 p_0개와 뒤의 p_1개로 분리하여 생각한다고

해보자. $p_0 + p_1 = p + 1$이다. 그렇다면 $n \times p_0$ 행렬 X_0과 $n \times p_1$ 행렬 X_1에 대하여,

$$X = \begin{pmatrix} X_0 & X_1 \end{pmatrix}$$

처럼 열끼리 붙여 표현할 수 있고, β도 마찬가지로 X_0에 곱해질 $\beta_0, \cdots, \beta_{p_0-1}$을 $\beta_0 \in \mathbb{R}^{p_0}$, X_1에 곱해질 $\beta_{p_0}, \cdots, \beta_p$를 β_1로 표현해

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

처럼 쓰기로 하자. 그렇다면 우리는

$$Y = X\beta + e = \begin{pmatrix} X_0 & X_1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + e = X_0\beta_0 + X_1\beta_1 + e$$

처럼 X_0와 X_1의 선형결합이 $E[Y]$를 결정한다고 결론지을 수 있다. 이제 우리는 아래처럼 hat matrix들을 정의하자.

$$\Pi_0 = X_0(X_0^T X_0)^{-1} X_0^T,\ X_{1|0} = (I - \Pi_0)X_1,\ \Pi_{1|0} = X_{1|0}(X_{1|0}^T X_{1|0})^{-1} X_{1|0}$$

이들의 함의는 사실 선형대수에서의 projection가 같다. Π_0는 어떠한 벡터 앞에 곱해질 경우 해당 벡터의 $col(X_0)$, 즉 X_0의 column space 성분을 보여주는 projection matrix와 같다. 한편 $X_{1|0}$은 직교화된 X_1으로, X_1에서 X_0에 나란한 성분 $\Pi_0 X_1$을 빼 만든 X_0에 수직인 부분이다. 마지막으로 $\Pi_{1|0}$은 X_0에 수직인 성분들이 모여 있는 orthogonal space $(col(X_0))^\perp$의 원소들을 $col(X_{1|0})$으로 사영하는 선형사상이다. 통계적 관점에서, projection은 반응변수 중에서 설명변수의 선형결합이 얼마나 그 변동을 설명할 수 있는지를 의미한다. 다르게 말하면, 우리가 기존의 오차항의 추정값으로 사용한 $\hat{e} = (I - \Pi)Y$는 실제 Y에서, 설명변수 X의 선형결합으로 설명 가능한 부분 ΠY를 뺀, 즉 설명 불가능한 부분이다. 한편 $X_{1|0}$은 X_1 중에서 X_0의 선형결합으로써 표현이 불가능한 부분을 의미한다.

Theorem 1.
$\gamma_0 = \beta_0 + (X_0^T X_0)^{-1} X_0^T X_1 \beta_1$ *일 때,*

$$\begin{aligned} X\hat{\beta}^{LSE} &= X_0\hat{\beta}_0^{LSE} + X_1\hat{\beta}_1^{LSE} = X_0\hat{\gamma}_0^{LSE} + X_{1|0}\hat{\beta}_1^{LSE} \\ X_0\hat{\gamma}_0^{LSE} &= \Pi_0 Y,\ X_{1|0}\hat{\beta}_1^{LSE} = \Pi_{1|0} Y \\ \hat{\beta}_0^{LSE} &= \hat{\gamma}_0^{LSE} - (X_0^T X_0)^{-1} X_0^T X_1 \hat{\beta}_1^{LSE} \end{aligned}$$

이다.

Proof.

$$X\beta = X_0\beta_0 + X_1\beta_1 = X_0(\beta_0 + (X_0^T X_0)^{-1} X_0^T X_1 \beta_1) + (I - \Pi_0)X_1\beta_1 = X_0\gamma_0 + X_{1|0}\beta_1$$

임을 알 수 있다. 한편 이제 우리의 선형모형은

$$Y = X\beta = X_0\gamma_0 + X_{1|0}\beta_1 + e$$

이 되는데, 최소제곱추정량에서 최소화하려는 항이

$$\begin{aligned} ||Y - X\beta||^2 &= ||Y - X_0\gamma_0 - X_{1|0}\beta_1||^2 \\ &= ||(\Pi_0 Y - X_0\gamma_0) + ((I - \Pi_0)Y - X_{1|0}\beta_1)||^2 \\ &= ||\Pi_0 Y - X_0\gamma_0||^2 + ||(I - \Pi_0)Y - X_{1|0}\beta_1||^2 + 2(\Pi_0 Y - X_0\gamma_0)^T((I - \Pi_0)Y - X_{1|0}\beta_1) \end{aligned}$$

과 같다. 그런데

$$\begin{aligned}(\Pi_0 Y - X_0\gamma_0)^T((I-\Pi_0)Y - X_{1|0}\beta_1) &= Y^T\Pi_0^T(I-\Pi_0)Y - \gamma_0 X_0^T(I-\Pi_0)Y \\ &\quad - Y^T\Pi_0^T X_{1|0}\beta_1 + \gamma_0^T X_0^T X_{1|0}\beta_1 \\ &= 0\end{aligned}$$

이 $\Pi_0^T(I-\Pi_0) = X_0^T(I-\Pi_0) = \Pi_0^T X_{1|0} = X_0^T X_{1|0} = 0$이므로 성립한다. 즉, 설명변수의 직교화라 함은 두 설명변수 X_0와 X_1을 직교화하여 X_0과 $X_{1|0}$으로 만듦으로써 위의 식들을 모두 0으로 만들고, 이로써 projection을 쉽게 해주는 방법을 의미한다.

$$\begin{aligned}||(I-\Pi_0)Y - X_{1|0}\beta_1||^2 &= ||(I-\Pi_0-\Pi_{1|0})Y + \Pi_{1|0}Y - X_{1|0}\beta_1||^2 \\ &= ||(I-\Pi_0-\Pi_{1|0})Y||^2 + ||\Pi_{1|0}Y - X_{1|0}\beta_1||^2 \\ &\quad + 2((I-\Pi_0-\Pi_{1|0})Y)^T(\Pi_{1|0}Y - X_{1|0}\beta_1)\end{aligned}$$

에서,

$$((I-\Pi_0-\Pi_{1|0})Y)^T(\Pi_{1|0}Y - X_{1|0}\beta_1) = Y^T(I-\Pi_0-\Pi_{1|0})^T X_{1|0}((X_{1|0}^T X_{1|0})^{-1}X_{1|0}^T Y - \beta_1)$$

인데

$$(I-\Pi_0-\Pi_{1|0})^T X_{1|0} = X_{1|0} - \Pi_0^T X_{1|0} - \Pi_{1|0}^T X_{1|0} = O$$

이므로 이것도 0이다. 따라서 우리는

$$||Y - X\beta||^2 = ||\Pi_0 Y - X_0\gamma_0||^2 + ||\Pi_{1|0}Y - X_{1|0}\beta_1||^2 + ||(I-\Pi_0-\Pi_{1|0})Y||^2$$

임을 안다. 마지막 항은 항상 존재하므로, 앞선 두 항을 없애기 위해 우리는

$$X_0\hat{\gamma}_0^{LSE} = \Pi_0 Y, \quad X_{1|0}\hat{\beta}_1^{LSE} = \Pi_{1|0}Y$$

인 LSE를 선택하게 된다. 다르게 말하면,

$$\hat{\gamma}_0^{LSE} = (X_0^T X_0)^{-1}X_0^T Y, \quad \hat{\beta}_1^{LSE} = (X_{1|0}^T X_{1|0})^{-1}X_{1|0}^T Y$$

를 LSE로 선택할 것이다.

이제 첫째 식을 증명하자. 둘째 식으로부터 첫째 식의 가장 우변이

$$\Pi_0 Y + \Pi_{1|0}Y = (\Pi_0 + \Pi_{1|0})Y$$

임을 알 수 있다. 한편 좌변은 $X = \begin{pmatrix} X_0 & X_1 \end{pmatrix}$에 대하여 ΠY와 같다. 즉 우리는

$$\Pi = \Pi_0 + \Pi_{1|0}$$

임을 보여 주자. $X_1^T(I-\Pi_0)X_1 = X_{1|0}^T X_{1|0}$이므로,

$$\begin{aligned}\Pi &= X(X^T X)^{-1}X \\ &= \begin{pmatrix} X_0 & X_1 \end{pmatrix}\left(\begin{pmatrix} X_0^T \\ X_1^T \end{pmatrix}\begin{pmatrix} X_0 & X_1 \end{pmatrix}\right)^{-1}\begin{pmatrix} X_0^T \\ X_1^T \end{pmatrix} \\ &= \begin{pmatrix} X_0 & X_1 \end{pmatrix}\left(\begin{pmatrix} X_0^T X_0 & X_0^T X_1 \\ X_1^T X_0 & X_1^T X_1 \end{pmatrix}\right)^{-1}\begin{pmatrix} X_0^T \\ X_1^T \end{pmatrix}\end{aligned}$$

$$
\begin{aligned}
&= \begin{pmatrix} X_0 & X_1 \end{pmatrix} \begin{pmatrix} (X_0^TX_0)^{-1}X_0^T(I + X_1(X_{1|0}^TX_{1|0})^{-1}X_1^T)X_0(X_0^TX_0)^{-1} & -(X_0^TX_0)^{-1}X_0^TX_1(X_{1|0}^TX_{1|0})^{-1} \\ (X_{1|0}^TX_{1|0})^{-1}X_1^TX_0(X_0^TX_0)^{-1} & (X_{1|0}^TX_{1|0})^{-1} \end{pmatrix} \begin{pmatrix} X_0^T \\ X_1^T \end{pmatrix} \\
&= \begin{pmatrix} X_0 & X_1 \end{pmatrix} \begin{pmatrix} (X_0^TX_0)^{-1}X_0^T(I + X_1(X_{1|0}^TX_{1|0})^{-1}X_1^T)\Pi_0 - (X_0^TX_0)^{-1}X_0^TX_1(X_{1|0}^TX_{1|0})^{-1}X_1^T \\ (X_{1|0}^TX_{1|0})^{-1}X_1^T\Pi_0 + (X_{1|0}^TX_{1|0})^{-1}X_1^T \end{pmatrix} \\
&= \Pi_0 + \Pi_0X_1(X_{1|0}^TX_{1|0})^{-1}X_1^T\Pi_0^T - \Pi_0X_1(X_{1|0}^TX_{1|0})^{-1}X_1^T + X_1(X_{1|0}^TX_{1|0})^{-1}X_1^T\Pi_0 + X_1(X_{1|0}^TX_{1|0})^{-1}X_1^T \\
&= \Pi_0 + (I - \Pi_0)X_1(X_{1|0}^TX_{1|0})^{-1}X_1^T(I - \Pi_0)^T \\
&= \Pi_0 + X_{1|0}(X_{1|0}^TX_{1|0})^{-1}X_{1|0}^{-1} = \Pi_0 + \Pi_{1|0}
\end{aligned}
$$

이다. 따라서 첫째 식도 쉽게 증명할 수 있다. 이제 마지막 식을 보이자.

$$\hat{\beta}_0^{LSE} = \begin{pmatrix} I_{p_0} & O_{p_0 \times p_1} \end{pmatrix} \hat{\beta}^{LSE} = T\hat{\beta}^{LSE}$$

라고 하자. 그러면

$$
\begin{aligned}
\hat{\beta}_0^{LSE} &= T(X^TX)^{-1}X^TY \\
&= T \begin{pmatrix} (X_0^TX_0)^{-1}X_0^T(I + X_1(X_{1|0}^TX_{1|0})^{-1}X_1^T)\Pi_0 - (X_0^TX_0)^{-1}X_0^TX_1(X_{1|0}^TX_{1|0})^{-1}X_1^T \\ (X_{1|0}^TX_{1|0})^{-1}X_1^T\Pi_0 + (X_{1|0}^TX_{1|0})^{-1}X_1^T \end{pmatrix} Y \\
&= ((X_0^TX_0)^{-1}X_0^T(I + X_1(X_{1|0}^TX_{1|0})^{-1}X_1^T)\Pi_0 - (X_0^TX_0)^{-1}X_0^TX_1(X_{1|0}^TX_{1|0})^{-1}X_1^T)Y \\
&= (X_0^TX_0)^{-1}X_0^T(\Pi_0 - X_1(X_{1|0}^TX_{1|0})^{-1}X_1^T(I - \Pi_0)^T)Y \\
&= (X_0^TX_0)^{-1}X_0^TY - (X_0^TX_0)^{-1}X_0^TX_1(X_{1|0}^TX_{1|0})^{-1}X_{1|0}Y \\
&= \hat{\gamma}_0^{LSE} - (X_0^TX_0)^{-1}X_0^TX_1\hat{\beta}_1^{LSE}
\end{aligned}
$$

이다. □

이걸 다르게 생각해 보자. 우리는 $E[Y]$를 $X\hat{\beta}^{LSE}$로 추정하게 된다. 따라서

$$\hat{E}[Y] = X\hat{\beta}^{LSE} = X_0\hat{\gamma}_0^{LSE} + X_{1|0}\hat{\beta}_1^{LSE}$$

이다. 이제 $X_0\hat{\gamma}_0^{LSE}$을 우변으로 넘기면,

$$\hat{E}[Y] - X_0\hat{\gamma}_0^{LSE} = X_{1|0}\hat{\beta}_1^{LSE}$$

을 얻는다. 이제 우리는 β_1의 의미를 정확히 알 수 있다. 먼저 좌변은 Y의 평균반응 중 X_0으로써 설명되는 부분을 뺀 항이다. 즉 Y의 평균반응 중 X_0으로 설명되지 않는 변동이다. 한편 우변의 $X_{1|0}$은 X_1 중 X_0으로 설명되지 않는 변동이다. 둘 사이의 회귀계수가 β_1으로 주어지는 것이다. 따라서 다중회귀분석에서 X_1의 회귀계수 β_1은 곧, '반응변수에서 X_0으로 설명되지 않는 변동을, X_1에서 X_0으로 설명되지 않는 부분으로 설명할 때의 회귀계수'와 같다.

Theorem 2.

$$Cov(\hat{\gamma}_0^{LSE}, \hat{\beta}_1^{LSE}) = O$$

$$E[\hat{\gamma}_0^{LSE}] = \gamma_0, \quad Var(\hat{\gamma}_0^{LSE}) = \sigma^2(X_0^TX_0)^{-1}$$

$$E[\hat{\beta}_1^{LSE}] = \beta_1, \quad Var(\hat{\beta}_1^{LSE}) = \sigma^2(X_{1|0}^TX_{1|0})^{-1}$$

Proof.

$$
\begin{aligned}
Cov(\hat{\gamma}_0^{LSE}, \hat{\beta}_1^{LSE}) &= Cov((X_0^TX_0)^{-1}X_0^TY, (X_{1|0}^TX_{1|0})^{-1}X_{1|0}^TY) \\
&= \sigma^2(X_0^TX_0)^{-1}X_0^TX_{1|0}(X_{1|0}^TX_{1|0})^{-1} = O
\end{aligned}
$$

$$E[\hat{\gamma}_0^{LSE}] = E[(X_0^T X_0)^{-1} X_0^T Y] = E[(X_0^T X_0)^{-1} X_0^T (X_0\beta_0 + X_1\beta_1 + e)] = \beta_0 + (X_0^T X_0)^{-1} X_0^T X_1 \beta_1 = \gamma_0$$

$$Var(\hat{\gamma}_0^{LSE}) = \sigma^2 (X_0^T X_0)^{-1} X_0^T X_0 (X_0^T X_0)^{-1} = \sigma^2 (X_0^T X_0)^{-1}$$

$$E[\hat{\beta}_1^{LSE}] = E[(X_{1|0}^T X_{1|0})^{-1} X_{1|0}^T (X_0\gamma_0 + X_{1|0}\beta_1 + e)] = \beta_1$$

$$Var(\hat{\beta}_1^{LSE}) = \sigma^2 (X_{1|0}^T X_{1|0})^{-1} X_{1|0}^T X_{1|0} (X_{1|0}^T X_{1|0})^{-1} = \sigma^2 (X_{1|0}^T X_{1|0})^{-1}$$

으로 간단하게 증명된다. □

위처럼 설명변수 X를 직교화하거나 분리하여 omitted variable 관련한 문제를 풀 때, 블럭행렬의 연산이 필요한 경우가 있다.

Sherman-Morrison-Woodbury Formula

한편 블럭행렬 이외에도 특정한 형태의 행렬에 대해 역행렬을 구하고 싶은 경우가 있다. 이때 사용할 수 있는 것이 셔먼-모리슨 공식이다.

Theorem 3. *가역행렬 A에 대해 $A + uv^T$ 형태의 행렬이 가역일 필요충분조건은 $1 + v^T A u \neq 0$인 것이며, 이때*

$$(A + uv^T)^{-1} = A^{-1} - A^{-1} u (1 + v^T A^{-1} u)^{-1} v^T A^{-1}$$

이다.

이를 확장하면 일반적인 $n \times m$ 행렬 U, V에 대하여 셔먼-모리슨-우드버리 공식

Theorem 4. *가역행렬 A에 대해 $A + UV^T$ 형태의 행렬이 가역일 필요충분조건은 $I + V^T A U$가 가역인 것이며, 이때*

$$(A + UV^T)^{-1} = A^{-1} - A^{-1} U (I + V^T A^{-1} U)^{-1} V^T A^{-1}$$

이다.

상황에 따라서는 행렬식을 구하는 데 사용하기도 한다. 특히 $\det(I + uv^T) = 1 + v^T u$임이 잘 알려져 있으므로,

$$\begin{aligned} \det(A + uv^T) &= \det(A)\det(I + A^{-1} uv^T) \\ &= \det(A)(1 + v^T A^{-1} u) \end{aligned}$$

이다. 더욱 확장하면

$$\det(A + UV^T) = \det(A)\det(I_m + V^T A^{-1} U)$$

라는 matrix determinant lemma를 얻을 수도 있다.

셔먼-모리슨-우드버리 공식을 쓸 수 있는 상황에는 cross-validation을 수행할 때이다. 대표적으로 LOOCV를 고려해 보자. X_{-i}와 Y_{-i}를 각각 i번째 observation을 제거하였을 때의 model matrix와 response라고 하자. 그렇다면 추정량은

$$\hat{\beta}_{-i} = (X_{-i}^T X_{-i})^{-1} X_{-i}^T Y_{-i}$$

로 나타날 것이다. 이때

$$X = \begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{pmatrix}$$

으로 쓰면

$$X^TX = \begin{pmatrix} X_1 & X_2 & \cdots & X_n \end{pmatrix} \begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{pmatrix} = X_1X_1^T + \cdots + X_nX_n^T$$

이고

$$X_{-i}^TX_{-i} = X^TX - X_iX_i^T$$

처럼 쓸 수 있음이 online regression을 비롯한 다양한 분야에서 잘 알려져 있다. 그렇다면 우리는

$$\begin{aligned}(X_{-i}^TX_{-i})^{-1} &= (X^TX - X_iX_i^T)^{-1} \\ &= (X^TX)^{-1} + (X^TX)^{-1}X_i(1 - X_i^T(X^TX)^{-1}X_i)^{-1}X_i^T(X^TX)^{-1} \\ &= (X^TX)^{-1} + \frac{(X^TX)^{-1}X_iX_i^T(X^TX)^{-1}}{1-h_{ii}}\end{aligned}$$

임을 알 수 있다. 이때 $h_{ii} = X_i^T(X^TX)^{-1}X_i$는 i번째 점의 leverage이다. 자연스럽게 여기에서는 X_{-i}를 기반으로 한 추론이 X를 기반으로 한 추론에 비해 벗어나는 정도가 $(1-h_{ii})^{-1}$에 비례함을 아는데, 이는 high leverage point가 회귀계수의 추정에 큰 영향을 미칠 수 있음을 암시한다.

한편 동일한 과정으로 $X_{-i}^TY_{-i} = X^TY - X_iy_i$이므로,

$$\begin{aligned}\hat{\beta}_{-i} &= \left((X^TX)^{-1} + \frac{(X^TX)^{-1}X_iX_i^T(X^TX)^{-1}}{1-h_{ii}}\right)(X^TY - X_iy_i) \\ &= \hat{\beta} - \left(\frac{(X^TX)^{-1}X_i}{1-h_{ii}}\right)(y_i(1-h_{ii}) - X_i^T\hat{\beta} + h_{ii}y_i) \\ &= \hat{\beta} - (X^TX)^{-1}X_ie_i/(1-h_{ii})\end{aligned}$$

를 얻는다. 따라서

$$\begin{aligned}e_{i,-i} &= y_i - X_i^T\hat{\beta}_{-i} \\ &= y_i - X_i^T\hat{\beta} + X_i^T(X^TX)^{-1}X_ie_i/(1-h_{ii}) \\ &= e_i + \frac{h_{ii}}{1-h_{ii}}e_i \\ &= \frac{e_i}{1-h_{ii}}\end{aligned}$$

를 얻는다. 따라서

$$\begin{aligned}\widehat{\mathrm{MSE}}_{\mathrm{LOOCV}} &= \frac{1}{n}\sum_{i=1}^n e_{i,-i}^2 \\ &= \frac{1}{n}\sum_{i=1}^n \left(\frac{e_i}{1-h_{ii}}\right)^2\end{aligned}$$

으로 n번의 회귀추정을 할 필요 없이 간단하게 구할 수 있다.

한편 $\hat{\beta}_{-i}$를 저렇게 구하면 영향점 탐지를 하는 데에도 유용하게 사용할 수 있다. 먼저 Cook's distance

에서는

$$\begin{aligned}
C_i &= \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j,-i})^2}{(p+1)s^2} \\
&= \frac{(\hat{\beta} - \hat{\beta}_{-i})^T X^T X(\hat{\beta} - \hat{\beta}_{-i})}{(p+1)s^2} \\
&= \frac{((X^TX)^{-1}X_i e_i/(1-h_{ii}))^T X^T X((X^TX)^{-1}X_i e_i/(1-h_{ii}))}{(p+1)s^2} \\
&= \frac{e_i^T X_i^T (X^TX)^{-1} X_i e_i}{(1-h_{ii})^2(p+1)s^2} \\
&= \frac{h_{ii}}{1-h_{ii}} \times \frac{e_{si}^2}{p+1}
\end{aligned}$$

으로 계산을 편히 할 수 있다. 이때

$$e_{si} = \frac{e_i}{s\sqrt{1-h_{ii}}}$$

로 internally studentized residual이 된다.

DFFITS에서는

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i}\sqrt{h_{ii}}}$$

인데, 이를 위해서는 externally studentized residual s_{-i}를 알아야 한다는 번거로움이 있다. 이때

$$\begin{aligned}
s_{-i}^2 &= \frac{\text{SSE}_{-i}}{n-p-2} \\
&= \frac{(Y_{-i} - X_{-i}\hat{\beta}_{-i})^T(Y_{-i} - X_{-i}\hat{\beta}_{-i})}{n-p-2} \\
&= \frac{Y^TY - Y_i^2 - 2\hat{\beta}_{-i}^T X_{-i}^T Y_{-i} + \hat{\beta}_{-i}^T X_{-i}^T X_{-i}\hat{\beta}_{-i}}{n-p-2} \\
&= \frac{Y^TY - Y_i^2 - 2(\hat{\beta} - (X^TX)^{-1}X_i e_i/(1-h_{ii}))^T(X^TY - X_iY_i)}{n-p-2} \\
&+ \frac{(\hat{\beta} - (X^TX)^{-1}X_i e_i/(1-h_{ii}))^T(X^TX - X_iX_i^T)(\hat{\beta} - (X^TX)^{-1}X_i e_i/(1-h_{ii}))}{n-p-2} \\
&= \frac{Y^TY - 2\hat{\beta}^T X^TY + \hat{\beta}^T X^TX\hat{\beta}}{n-p-2} \\
&+ \frac{-Y_i^2 + 2e_iX_i^T(X^TX)^{-1}X^TY/(1-h_{ii}) + 2\hat{\beta}^T X_iY_i - 2e_iX_i^T(X^TX)^{-1}X_iY_i/(1-h_{ii})}{n-p-2} \\
&+ \frac{-\hat{\beta}^T X_iX_i^T\hat{\beta} - 2e_iX_i^T(X^TX)^{-1}(X^TX - X_iX_i^T)\hat{\beta}/(1-h_{ii})}{n-p-2} \\
&+ \frac{2e_iX_i^T(X^TX)^{-1}(X^TX - X_iX_i^T)(X^TX)^{-1}X_ie_i/(1-h_{ii})^2}{n-p-2} \\
&= \frac{\text{SSE}}{n-p-2} \\
&+ \frac{-Y_i^2 + 2e_iX_i^T\hat{\beta}/(1-h_{ii}) + 2\hat{\beta}^T X_iY_i - 2e_ih_{ii}Y_i/(1-h_{ii})}{n-p-2} \\
&+ \frac{-\hat{\beta}^T X_iX_i^T\hat{\beta} - 2e_iX_i^T\hat{\beta}/(1-h_{ii}) + 2e_ih_{ii}X_i^T\hat{\beta}/(1-h_{ii})}{n-p-2} \\
&+ \frac{e_ih_{ii}e_i/(1-h_{ii})^2 - e_ih_{ii}^2e_i/(1-h_{ii})^2}{n-p-2}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\text{SSE}}{n-p-2} + \frac{-Y_i^2 + 2e_i\hat{Y}_i/(1-h_{ii}) + 2Y_i^T\hat{Y}_i - 2e_ih_{ii}Y_i/(1-h_{ii})}{n-p-2} \\
&+ \frac{-\hat{Y}_i^2 - 2e_i\hat{Y}_i/(1-h_{ii}) + 2e_ih_{ii}\hat{Y}_i/(1-h_{ii}) + e_i^2h_{ii}/(1-h_{ii})^2 - e_i^2h_{ii}^2/(1-h_{ii})^2}{n-p-2} \\
&= \frac{SSE - e_i^2/(1-h_{ii})}{n-p-2} \\
&= \frac{(n-p-1)s^2 - s^2e_{si}^2}{n-p-2} \\
&= s^2 \times \frac{n-p-1-e_{si}^2}{n-p-2}
\end{aligned}
$$

임을 계산을 통해 알 수 있기 때문에,

$$
\begin{aligned}
\text{DFFITS}_i &= \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i}\sqrt{h_{ii}}} \\
&= \frac{X_i^T(\hat{\beta} - \hat{\beta}_{-i})}{s\sqrt{h_{ii}}} \times \sqrt{\frac{n-p-2}{n-p-1-e_{si}^2}} \\
&= \frac{e_{i,-i} - e_i}{s\sqrt{h_{ii}}} \times \sqrt{\frac{n-p-2}{n-p-1-e_{si}^2}} \\
&= \frac{e_i/(1-h_{ii}) - e_i}{s\sqrt{h_{ii}}} \times \sqrt{\frac{n-p-2}{n-p-1-e_{si}^2}} \\
&= \sqrt{\frac{h_{ii}}{1-h_{ii}}}e_{si} \times \sqrt{\frac{n-p-2}{n-p-1-e_{si}^2}} \\
&= \sqrt{\frac{h_{ii}}{1-h_{ii}}}e_{si,-i}
\end{aligned}
$$

로 LOO 없이 DFFITS를 쉽게 구해줄 수 있다. 이외에도 CV에 기반한 추정량들은 셔먼-모리슨-우드버리 공식을 통하여 간단하게 표현할 수 있다.

행렬미분

다변량분포 하에서 MLE 등을 구할 때 행렬미분을 자주 수행한다. 알아두면 좋을 공식은 아래와 같다.

행렬 혹은 스칼라를 스칼라로 미분할 때

$$
\begin{aligned}
\frac{\partial}{\partial x}A^{-1}(x) &= -A^{-1}(x)\frac{\partial A(x)}{\partial x}A^{-1}(x) \\
\frac{\partial}{\partial x}|A(x)| &= |A(x)|\text{tr}\left(A^{-1}(x)\frac{\partial A(x)}{\partial x}\right) \\
\frac{\partial}{\partial x}\ln|A(x)| &= \text{tr}\left(A^{-1}(x)\frac{\partial A(x)}{\partial x}\right) \\
\frac{\partial}{\partial x}\text{tr}(A(x)) &= \text{tr}\left(\frac{\partial A(x)}{\partial x}\right)
\end{aligned}
$$

스칼라를 행렬로 미분할 때

$$
\begin{aligned}
\frac{\partial u^TAv}{\partial A} &= vu^T \\
\frac{\partial \text{tr}(A)}{\partial A} &= I
\end{aligned}
$$

$$\frac{\partial \mathrm{tr}(g(A))}{\partial A} = g'(A)$$
$$\frac{\partial \mathrm{tr}(UA)}{\partial A} = U$$
$$\frac{\partial \mathrm{tr}(A^T UA)}{\partial A} = A^T(U + U^T)$$
$$\frac{\partial \mathrm{tr}(A^{-1}U)}{\partial A} = -A^{-1}UA^{-1}$$
$$\frac{\partial \mathrm{tr}(UAV)}{\partial A} = VU$$
$$\frac{\partial |A|}{\partial A} = |A|A^{-1}$$
$$\frac{\partial \ln |A|}{\partial A} = A^{-1}$$

벡터 혹은 스칼라를 벡터로 미분할 때

$$\frac{\partial U\mathbf{x}}{\partial \mathbf{x}} = U$$
$$\frac{\partial g(u(\mathbf{x}))}{\partial \mathbf{x}} = \frac{\partial g(u(\mathbf{x}))}{\partial u}\frac{\partial u}{\partial \mathbf{x}}$$
$$\frac{\partial \mathbf{x}^T S\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T(S + S^T)$$
$$\frac{\partial u^T \mathbf{x}\mathbf{x}^T v}{\partial \mathbf{x}} = \mathbf{x}^T(uv^T + vu^T)$$
$$\frac{\partial \|\mathbf{x}\|}{\partial \mathbf{x}} = \frac{\mathbf{x}^T}{\|\mathbf{x}\|}$$

대표적으로 다변량정규분포에서 얻은 표본으로부터 Σ에 대한 추론을 수행할 때를 고려하자. $X_1, \cdots, X_n$이 $N_p(\mu, \Sigma)$로부터의 랜덤표본이라고 할 때, 가능도함수는

$$L(\Sigma; x) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu)\right)$$

으로 주어지고 로그가능도함수는

$$l(\Sigma; x) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu)$$

를 얻는다. 한편 Σ가 존재하는 모수공간은 $p \times p$ symmetric, positive definite matrix의 공간이다. 동시에 Σ^{-1} 역시도 Σ와 일대일대응되며 같은 모수공간에 존재하므로, Σ 대신 Σ^{-1}을 찾아도 된다.

로그가능도함수를 μ로 미분하면

$$\begin{aligned}\frac{\partial l}{\partial \mu} &= -\frac{n}{2}\mu^T(\Sigma^{-1} + \Sigma^{-1}) + \sum_{i=1}^{n} x_i^T\Sigma^{-1} \\ &= -n\mu^T\Sigma^{-1} + n\bar{x}\Sigma^{-1}\end{aligned}$$

이므로 일계조건에서 $\hat{\mu}^{\mathrm{mle}} = \bar{x}$이다. ($\Sigma^{-1}$은 가역)

로그가능도함수를 Σ^{-1}로 미분하면

$$\frac{\partial l}{\partial \Sigma^{-1}} = \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$$

이며, 일계조건으로부터

$$\hat{\Sigma}^{\text{mle}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

를 얻는다.

둘째 응용은 직교인자모형의 적합 과정이다. 직교인자모형에서 CMLE를 통해 추정하는 상황을 고려하여 보자. 참모형이

$$X = \mu + Lf + \epsilon$$

이라 하고 $f \sim N(0_m, I_m)$, $\epsilon \sim N(0, \Phi)$이라 하자. f와 ϵ은 독립이다. 그렇다면

$$L(\mu, \Sigma) = (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^T\Sigma^{-1}(X_i - \mu)\right)$$

인데, $\Sigma = LL^T + \Phi$가 될 것이다. 앞서 Σ^{-1}는 셔먼-모리슨-우드버리 공식에 의하여

$$\Sigma^{-1} = \Phi^{-1} - \Phi^{-1}L(I - L^T\Phi^{-1}L)^{-1}L^T\Phi^{-1}$$

임을 안다. 계산적으로 이를 구하기 쉬우면서 유일하게 해를 찾을 수 있으려면, $L^T\Phi^{-1}L$이 대각행렬이면 된다. 따라서 일반적으로 CMLE에서는 대각행렬 Δ에 대해 $L^T\Phi^{-1}L = \Delta$와 같은 제약을 건다. 한편 μ의 최소제곱추정량은 $\bar{X}$로 쉽게 구해지니 생략하고 해당 조건 하에서

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$$

로 정의하면, 로그가능도함수에서 상수항을 제거한 최적화하고자 하는 항은

$$-\frac{n}{2}\log(\|LL^T + \Phi|) - \frac{1}{2}\text{tr}((LL^T + \Phi)^{-1}S_n)$$

처럼 쓸 수 있고, 이로부터 L, Φ에 대해 풀려는 문제는

$$\begin{cases}\text{maximize} & \log|(LL^T + \Phi)^{-1}| - \text{tr}((LL^T + \Phi)^{-1}S_n) \\ \text{subject to} & L^T\Phi^{-1}L = \Delta\end{cases}$$

으로 주어진다.

목적함수를 Φ로 미분하여 보자. 그러면

$$\begin{aligned}&\frac{\partial}{\partial\Phi}(\log|(LL^T + \Phi)^{-1}| - \text{tr}((LL^T + \Phi)^{-1}S_n)) \\ &= (LL^T + \Phi)^{-1} - (LL^T + \Phi)^{-1}S_n(LL^T + \Phi)^{-1}\end{aligned}$$

으로 주어진다. 따라서 일계조건을 $\text{diag}(\Sigma^{-1}(\Sigma - S_n)\Sigma^{-1}) = O$처럼 적을 수도 있다. 이때 Φ에는 p개의 값만 있고 off-diagonal들은 0으로 제약이 걸어져 있기 때문에 저 값이 정확히 O가 아니라 그 대각원소들이 0이라는 조건만 얻는다. 적절한 행렬 연산을 통하여 $\Sigma^{-1}(\Sigma - S_n)\Sigma^{-1} = \Phi^{-1}(\Sigma - S_n)\Phi^{-1}$를 얻고, Φ는 대각행렬이므로 결국

$$\text{diag}(LL^T + \Phi) = \text{diag}(S_n)$$

을 조건으로 얻는다.

목적함수를 L로 미분하여 보자. 그러면

$$\frac{\partial}{\partial L}(\log|(LL^T + \Phi)^{-1}| - \text{tr}((LL^T + \Phi)^{-1}S_n))$$

$$= L^T(LL^T+\Phi)^{-1} - L^T(LL^T+\Phi)^{-1}S_n(LL^T+\Phi)^{-1}$$

이며, 우리는 일계조건을 $\Sigma^{-1}L = \Sigma^{-1}S_n\Sigma^{-1}L$처럼 쓸 수 있다. 이 식을 조금 더 정리하여 보자.

$$\Sigma\Phi^{-1}L = (LL^T+\Phi)\Phi^{-1}L = LL^T\Phi^{-1}L + L = L(\Delta+I)$$

에서 양변을 정리하면

$$\Sigma^{-1}L = \Phi^{-1}L(\Delta+I)^{-1}$$

을 얻는다. 따라서 위 식은

$$L = S_n\Sigma^{-1}L = S_n\Phi^{-1}L(\Delta+I)^{-1}$$

, 더 나아가

$$L(\Delta+I) = S_n\Phi^{-1}L$$

처럼 쓸 수 있다. 한편 이를 푸는 것은

$$\underbrace{\Phi^{-1/2}S_n\Phi^{-1/2}}_{p\times p}\underbrace{\Phi^{-1/2}L}_{p\times m} = \underbrace{\Phi^{-1/2}L}_{p\times m}\underbrace{(\Delta+I)}_{m\times m}$$

을 푸는 것과 동치인데, $\Phi^{-1/2}S_n\Phi^{-1/2} = PDP^T$로 직교대각화한 다음 $y = P^T\Phi^{-1/2}L$으로 정의하면 $y(\Delta+I) = Dy$의 형태가 된다. $(\Delta+I)$와 D는 모두 대각행렬이므로, 이를 이용하여 쉽게 해를 구할 수 있다. 따라서 CMLE의 적합은 수치적으로

$$\begin{aligned}\mathrm{diag}(LL^T+\Phi) &= \mathrm{diag}(S_n)\\ L(\Delta+I) &= S_n\Phi^{-1}L\end{aligned}$$

을 번갈아 풀어나가며 L, Φ를 찾아나가는 방식으로 이루어진다.

Matrix Norm

행렬에서 노름은 submultiplicativity나 consistency 등의 좋은 성질들을 만족하도록 정의된다. 대표적인 행렬노름은 아래가 있다. $A \in \mathbb{R}^{m\times n}$이라고 하자.

1. **Frobenius norm**
$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\right)^{1/2}$$
2. **p-norm**
$$\|A\|_p = \sup_{x\neq 0}\frac{\|Ax\|_p}{\|x\|_p}$$

Theorem 5. *$\|\cdot\|_2$ 행렬노름에 대하여, $\|A\|_2 = \sigma_1$이다. 이때 σ_1은 A의 가장 큰 singular value이다. 또한 A가 정방행렬이라면 $\|Ax\|_2/\|x\|_2$가 최대화가 되는 x는 고유값 σ_1에 상응하는 고유벡터이다.*

Theorem 6.

$$\|A\|_1 = \max_{j=1,2,\cdots,n}\sum_{i=1}^{m}|a_{ij}|$$
$$\|A\|_\infty = \max_{i=1,2,\cdots,m}\sum_{j=1}^{n}|a_{ij}|$$

이다. 즉 행렬의 1-norm은 모든 열의 1-norm 중 최대값이고, ∞-norm은 모든 행의 1-norm 중 최대값이다.

Theorem 7. *$m \times m$ 유니타리 행렬 Q와 $n \times n$ 유니타리 행렬 Z에 대하여,*

$$\|QAZ\|_2 = \|A\|_2$$
$$\|QAZ\|_F = \|A\|_F$$

Theorem 8.

$$\|A\|_2 \leq \|A\|_F = \sqrt{\mathrm{tr}(AA^T)}$$

Theorem 9. *2-norm에 대하여,*

$$\|Ax\|_2 \leq \|A\|_2\|x\|_2$$

가 성립한다.

대표적으로 행렬노름을 알면 간단한 부분은 classification에서의 maximally separating hyperplane을 계산할 때이다. 이때 우리는

$$\mathrm{argmax}_a \frac{(a^T\mu_2 - a^T\mu_1)^2}{a^T\Sigma a}$$

인 a를 찾고자 한다. $b = \Sigma^{1/2}a$를 정의하면, 이 문제는

$$\mathrm{argmax}_b \frac{\|(\mu_2 - \mu_1)^T\Sigma^{-1/2}b\|_2}{\|b\|_2}$$

처럼 쓸 수 있기에 최대값은 $\|(\mu_2 - \mu_1)^T\Sigma^{-1/2}\|_2$이며 등호조건은 b가 $\Sigma^{-1/2}(\mu_2 - \mu_1)$와 나란한 것이다. 따라서 a는

$$\Sigma^{-1}(\mu_2 - \mu_1)$$

에 나란해야 한다.

한편 PCA에서도 이 개념을 이용할 수 있다. PCA에서 우리는 첫째 PC를

$$a_1 = \mathrm{argmax}_{\|b\|=1}\mathrm{Var}(b^TX)$$

으로 찾는데, 이는 사실

$$a_1 = \mathrm{argmax}_{\|b\|=1}\frac{\|\Sigma^{1/2}b\|_2}{\|b\|_2}$$

으로 쓸 수 있다. 그런데 $\Sigma^{1/2}$의 가장 큰 eigenvalue는 Σ의 가장 큰 eigenvalue에 루트를 씌운 값과 같고, 상응하는 eignvector는 동일하다. 따라서 a_1은 Σ에 직교대각화를 수행한 뒤 가장 큰 고유값에 대응되는 고유벡터 e_1과 동일하게 된다. 단, 이 방법은 a_2부터 구하기가 살짝 어려워진다는 단점이 있기는 하다.

한편 우리가 다중공선성을 파악할 때 사용하는 condition number를 이해하기 위해서도 matrix norm을 아는 것이 좋다. 원래 정방행렬 A에 대한 condition number의 정의는

$$\kappa(A) = \|A\|\|A^{-1}\|$$

으로, 오차가 A를 곱함에 따라 얼마나 증폭되는지를 계측한다. 이는 2-norm 만이 아니라 일반적인 matrix norm에 대해 적용될 수 있다. 한편 condition number가 클수록 수치적으로 불안정해짐을 의미하게 되고, 그런 행렬을 ill-condition을 가지는 행렬이라고 부른다. 이러한 현상은 A가 invertible하지 않은 편에 가까울수록 나타난다. 따라서 회귀분석에서 X^TX의 다중공선성을 파악하기 위하여 그 condition number를 하나의 바로미터로써 사용하는 것이다. 2-norm을 사용하는 경우,

$$\|X^TX\|_2 = \lambda_{\max}$$

이며, $X^TX = PDP^T$로 하면 $(X^TX)^{-1} = PD^{-1}P^T$으로 eigenvalue가 $1/\lambda_1, \cdots, 1/\lambda_n$이 될 것이기에

순서가 뒤집혀

$$\|(X^TX)^{-1}\|_2 = \frac{1}{\lambda_{\min}}$$

를 얻는다. 따라서 우리는

$$\kappa(X^TX) = \frac{\lambda_{\max}}{\lambda_{\min}}$$

을 얻게 되는 것이다.

이외에도 전산통계 및 통계계산 분야에서 다양한 행렬계산 지식들이 사용될 수 있다.