SF 2930 Regression Analysis
Lecture 13.1
Generalized Inverses & Multiple Regression

Timo Koski

KTH Royal Institute of Technology

2023

Learning outcomes

- Generalized Inverses
- Interpolation Limit

Multiple Regression $k >> n$

Let us center the model. Then X is an $n \times k$ matrix. The normal equations are

$$X^T X \boldsymbol{\beta} = X^T \mathbf{y}$$

When $k > n$, the column rank of X cannot be full, since the row rank equals the column rank. (The rank of an $n \times (k+1)$ matrix is the size of the largest invertible square matrix that can be found inside X.) Hence the rank of the $k \times k$ matrix $X^T X$ is smaller than k and the matrix $X^T X$ is not invertible.

Consider the general equation with an $n \times k$ matrix A

$$A\mathbf{x} = \mathbf{z}$$

There are three possibilities:

- There is no solution $\Leftrightarrow \mathbf{z} \notin \mathrm{sp}(A)$.
- There is one solution $\Leftrightarrow A$ is invertible.
- There are many solutions.

We try to find an $k \times n$ matrix G, which would behave as much like A^{-1} is such that if there are many solutions, then $G\mathbf{z}$ is one of them, i.e.,

$$AG\mathbf{z} = \mathbf{z}$$

Arne Bjerhammar: "Application of calculus of matrices to method of least squares; with special references to geodetic calculations". *Transactions of the Royal Institute of Technology* Stockholm. 49, 1951

Generalized Inverse

An $k \times n$ matrix G is called a generalized inverse of an $n \times k$ matrix A if any of the following equivalent conditions hold [1]:

1. $G\mathbf{z}$ is a solution to $A\mathbf{x} = \mathbf{z}$ if solutions exist.
2. GA is idempotent and $\operatorname{rank} GA = \operatorname{rank} A \Leftrightarrow AG$ is idempotent and $\operatorname{rank} AG = \operatorname{rank} A$
3. $AGA = A$

If $n = k$ and the inverse A^{-1} exists, then $G = A^{-1}$, since if we leftmultiply in $AGA = A$

$$A^{-1}AGA = A^{-1}A \Leftrightarrow GA = \mathbb{I}$$

For any given $n \times k$ matrix A there are many generalized inverses.

[1] Proof on p. 106 in S. Puntanen, G.P.H. Styan, J. Isotalo: *Matrix Tricks for Linear Statistical Models. Our Personal Top Twenty* Springer 2011.

Moore-Penrose Generalized Inverse

If a generalized inverse G of A satisfies the four conditions below, then G is called the Bjerhammar -Moore-Penrose (BMP) inverse.

MP1 $AGA = A$

MP2 $GAG = G$

MP3 $(AG)^T = AG$

MP4 $(GA)^T = GA$

Moore (1935), Penrose (1955) showed that for a given A there is only one matrix that satisfies MP1-MP4. We set $A^+ := G$ to denote the Moore-Penrose inverse of A. Arne Bjerhammar found A^+ independently of Moore and Penrose. We shall talk about the Bjerhammar-Moore-Penrose (BMP) inverse.

It can be shown that a generalized inverse always exists, but it is unique if and only if A^{-1} exists. By MP! we get that $G = A^{-1}$.

$$A^{-} = \{G | AGA = A\} \tag{1}$$

is the set of all generalized inverses of A.

Assume that A has full column rank, $\operatorname{rank} A = k$. Then the BMP inverse of A is

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

Proof: Check that MP1-MP4 hold. $\square$

EXAMPLE

Hence, if X has full column rank, then

$$X^+ = (X^T X)^{-1} X^T$$

is the BMP inverse of X.

A^+ is the left inverse of A, since

$$A^+ A = (A^T A)^{-1} A^T A = \mathbb{I}_k.$$

If A has full row rank, $\mathrm{rank}\, A = n$, then

$$A^{+} = A^{T}(AA^{T})^{-1}$$

is the BMP inverse of A. A^{+} is a right inverse of A, since

$$AA^{+} = \mathbb{I}_{n}.$$

Interpolation Limit with Full Row Rank

Back to the normal equations:

$$X^T X \boldsymbol{\beta} = X^T \mathbf{y} \tag{2}$$

When X has full row rank, $\operatorname{rank} X = n$, then we define

$$\boldsymbol{\beta}^+ = X^+ \mathbf{y}.$$

Then

$$X^T X \boldsymbol{\beta}^+ = X^T \underbrace{XX^T (XX^T)^{-1}}_{=\mathbb{I}_n} \mathbf{y} = X^T \mathbf{y}.$$

Hence $\boldsymbol{\beta}^+$ is a solution to (2).

Interpolation Limit

We have the BMP the predictor

$$\widehat{\mathbf{y}}^{+} = X\boldsymbol{\beta}^{+} = XX^{+}\mathbf{y}$$

$$= XX^{T}(XX^{T})^{-1}\mathbf{y} = \mathbf{y}.$$

X has full row rank n, then $\widehat{\mathbf{y}}^{+}$ is an interpolation of the training set.

More on generalized inverses

We realize also that if $G = \left(X^T X\right)^-$ is any generalized inverse of $X^T X$, then

$$\widehat{\boldsymbol{\beta}}^{\dagger} = G\mathbf{y}$$

solves the normal equations (2). We shall next study further expressions that will contain $\left(X^T X\right)^-$.

More on generalized inverses

The condition defining a generalized inverse of a matrix A in (1) is $AGA = A$. Then

$$A = AGA = (AGA)GA = A(GAG)A$$

and hence this implies

$$GAG = G. \tag{3}$$

Conversely, $GAG = G$ implies $AGA = G$.

MORE ON GENERALIZED INVERSES

PROPOSITION

X is $n \times k$ and $\operatorname{rank} X = r \leq n < k$. A generalized inverse of X is

$$X^{-} = \left(X^T X\right)^{-} X^T. \tag{4}$$

Proof Set $G = \left(X^T X\right)^{-} X^T$. First, X is an $n \times k$ matrix and G is an $k \times n$ matrix. We check (3).Then

$$GXG = \left(X^T X\right)^{-} X^T X \left(X^T X\right)^{-} X^T = \underbrace{\left(\left(X^T X\right)^{-} X^T X \left(X^T X\right)^{-}\right)}_{=\left(X^T X\right)^{-}} X^T$$

by (3), and thus

$$GXG = \left(X^T X\right)^{-} X^T = G.$$

and we invoke the condition in (3) to prove the assertion as claimed.

More on generalized inverses

$X^{-} = (X^T X)^{-} X^T$. Hence it follows by the definition in (1)

$$XGX = X \Leftrightarrow X = X\left(X^T X\right)^{-} X^T X. \qquad (5)$$

We observe also

1. $XX^{-} = X(X^T X)^{-} X^T$ is symmetric.
2. $\text{rank}\, XX^{-} = r$
3. $X(X^T X)^{-} X^T$ is the same independently of what $(X^T X)^{-}$ is used.

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \boldsymbol{\epsilon},$$

where $\operatorname{rank} X = r \leq n \leq k$. $\widehat{\boldsymbol{\beta}}^{\dagger} = \left(X^T X\right)^{-} X^T \mathbf{y}$. We have $\| \mathbf{x} \| = \sqrt{\sum_{i=1}^{n} x_i^2}$. We want to study

$$\frac{1}{n} \mathsf{E} \| X\widehat{\boldsymbol{\beta}}^{\dagger} - X\boldsymbol{\beta}_* \|^2 .$$

.

$$X\widehat{\boldsymbol{\beta}}^{\dagger} - X\boldsymbol{\beta}_* = X\left(X^T X\right)^{-} X^T \mathbf{y} - X\boldsymbol{\beta}_*$$

$$= \underbrace{X\left(X^T X\right)^{-} X^T X}_{=X \quad \text{by (5)}} \boldsymbol{\beta}_* - X\boldsymbol{\beta}_* + X\left(X^T X\right)^{-} X^T \boldsymbol{\epsilon}$$

$$= X\left(X^T X\right)^{-} X^T \boldsymbol{\epsilon}$$

Hence

$$\frac{1}{n} \mathrm{E} \parallel X\widehat{\boldsymbol{\beta}}^{\dagger} - X\boldsymbol{\beta}_{*} \parallel^2 =$$

$$= \frac{1}{n} \mathrm{E} \parallel X \left(X^T X \right)^{-} X^T \boldsymbol{\epsilon} \parallel^2 .$$

$$= \frac{1}{n} \mathrm{Tr}\, \mathrm{E} \left[X \left(X^T X \right)^{-} X^T \boldsymbol{\epsilon} \boldsymbol{\epsilon}^T X \left(X^T X \right)^{-} X^T \right]$$

where we used the property $\left((X^T X)^{-} \right)^T = \left((X^T X)^T \right)^{-} = (X^T X)^{-}$, so that

$$= \frac{\sigma^2}{n} X \underbrace{\left(X^T X \right)^{-} X^T X \left(X^T \right)^{-}}_{=\left(X^T X \right)^{-} \quad \text{by (3)}} X^T$$

$$= \frac{\sigma^2}{n} \mathrm{Tr}\, X \left(X^T X \right)^{-} X^T$$

$$= \frac{\sigma^2}{n} \mathrm{Tr}\, X X^{-} = \frac{\sigma^2}{n} r.$$

$$H^{-} = X\left(X^{T}X\right)^{-}X^{T}$$

is the orthogonal projector onto the range of X