SMIII lectures

Week 12

The Geometry of Least Squares

- The discussion of linear models so far has been formulated in terms of matrices.
- This is adequate from a computational perspective.
- But there is an important logical gap to be filled.
- We know the same model can be specified with different model matrices.
- For example, when dealing with factors, we can use different constraints and we showed that the models were equivalent.

The Geometry of Least Squares

We need to answer the question:

What is the unique definition of a linear model?

Overview

- In this section, we use concepts of linear algebra and, in particular, vector subspaces to provide a precise specification of the linear model.
- Hypothesis testing and the analysis of variance will be described in terms of linear algebra.
- We will consider only n-dimensional Euclidean space.
- Much of the theory extends to abstract vector spaces, but this is not considered here.

Subspaces

$$\mathbb{R}^n = \left\{ \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} : x_i \in \mathbb{R} \right\}.$$

Vector Addition

$$\boldsymbol{x} + \boldsymbol{y} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix}$$

for

$$\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

Scalar Multiplication

$$a\boldsymbol{x} = \begin{bmatrix} ax_1 \\ ax_2 \\ \vdots \\ ax_n \end{bmatrix}$$

for

$$a \in \mathbb{R}, \;\; \boldsymbol{x} \in \mathbb{R}^n.$$

The dot product

The **dot product** or **inner product** of two vectors is defined in the usual way.

$$\boldsymbol{x}.\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{y} = \sum_{i=1}^{n} x_i y_i.$$

The norm

The dot product can be used to define distances, via the **norm**,

$$\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}.\boldsymbol{x}} = \sqrt{\sum_{i=1}^{n} x_i^2},$$

so that the **Euclidean distance** between two points is

$$\|\boldsymbol{x} - \boldsymbol{y}\|.$$

Orthogonal vectors

The dot product is also used to define the angle, θ between vectors via the relation

$$\frac{\boldsymbol{x}.\boldsymbol{y}}{\|\boldsymbol{x}\|.\|\boldsymbol{y}\|} = \cos\theta.$$

The vectors $\boldsymbol{x}$, $\boldsymbol{y}$ are said to be **orthogonal** or **perpendicular** if

$$\boldsymbol{x}.\boldsymbol{y} = 0.$$

In this case, we write

$$\boldsymbol{x} \perp \boldsymbol{y}.$$

Definition 12.1: Vector Subspaces

A set

$$\mathcal{M} \subseteq \mathbb{R}^n$$

is said to be a **vector subspace** if

$$a\boldsymbol{x} + b\boldsymbol{y} \in \mathcal{M}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$ and $a, b \in \mathbb{R}$.

$\square$

Example 1

$$\mathcal{M}_1 \subset \mathbb{R}^3$$

defined by

$$\mathcal{M}_1 = \left\{ \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix} : x_1, x_2 \in \mathbb{R} \right\}.$$

Example 2

Let X be an $n \times p$ matrix and let

$$\mathcal{M}_2 = \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = X\boldsymbol{\beta} \text{ for } \boldsymbol{\beta} \in \mathbb{R}^p\}.$$

Definition 12.2: Span

Let

$$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p\} \subset \mathbb{R}^n$$

be a set of vectors.

The **span** is defined by

$$\mathcal{S}\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p\} = \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + \ldots + a_p\boldsymbol{x}_p\}$$

for

$$a_1, a_2, \ldots, a_p \in \mathbb{R}.$$

□

The span,

$$\mathcal{S}\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p\},$$

is always a vector subspace.

The vectors

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$$

is said to span the subspace $\mathcal{M}$ if

$$\mathcal{M} = \mathcal{S}\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p\}.$$

Basis

Recall that a set of vectors,

$$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p\} \subset \mathbb{R}^n$$

is said to be linearly independent if

$$\sum_{i=1}^{p} a_i \boldsymbol{x}_i = \mathbf{0} \Rightarrow a_1 = a_2 = \ldots = a_p = 0.$$

Definition 12.3

Let

$$\mathcal{M} \subseteq \mathbb{R}^n$$

be a vector subspace.

A set of linearly independent vectors,

$$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p\},$$

such that

$$\mathcal{M} = \mathcal{S}\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p\}$$

is said to form a **basis** for $\mathcal{M}$.

$\square$

Example 1

The two vectors

$$\left\{ \boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$$

form a basis for $\mathcal{M}_1$.

Example 2

If the columns

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$$

of the matrix, X, are **linearly independent**, then they form a basis for $\mathcal{M}_2$.

Alternative Bases

The basis of a vector subspace is not unique. In our two examples:

- The two vectors

$$\left\{ \boldsymbol{y}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \boldsymbol{y}_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \right\}$$

form a basis for $\mathcal{M}_1$. - For any $p \times p$ invertible matrix, A, the columns of the matrix XA form a basis for $\mathcal{M}_2$, provided the columns of X are linearly independent.

Theorem 12.1: Dimension of a subspace

Suppose

$$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p\}$$

and

$$\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r\}$$

are bases for the vector subspace $\mathcal{M}$. The $r = p$.

$\square$

Definition 12.4

The the vector subspace, $\mathcal{M}$, is said to have dimension, p, if any basis for $\mathcal{M}$ has p elements.

$\square$

Linear Models

Linear Models

Consider data $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ and the statistical model,

$$\boldsymbol{y} = \boldsymbol{\eta} + \boldsymbol{e}$$

where $e_1, e_2, \ldots, e_n$ are such that

$$E(\boldsymbol{e}) = 0 \text{ and } \mathrm{Var}(\boldsymbol{e}) = \sigma^2 I.$$

A model of the form,

$$M : \boldsymbol{\eta} \in \mathcal{M}$$

where $\mathcal{M} \subseteq \mathbb{R}^n$ is called a linear model.

Linear Models

A linear model is specified at the most fundamental level by the subspace $\mathcal{M}$.

In practice, linear models can be specified by a model matrix, X, such that

$$\boldsymbol{\eta} \in \mathcal{M} \Leftrightarrow \boldsymbol{\eta} = X\boldsymbol{\beta}.$$

That is, where the columns of X form a basis for $\mathcal{M}$.

We have seen previously examples where the same model can specified with different model matrices.

In general two models of the form

$$\boldsymbol{\eta} = X\boldsymbol{\beta} \text{ and } \boldsymbol{\eta} = X^*\boldsymbol{\beta}^*$$

will be equivalent when the columns of X and X^* are bases for the same subspace $\mathcal{M}$.

The projection matrix

In the matrix treatment of the linear model, the projection matrix

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

plays a key role.

P can also be defined as a projection operator on the underlying subspace without reference to a model matrix.

This is an important result because it implies that the projection P obtained from X and X^* is the same.

Definition 12.5 Orthogonal Complement

The **orthogonal complement** of the vector subspace $\mathcal{M}$ is the set

$$\mathcal{M}^{\perp} = \{\boldsymbol{v} \in \mathbb{R}^{\boldsymbol{n}} : \boldsymbol{v} \perp \boldsymbol{x} \text{ for all } \boldsymbol{x} \in \mathcal{M}\}.$$

$\square$

Theorem 12.2

If $\mathcal{M}$ is a vector subspace of dimension p, then $\mathcal{M}^{\perp}$ is a vector subspace of dimension $n - p$.

$\square$

Example

Let $\mathcal{M}_1$ be as previously.

$$\mathcal{M}_1 = \left\{ \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix} : x_1, x_2 \in \mathbb{R} \right\}.$$

$$\mathcal{M}_1^{\perp} = \left\{ \boldsymbol{v} \in \mathbb{R}^3 : \boldsymbol{v}.\boldsymbol{x} = 0 \text{ for all } \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix} \right\}.$$

$$\boldsymbol{v}.\boldsymbol{x} = x_1 v_1 + x_2 v_2 = 0 \text{ for all } x_1, x_2 \Leftrightarrow v_1 = v_2 = 0.$$

That is,

$$\mathcal{M}_1^{\perp} = \left\{ \boldsymbol{v} = \begin{bmatrix} 0 \\ 0 \\ v_3 \end{bmatrix} : v_3 \in \mathbb{R} \right\} = \mathcal{S}\left\{ \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

and $\dim(\mathcal{M}_1^{\perp}) = 1 = 3 - 2$ as required.

Orthogonal Projection

Let $\mathcal{M} \subseteq \mathbb{R}^n$ be a vector subspace.

It can be proved that every $\boldsymbol{y} \in \mathbb{R}^n$ can be expressed **uniquely** as

$$\boldsymbol{y} = \boldsymbol{u} + \boldsymbol{v} \text{ where } \boldsymbol{u} \in \mathcal{M},\ \boldsymbol{v} \in \mathcal{M}^{\perp}.$$

Definition 12.6

The orthogonal projection P on the subspace $\mathcal{M}$ is the linear mapping $P : \mathbb{R}^n \mapsto \mathcal{M}$ defined by

$$P\boldsymbol{y} = \boldsymbol{u}.$$

$\square$

Note $I - P$ is the orthogonal projection on $\mathcal{M}^\perp$.

Example

Let $\mathcal{M}_1 \subset \mathbb{R}^3$ be as before and let P be the orthogonal projection.

Consider

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \in \mathbb{R}^3.$$

The unique decomposition of

$$\boldsymbol{y} = \boldsymbol{u} + \boldsymbol{v}$$

is

$$\boldsymbol{u} = \begin{bmatrix} y_1 \\ y_2 \\ 0 \end{bmatrix} \in \mathcal{M}_1 \text{ and } \boldsymbol{v} = \begin{bmatrix} 0 \\ 0 \\ y_3 \end{bmatrix} \in \mathcal{M}_1^{\perp}.$$

The orthogonal projection is

$$P\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ 0 \end{bmatrix}$$

Theorem 12.3 The least squares property

Let P be the orthogonal projection on $\mathcal{M}$ and consider $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{w} \in \mathcal{M}$.

Then

$$\|\boldsymbol{y} - \boldsymbol{w}\|^2 \geq \|\boldsymbol{y} - P\boldsymbol{y}\|^2$$

with equality if and only if $\boldsymbol{w} = P\boldsymbol{y}$.

Theorem 12.4 The normal equations

The linear mapping P is the orthogonal projection on $\mathcal{M}$ if and only if

$$P\boldsymbol{y} \in \mathcal{M} \text{ and } \boldsymbol{y} - P\boldsymbol{y} \in \mathcal{M}^{\perp}$$

for all $\boldsymbol{y} \in \mathbb{R}^n$.

Note, the conditions

$$P\boldsymbol{y} \in \mathcal{M} \text{ and } \boldsymbol{y} - P\boldsymbol{y} \in \mathcal{M}^{\perp}$$

for all $\boldsymbol{y} \in \mathbb{R}^n$ are called the **normal equations**.

Theorem 12.5 Symmetry and Idempotence

Consider an $n \times n$ matrix, P. The mapping $P\boldsymbol{y}$ is the orthogonal projection on the range of P if and only if $P = P^2 = P^T$.

The matrix formulation

Consider the matrix formulation,

$$\boldsymbol{y} = \boldsymbol{\eta} + \boldsymbol{e}$$

such that $E(\boldsymbol{e}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{e}) = \sigma^2 I$, and consider the linear model

$$\boldsymbol{\eta} = X\beta$$

where X an $n \times p$ matrix with linearly independent columns.

- Have seen previously that $P = X(X^T X)^{-1} X^T$ is symmetric and idempotent.
- Moreover the range of P is the column space of X, or the linear subspace $\mathcal{M}$.
- Least squares estimation is the orthogonal projection on $\mathcal{M}$.

Test of Hypotheses and ANOVA

Consider vector subspaces

$$\mathcal{H} \subset \mathcal{M} \subset \mathbb{R}^n$$

such that $p_0 = \dim(\mathcal{H}) < p = \dim(\mathcal{M}) < n$.

Consider the linear model

$$M : \boldsymbol{\eta} \in \mathcal{M}$$

and the hypothesis

$$H_0 : \boldsymbol{\eta} \in \mathcal{H}.$$

The ANOVA table

The analysis of variance table is

Source	SS	DF	MS	F-ratio
H_0 vs M	$\|(P-P_0)\boldsymbol{y}\|^2$	$p-p_0$	$MS_H = \frac{\|(P-P_0)\boldsymbol{y}\|^2}{p-p_0}$	$F = \frac{MS_H}{MS_E}$
Residual	$\|(I-P)\boldsymbol{y}\|^2$	$n-p$	$MS_E = \frac{\|(I-P)\boldsymbol{y}\|^2}{n-p}$	
Total	$\|(I-P_0)\boldsymbol{y}\|^2$	$n-p_0$		

} \end{center}

where P is the orthogonal projection on $\mathcal{M}$ and P_0 is the orthogonal projection on $\mathcal{H}$.

The expected mean squares

The key to understanding ANOVA lies in the expected value of the mean square entries.

Suppose the model $\mathcal{M}$ holds.

We have

$$E(\|\boldsymbol{Y} - X\hat{\beta}\|^2) = E(\|(I - P)\boldsymbol{Y}\|^2) = (n - p)\sigma^2$$

so that

$$E(MS_E) = \sigma^2.$$

The expected mean squares

Consider now the hypothesis sum squares.

$$
\begin{aligned}
\|(P-P_0)\boldsymbol{Y}\|^2 &= \|((P-P_0)(\boldsymbol{Y}-\boldsymbol{\eta})+(P-P_0)\boldsymbol{\eta}\|^2 \\
&= \|((P-P_0)(\boldsymbol{Y}-\boldsymbol{\eta})\|^2+\|(P-P_0)\boldsymbol{\eta}\|^2 \\
&+ 2\boldsymbol{\eta}^T(P-P_0)(\boldsymbol{Y}-\boldsymbol{\eta})
\end{aligned}
$$

Since $P - P_0$ is the orthogonal projection on the vector subspace, $\mathcal{M} \cap \mathcal{H}$,

$$E(\|(P - P_0)(\boldsymbol{Y} - \boldsymbol{\eta})\|^2) = (p - p_0)\sigma^2.$$

Since $\boldsymbol{\eta}$ is a constant,

$$E(\|(P - P_0)\boldsymbol{\eta}\|^2) = \|(P - P_0)\boldsymbol{\eta}\|^2.$$

$$E(\boldsymbol{\eta}^T(P - P_0)(\boldsymbol{Y} - \boldsymbol{\eta})) = \boldsymbol{\eta}^T(P - P_0)E(\boldsymbol{Y} - \boldsymbol{\eta}) = 0$$

Hence

$$E(MS_H) = \sigma^2 + \frac{1}{p - p_0}\|(P - P_0)\boldsymbol{\eta}\|^2.$$

The expected mean squares

Observe that $H_0 : \boldsymbol{\eta} \in \mathcal{H}$ is true if and only if $(P - P_0)\boldsymbol{\eta} = \mathbf{0}$

Hence $\delta^2 = \frac{1}{p-p_0}\|(P - P_0)\boldsymbol{\eta}\|^2$ is a measure of the magnitude of any departure from H_0.

To summarise

$$E(MS_H) = \begin{cases} \sigma^2 & \text{if } H_0 \text{ holds} \\ \sigma^2 + \delta^2 & \text{otherwise.} \end{cases}$$

The F-statistic

The F-statistic is the ratio

$$F = \frac{MS_H}{MS_E}.$$

- When H_0 is true, we would expect the F-statistic to be close to 1.
- When H_0 is not true, we expect the F-statistic to exceed 1.
- The greater the departure from H_0, the larger we expect F-statistic to be.

This explains the logic of the F-test, where H_0 is rejected only for large values of the F-statistic.

It can also be proved that the null distribution of the F-statistic is $F_{p-p_0,n-p}$ but this is beyond the scope of the course.

Geometry of generalised least squares

Geometry of generalised least squares

Generalised least squares also has a geometrical interpretation.

The same theory applies but instead of using the standard inner product (dot product)

$$\boldsymbol{x}_1 . \boldsymbol{x}_2 = \langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle = \boldsymbol{x}_1^T \boldsymbol{x}_2$$

to define angle and distance, we use

$$\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle_* = \boldsymbol{x}_1^T V^{-1} \boldsymbol{x}_2$$

Geometry of generalised least squares

The $\mathrm{orthogonal}_*$ projection

$$P = X(X^T V^{-1} X)^{-1} X^T V^{-1}$$

is idempotent, i.e. $P^2 = P$.

However, symmetry is generalised to the self-adjoint property,

$$\langle P\boldsymbol{x}_1, \boldsymbol{x}_2 \rangle_* = \langle \boldsymbol{x}_1, P\boldsymbol{x}_2 \rangle_*$$

Geometry of generalised least squares

It can be checked that

$$P\boldsymbol{y} = \mathrm{argmin}_{\boldsymbol{\eta} \in \mathcal{L}} \|\boldsymbol{y} - \boldsymbol{\eta}\|_*^2$$

where $\mathcal{L}$ is the column space of X and

$$\|\boldsymbol{x}\|_*^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle_* = \boldsymbol{x}^T V^{-1} \boldsymbol{x}.$$

Geometry of generalised least squares

Finally, the definition of

$$P\boldsymbol{x} = \boldsymbol{v}$$

from the decomposition

$$\boldsymbol{x} = \boldsymbol{v} + \boldsymbol{w}$$

with

$$\boldsymbol{v} \in \mathcal{L} \text{ and } \boldsymbol{w} \in \mathcal{L}^{\perp}$$

is the same.

The only difference is that the space $\mathcal{L}^{\perp}$ is orthogonal with respect to the * inner product.