Some Key Linear Models Results

Ulrike Genschel

January 23, 2025

A General Linear Model (GLM)

Suppose

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \text{where} \tag{1}$$

- $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector,
- $\boldsymbol{X}$ is an $n \times p$ matrix of known/fixed constants,

known

unknown

- $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown parameter vector, and
- $\boldsymbol{\epsilon}$ is a vector of unobserved random “errors” satisfying $\mathrm{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}$.

The model is called a linear model because the mean of the response vector $\cancel{\boldsymbol{Y}}$ is linear in the unknown parameter vector $\boldsymbol{\beta}$. $(\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta})$

y

A General Linear Model

- This GLM says simply that $\boldsymbol{y}$ is a random vector with expectation $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^p$.
- The distribution of $\boldsymbol{y}$ is left unspecified but generally depends on the distribution of $\boldsymbol{\epsilon}$.
- Goal: estimate $\mathrm{E}(\boldsymbol{y})$
- Available: <u>observed</u> values of $\boldsymbol{y}$ and $\boldsymbol{X}$,
- Estimate $\boldsymbol{X}\boldsymbol{\beta}$, which by definition corresponds to the mean of $\boldsymbol{y}$, i.e., $\mathrm{E}(\boldsymbol{y})$.

Examples

There are many special cases of (1) depending on the distribution of ϵ, the structure of the Σ, and the rank and the structure of $\boldsymbol{X}$.

We will start out by considering the following two cases generally known as the Gauss-Markov Model:

GMM

GMMNE

1. the distribution of ϵ is Normal with $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \sigma^2 \boldsymbol{I}$, where $\sigma^2 > 0$ is unknown; $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$
2. the distribution of ϵ is unknown with $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \sigma^2 \boldsymbol{I}$, where $\sigma^2 > 0$ is unknown

We will later relax the form of $\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ to allow for more flexibility, e.g., $\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \sigma^2 \boldsymbol{V}$, where $\boldsymbol{V}$ is known and $\sigma^2 > 0$ is unknown. This model is known as the Aitken model.

Ordinary Least Squares (OLS) Estimation

Suppose $\quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathrm{E}(\boldsymbol{\epsilon}) = 0, \quad \mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$

column space of X

- $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta} \in \mathcal{C}(\boldsymbol{X})$ with $\boldsymbol{\beta}$ unknown, $\boldsymbol{X}$ is full-rank
- To estimate $\mathrm{E}(\boldsymbol{y})$, consider $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$. estimate : $\hat{y}$
- To estimate $\mathrm{E}(\boldsymbol{y})$, find the vector in $\mathcal{C}(\boldsymbol{X})$ that is closest to $\boldsymbol{y}$.
- Let $\mathcal{N}(\boldsymbol{X}^\top)$ denote the null space of $\boldsymbol{X}^\top$ and note that $\mathcal{N}(\boldsymbol{X}^\top)$ and $\mathcal{C}(\boldsymbol{X})$ are orthogonal to each other, i.e., $\mathcal{N}(\boldsymbol{X}^\top) \perp \mathcal{C}(\boldsymbol{X})$

residuals $\hat{e}$ are in the null space of X^T

The null space of a matrix $\boldsymbol{A}$, denoted by $\mathcal{N}(\boldsymbol{A})$, is given as

$\mathcal{N}(\boldsymbol{A}) = \{\boldsymbol{x} : \boldsymbol{x}\boldsymbol{A} = \boldsymbol{0}.\}$

allowing for orthogonal decomposition of y into $\hat{y} + \hat{e}$

Ordinary Least Squares (OLS) Estimation

ordinary

An estimate $\widehat{\boldsymbol{\beta}}$ is a **least squares estimate** (LSE) of $\boldsymbol{\beta}$ if $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is the vector in $\mathcal{C}(\boldsymbol{X})$ that is closes to $\boldsymbol{y}$

$$\widehat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta} \in \mathbb{R}^p} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Method of least squares identifies the value of $\boldsymbol{\beta}$ for which the squared Euclidean norm of the residual vector, i.e., **error sum of squares**

$$\mathcal{Q}(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

is minimized.

Ordinary Least Squares (OLS) Estimation

There exist two distinct ways to identify the LSE:

- algebraically: normal equations
- geometrically: orthogonal projection of $\boldsymbol{y}$ onto $\mathcal{C}(\boldsymbol{X})$

OLS Estimation: Normal Equations

Recall that the method of least squares seeks the β that minimizes the Euclidean norm of the residual vector

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\beta}) &= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}. \end{aligned}$$

To find the minimum, we take the derivative and set the gradient equal to the null vector

$$\nabla\mathcal{Q}(\boldsymbol{\beta}) = -2\boldsymbol{X}^\top\boldsymbol{y} + 2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$$

leading to the **normal equations**

$$\boxed{\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top\boldsymbol{y}.} \tag{2}$$

OLS Estimation: Solutions to the Normal Equations

The normal equations

$$\underline{\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{y}}$$

have $\boxed{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}}$ as the **unique** solution for $\boldsymbol{\beta}$ if $rank(\boldsymbol{X}) = p$.

$(X^T X)^{-1}$ unique inverse when X is full rank

The normal equations have infinitely many solutions for $\boldsymbol{\beta}$ if $rank(\boldsymbol{X}) < p$.

generalized inverse

While $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-}\boldsymbol{X}^\top \boldsymbol{y}$ may not always be a unique solution, $\boxed{\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{y}}}$ will be unique.

OLS Estimation: Geometric Approach

Let $\boldsymbol{P_X}$ denote the orthogonal projection matrix onto $\mathcal{C}(\boldsymbol{X})$

$$\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^- \boldsymbol{X}^\top.$$

Properties:

- $\boldsymbol{P_X}$ is idempotent, (i.e., $\boldsymbol{P_X P_X} = \boldsymbol{P_X}$)
- $\boldsymbol{P_X}$ projects onto $\mathcal{C}(\boldsymbol{X})$
- $\boldsymbol{P_X}$ is invariant to the choice of $(\boldsymbol{X}^\top \boldsymbol{X})^-$, i.e., it is the same matrix for all generalized inverses $(\boldsymbol{X}^\top \boldsymbol{X})^-$ of $\boldsymbol{X}^\top \boldsymbol{X}$
- $\boldsymbol{P_X}$ is symmetric (i.e., $\boldsymbol{P_X} = \boldsymbol{P_X}^\top$) and unique
- $\boldsymbol{P_X X} = \boldsymbol{X}$ and $\boldsymbol{X}^\top \boldsymbol{P_X} = \boldsymbol{X}^\top$. trace
- $rank(\boldsymbol{X}) = rank(\boldsymbol{P_X}) = tr(\boldsymbol{P_X})$.

OLS Estimation: Geometric Approach

An estimate $\widehat{\boldsymbol{\beta}}$ is a least squares estimate if and only if

$$\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{P_X y}.$$

projecting y onto
Column space of X

The OLS Estimator of $\mathrm{E}(\boldsymbol{y})$ is thus given by

$$\boldsymbol{P_X y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \equiv \underline{\underline{\widehat{\boldsymbol{y}}}} = \widehat{E(y)} \qquad (3)$$

because $\boldsymbol{P_X y} \in \mathcal{C}(\boldsymbol{X})$ and

$$||\boldsymbol{y} - \boldsymbol{P_X y}||^2 < ||\boldsymbol{y} - \boldsymbol{z}||^2 \ \forall\ \boldsymbol{z} \in \mathcal{C}(\boldsymbol{X}) \setminus \{\boldsymbol{P_X y}\}.$$

Even when $\widehat{\boldsymbol{\beta}}$ is not unique, $\boldsymbol{P_X y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{y}}$ always will.

OLS Estimation: Fitted Values

$\widehat{\boldsymbol{y}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$ is the vector of fitted values. Recall that geometrically, $\widehat{\boldsymbol{y}}$ is the point in $\mathcal{C}(\boldsymbol{X})$ that is closest to $\boldsymbol{y}$. Now, note that $\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}$ is the perpendicular projection matrix onto $\mathcal{N}(\boldsymbol{X}^\top)$ and

$$(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{y} - \widehat{\boldsymbol{y}} \equiv \widehat{\mathrm{e}}.$$

$\widehat{\mathrm{e}}$ is the vector of **residuals** and $\widehat{\mathrm{e}} \in \mathcal{N}(\boldsymbol{X}^\top)$. Because $\mathcal{C}(\boldsymbol{X})$ and $\mathcal{N}(\boldsymbol{X}^\top)$ are orthogonal complements, we can uniquely decompose $\boldsymbol{y}$ as

$$\boldsymbol{y} = \widehat{\boldsymbol{y}} + \widehat{\mathrm{e}}.$$

OLS Estimation: Orthogonal Decomposition of $\boldsymbol{y}^\top \boldsymbol{y}$

Projects onto C(X)

Projects onto N(X^T)

We know that $\widehat{\boldsymbol{y}}$ and $\widehat{\mathrm{e}}$ are orthogonal vectors. Thus,

adding zero

$$
\begin{aligned}
\boldsymbol{y}^\top \boldsymbol{y} = \boldsymbol{y}^\top \boldsymbol{I} \boldsymbol{y} &= \boldsymbol{y}^\top (\boldsymbol{P_X} + \boldsymbol{I} - \boldsymbol{P_X}) \boldsymbol{y} \\
&= \boldsymbol{y}^\top \boldsymbol{P_X} \boldsymbol{y} + \boldsymbol{y}^\top (\boldsymbol{I} - \boldsymbol{P_X}) \boldsymbol{y} \\
&= \boldsymbol{y}^\top \boldsymbol{P_X} \boldsymbol{P_X} \boldsymbol{y} + \boldsymbol{y}^\top (\boldsymbol{I} - \boldsymbol{P_X})(\boldsymbol{I} - \boldsymbol{P_X}) \boldsymbol{y} \\
&= \widehat{\boldsymbol{y}}^\top \widehat{\boldsymbol{y}} + \widehat{\mathrm{e}}^\top \widehat{\mathrm{e}},
\end{aligned}
$$

P_X

since $\boldsymbol{P_X}$ and $(\boldsymbol{I} - \boldsymbol{P_X})$ are both symmetric and idempotent.

Orthogonal Decomposition of $\boldsymbol{y}^\top \boldsymbol{y}$ & ANOVA Table

This orthogonal decomposition of $\boldsymbol{y}^\top \boldsymbol{y}$ is often given in a tabular display called an analysis of variance (ANOVA) table.

Suppose $\boldsymbol{y}$ is $n \times 1$, $\boldsymbol{X}$ is $n \times p$ with rank $r \leq p$, $\boldsymbol{\beta}$ is $p \times 1$, and $\boldsymbol{\epsilon}$ is $n \times 1$. We assume the the model given in (1): $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Then, the ANOVA table looks as follows

Source	df	Sum of Squares
Model	r	$\widehat{\boldsymbol{y}}^\top \widehat{\boldsymbol{y}} = \boldsymbol{y}^\top \boldsymbol{P_X} \boldsymbol{y}$
Residual	$n - r$	$\widehat{\mathbf{e}}^\top \widehat{\mathbf{e}} = \boldsymbol{y}^\top (\boldsymbol{I} - \boldsymbol{P_X}) \boldsymbol{y}$
Total	$n - 1$	$\boldsymbol{y}^\top \boldsymbol{y} = \boldsymbol{y}^\top \boldsymbol{I} \boldsymbol{y}$

Table: ANOVA Table

The OLS Estimator of a Linear Function of $\mathrm{E}(\boldsymbol{y})$

For any $q \times n$ matrix $\boldsymbol{A}$, $\boldsymbol{A}\mathrm{E}(\boldsymbol{y})$ is a linear function of $\mathrm{E}(\boldsymbol{y})$.

For any $q \times n$ matrix $\boldsymbol{A}$, the OLS Estimator of $\boldsymbol{A}\mathrm{E}(\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}$ is

$$\begin{aligned}\boldsymbol{A}\left[\text{OLS Estimator of } \mathrm{E}(\boldsymbol{y})\right] &= \boldsymbol{A}\hat{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} \\ &= \boldsymbol{A}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{y}.\end{aligned}$$

- $\boldsymbol{A}\mathrm{E}(\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}$ is automatically a linear function of $\boldsymbol{\beta}$ of the form $\boldsymbol{C}\boldsymbol{\beta}$, where $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{X}$.
- If $\boldsymbol{C}$ is any $q \times p$ matrix, we say that the linear function of $\boldsymbol{\beta}$ given by $\boldsymbol{C}\boldsymbol{\beta}$ is estimable if and only if $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{X}$ for some matrix $q \times n$ matrix $\boldsymbol{A}$.
- The OLS Estimator of an estimable linear function $\boldsymbol{C}\boldsymbol{\beta}$ is $\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{y}$.

Uniqueness of the OLS Estimator of an Estimable $C\beta$

If $\boldsymbol{C}\boldsymbol{\beta}$ is estimable, then $\boldsymbol{C}\widehat{\boldsymbol{\beta}}$ is the same for all solutions $\widehat{\boldsymbol{\beta}}$ to the Normal Equations.

In particular, the unique OLS Estimator of $\boldsymbol{C}\boldsymbol{\beta}$ is

$$\boldsymbol{C}\widehat{\boldsymbol{\beta}} = \boldsymbol{C}\underbrace{\boxed{(\boldsymbol{X}^\top \boldsymbol{X})^- \boldsymbol{X}^\top \boldsymbol{y}}}_{\widehat{\beta}} = \boldsymbol{A}\boxed{\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^- \boldsymbol{X}^\top} \boldsymbol{y} = \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y},$$

where $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{X}$.

The OLS Estimator is a Linear Unbiased Estimator

If $C\beta$ is estimable, then $C\widehat{\beta}$ is a linear unbiased estimator of $C\beta$.

The OLS Estimator is a linear estimator because it is a linear function of $\boldsymbol{y}$:

$$\boldsymbol{C}\widehat{\boldsymbol{\beta}} = \boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y}, \text{ where } \boldsymbol{M} = \boxed{\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top.}$$

The OLS Estimator is unbiased because, for all $\boldsymbol{\beta} \in \mathbb{R}^p$,

My

constant

$$\begin{aligned} \mathrm{E}(\boldsymbol{C}\widehat{\boldsymbol{\beta}}) &= \mathrm{E}(\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{y}) = \boxed{\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top}\mathrm{E}(\boldsymbol{y}) \\ &= \boldsymbol{A}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\mathrm{E}(\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{X}}\boxed{\mathrm{E}(\boldsymbol{y})} = X\beta \\ &= \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{\beta}. \end{aligned}$$

$C = AX$

Slide 10

The Gauss-Markov Model (GMM)

Suppose $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

- $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector,
- $\boldsymbol{X}$ is an $n \times p$ matrix of known constants,
- $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown parameter vector, and
- $\boldsymbol{\epsilon}$ is a vector of random “errors” satisfying $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$ for some unknown variance parameter $\sigma^2 \in \mathbb{R}^+$.

The GMM is a Special Case of the GLM

The GMM is a special case of the GLM presented previously.

We have added the assumption $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$; i.e., we assume the errors are uncorrelated and have constant variance.

All the results presented for the GLM hold for the GMM.

For the GMM, we have an additional result provided by the *Gauss-Markov Theorem*:

The Gauss-Markov Theorem

The OLS Estimator of an estimable function $C\beta$ is the

Best Linear Unbiased Estimator $\boxed{\textit{(BLUE)}}$ of $C\beta$

in the sense that the OLS Estimator $C\widehat{\beta}$ has the smallest variance among all linear unbiased estimators of $C\beta$.

end lecture

2

