MAST90104 - Lecture 1

Weichang Yu

Room 108, Old Geology South Bldg
School of Mathematics and Statistics, University of Melbourne

22 Jul, 2024

Content acknowledgement

Course content is largely based on previous set of slides from:

- Tim Brown
- Yao-ban Chan
- Owen Jones
- Susan Wei
- Mingming Gong
- KD Dang

Admin

Ultra-important: remember to read through details on Course Information Sheet (Canvas) soon!

Statistics and Data Science

Statistics is a collection of tools for quantitative research, the main aspects of which are:

Questions ⇒ **Study design** ⇒ **Data collection (Prospective or Experimental Data)** ⇒ **Exploratory data analysis** ⇒ **Modelling**
- Formulation
- Estimation
- Checking
- Inference

⇒ **Conclusions**

Real world situation: many companies have already collected data before they hired a statistician or data scientist.

Data collected without clear purpose (Observational data) ⇒ **Exploratory data analysis** ⇒ **Questions** ⇒ **Study design** ⇒ **Modelling**
- Formulation
- Estimation
- Checking
- Inference

⇒ **Conclusions**

- Estimation is also called "learning" in statistical/machine learning.
- Machine learning focuses more on the prediction performance of a learned model on new test data.
- Statistics cares more about inference, e.g., confidence intervals, hypothesis testing.

Section 1: Linear models (Weeks 1 to 6)

What are linear models?

A linear model is one of many types of models that we can use in the modelling phase.

It assumes that the data variables of interest have a *linear* relationship to other explanatory sets of data (give or take a small amount of error).

Fancy modern methods such as deep neural networks and random forest beat linear models in terms of predictive accuracy. So why should we bother with linear models?

What are linear models?

A linear model is one of many types of models that we can use in the modelling phase.

It assumes that the data variables of interest have a *linear* relationship to other explanatory sets of data (give or take a small amount of error).

Fancy modern methods such as deep neural networks and random forest beat linear models in terms of predictive accuracy. So why should we bother with linear models?

- Linear models are more interpretable than many modern methods: easy to understand how change in explanatory is associated with change in variable of interest.

The Linear Model

We have n subjects (or objects), for each we observe a measurement (or a property) Y_i, $i = 1, \ldots, n$. Our aim is to analyse or predict the behaviour of Y.

- The Y's are *random variables*. Whether Y_i is a random variable, a value or data will depend on context.

Each subject also has $k > 0$ other properties that we know or have pre-determined (x variables). We denote these properties as:

$x_{i1}, x_{i2}, \ldots, x_{ik}$.

The Linear Model

We have n subjects (or objects), for each we observe a measurement (or a property) Y_i, $i = 1, \ldots, n$. Our aim is to analyse or predict the behaviour of Y.

- The Y's are *random variables*. Whether Y_i is a random variable, a value or data will depend on context.

Each subject also has $k > 0$ other properties that we know or have pre-determined (x variables). We denote these properties as: $x_{i1}, x_{i2}, \ldots, x_{ik}$.

- In practice, the x's might also be random but we condition on their values in the estimation and inference. For example, (x_1, Y_1) might be the height and weight of a person - our model predicts a person's weight given their height.

The Linear Model

The general (as opposed to generalized - to be studied in GLMs) linear model is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i$$

for all $i = 1, 2, \ldots, n$.

The Linear Model

The general (as opposed to generalized - to be studied in GLMs) linear model is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i$$

for all $i = 1, 2, \ldots, n$.

We call Y the *dependent* variable (or outcome variable) and the x's the *predictor* (or explanatory) variables.

The β's are *coefficients* of the model, and ϵ is a random *error* term. We assume ϵ has mean 0 and variance σ^2 (don't need normal distribution assumption for now).

The Linear Model

The model attempts to explain the variation in Y's using the predictors $(x_1, \ldots, x_k)$.

The Linear Model

The model attempts to explain the variation in Y's using the predictors $(x_1, \ldots, x_k)$.

However, not all variation can be explained by deterministic data alone. The error term ϵ captures the unexplained variation in the population.

Plant data: variation explained without predictor

Consider the dataset

Moisture (x)	Height (Y)
204	22
121	13
261	24
460	35
468	29
299	27
308	29
235	18
188	23

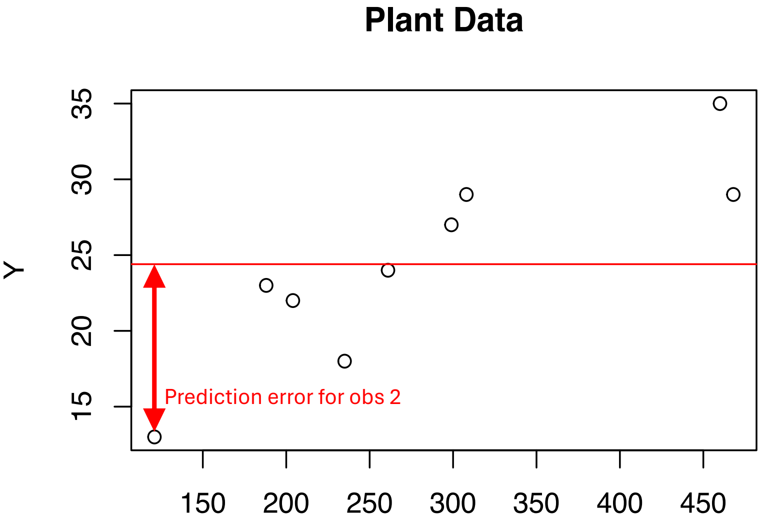
Predicted height without using moisture
$= \overline{Y} = (22 + 13 \ldots + 23)/9 = 24.4$

Plant data: variation explained without predictor

Consider the dataset

Moisture (x)	Height (Y)	Pred. w/o x
204	22	24.4
121	13	24.4
261	24	24.4
460	35	24.4
468	29	24.4
299	27	24.4
308	29	24.4
235	18	24.4
188	23	24.4

Plant data: variation explained without predictor



Plant data: variation explained with predictor

Consider the dataset

Moisture (x)	Height (Y)
204	22
121	13
261	24
460	35
468	29
299	27
308	29
235	18
188	23

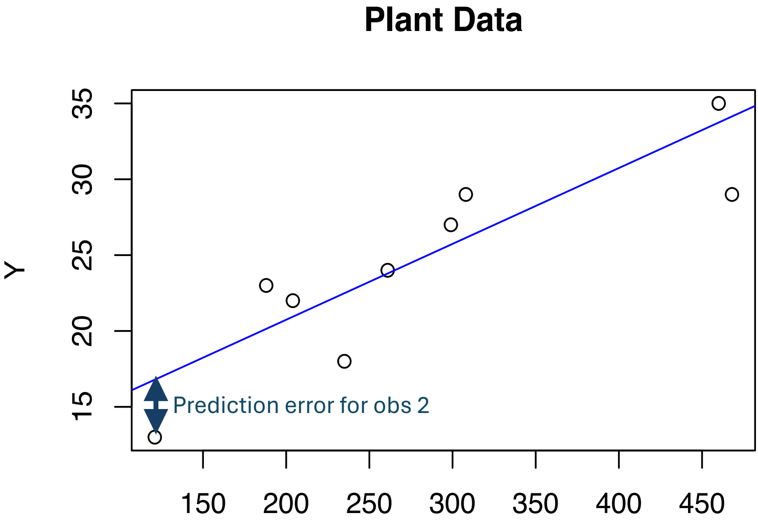
Predicted height $= \widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$. From MAST90105's least squares regression estimate , $\widehat{\beta}_0 = 10.74$ and $\widehat{\beta}_1 = 0.05$ (recap later).

Plant data: variation explained with predictor

Consider the dataset

Moisture (x)	Height (Y)	Pred. with x
204	22	20.6
121	13	16.6
261	24	23.4
460	35	33.0
468	29	33.4
299	27	25.2
308	29	25.7
235	18	22.1
188	23	19.9

Plant data: variation explained with predictor



Recap of estimation for single-predictor regression ($k = 1$)

Consider the model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\mathbb{E}(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$.

Recap of estimation for single-predictor regression ($k = 1$)

Consider the model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\mathbb{E}(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$.

We estimate β_0 and β_1 with the least squares estimate $\widehat{\beta}_0$ and $\widehat{\beta}_1$, where $(\widehat{\beta}_0, \widehat{\beta}_1)$ minimises

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2,$$

i.e.,

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \underset{(\beta_0, \beta_1)}{\text{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2.$$

Recap of estimation for single-predictor regression ($k = 1$)

Consider the model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$.
We estimate β_0 and β_1 with the least squares estimate $\widehat{\beta}_0$ and $\widehat{\beta}_1$, where $(\widehat{\beta}_0, \widehat{\beta}_1)$ minimises

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2,$$

i.e.,

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \underset{(\beta_0, \beta_1)}{\mathrm{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2.$$

The solution (after making assumptions which we'll cover later):

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x}, \quad \widehat{\beta}_0 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

Recap of estimation for single-predictor regression ($k = 1$)

In plant data, we have $\overline{Y} = 24.4444$, $\overline{x} = 283.6667$, $\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) = 5464.333$, and $\sum_{i=1}^{n}(x_i - \overline{x})^2 = 113996$. Hence,

$$\widehat{\beta}_1 = 5464.333/113996 = 0.05,$$

and

$$\widehat{\beta}_0 = 24.4444 - (5464.333/113996) \times 283.6667 = 10.74.$$

Caution!

In general, prediction error of observation $i \neq \epsilon_i$.

Caution!

In general, prediction error of observation $i \neq \epsilon_i$.

Random error term:

$$\epsilon_i = Y_i - \beta_0 - \beta_1 x_i,$$

where β_0 and β_1 are population regression coefficients.

Caution!

In general, prediction error of observation $i \neq \epsilon_i$.

Random error term:

$$\epsilon_i = Y_i - \beta_0 - \beta_1 x_i,$$

where β_0 and β_1 are population regression coefficients.

Prediction error:

$$e_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i,$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are estimates of β_0 and β_1 based on your sample.

Multiple predictors linear regression ($k > 1$)

The general linear model is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i$$

for all $i = 1, 2, \ldots, n$.

Multiple predictors linear regression ($k > 1$)

The general linear model is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i$$

for all $i = 1, 2, \ldots, n$.

The least squares estimate of the coefficients is

$$(\widehat{\beta}_0, \ldots, \widehat{\beta}_k) = \underset{(\beta_0, \beta_1, \ldots, \beta_k)}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2.$$

Multiple predictors linear regression ($k > 1$)

For $k = 2$ the solution is:

$$\widehat{\beta}_1 = \frac{S_{x_1 y} S_{x_2 x_2} - S_{x_2 y} S_{x_1 x_2}}{S_{x_1 x_1} S_{x_2 x_2} - S_{x_1 x_2}^2}, \quad \widehat{\beta}_2 = \frac{S_{x_2 y} S_{x_1 x_1} - S_{x_1 y} S_{x_1 x_2}}{S_{x_1 x_1} S_{x_2 x_2} - S_{x_1 x_2}^2},$$

and $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x}_1 - \widehat{\beta}_2 \overline{x}_2$. Here, $S_{x_1 x_2} = \sum_{i=1}^{n} x_{i1} x_{i2} - n\overline{x}_1 \overline{x}_2$, $S_{x_1 Y} = \sum_{i=1}^{n} x_{i1} Y_i - n\overline{x}_1 \overline{Y}$, $S_{x_2 Y} = \sum_{i=1}^{n} x_{i2} Y_i - n\overline{x}_2 \overline{Y}$,.....*and you get the pattern there.....*

Multiple predictors linear regression ($k > 1$)

For $k = 2$ the solution is:

$$\widehat{\beta}_1 = \frac{S_{x_1 y} S_{x_2 x_2} - S_{x_2 y} S_{x_1 x_2}}{S_{x_1 x_1} S_{x_2 x_2} - S_{x_1 x_2}^2}, \quad \widehat{\beta}_2 = \frac{S_{x_2 y} S_{x_1 x_1} - S_{x_1 y} S_{x_1 x_2}}{S_{x_1 x_1} S_{x_2 x_2} - S_{x_1 x_2}^2},$$

and $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x}_1 - \widehat{\beta}_2 \overline{x}_2$. Here, $S_{x_1 x_2} = \sum_{i=1}^n x_{i1} x_{i2} - n\overline{x}_1 \overline{x}_2$, $S_{x_1 Y} = \sum_{i=1}^n x_{i1} Y_i - n\overline{x}_1 \overline{Y}$, $S_{x_2 Y} = \sum_{i=1}^n x_{i2} Y_i - n\overline{x}_2 \overline{Y}$,*.....and you get the pattern there.....*

Larger k will lead to an even more complicated looking expression!

Multiple predictors linear regression ($k > 1$)

For $k = 2$ the solution is:

$$\widehat{\beta}_1 = \frac{S_{x_1 y} S_{x_2 x_2} - S_{x_2 y} S_{x_1 x_2}}{S_{x_1 x_1} S_{x_2 x_2} - S_{x_1 x_2}^2}, \quad \widehat{\beta}_2 = \frac{S_{x_2 y} S_{x_1 x_1} - S_{x_1 y} S_{x_1 x_2}}{S_{x_1 x_1} S_{x_2 x_2} - S_{x_1 x_2}^2},$$

and $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x}_1 - \widehat{\beta}_2 \overline{x}_2$. Here, $S_{x_1 x_2} = \sum_{i=1}^n x_{i1} x_{i2} - n\overline{x}_1 \overline{x}_2$, $S_{x_1 Y} = \sum_{i=1}^n x_{i1} Y_i - n\overline{x}_1 \overline{Y}$, $S_{x_2 Y} = \sum_{i=1}^n x_{i2} Y_i - n\overline{x}_2 \overline{Y}$,.....*and you get the pattern there.....*

Larger k will lead to an even more complicated looking expression!

Linear algebra to the rescue!!!

Multiple predictors linear regression ($k > 1$)

Preview of solution: Assume that $\mathbf{X}$ is a *full rank* matrix, then

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}.$$

Details will be unpacked shortly.

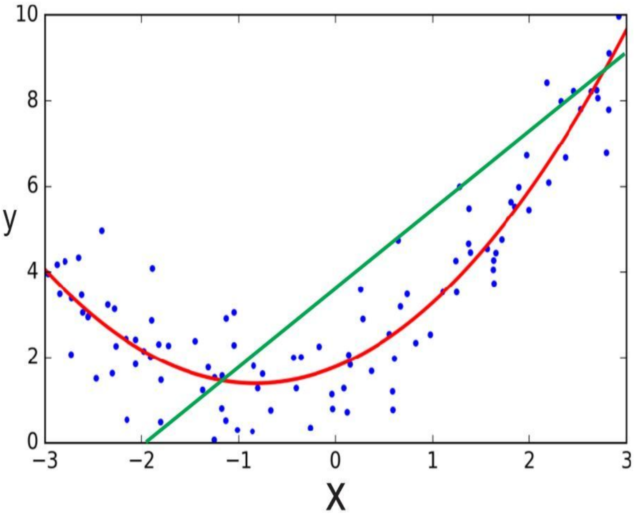
Polynomial regression

A model is linear when the response variable Y is predicted to be a linear form of the parameters β. Linearity in x is not needed.
For example, the model $Y = \beta_0 + \beta_1 x + \beta_2 x^2$ is a linear model. We just take different design variables!
The model

$$Y = \frac{\beta_1 x}{\beta_2 + x}$$

is NOT a linear model

Polynomial regression



Linear algebra

The matrix (interesting version)

The matrix (very interesting version)

A n by m matrix is a rectangular array of numbers of the form

$$\underbrace{\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}}_{\text{columns}} \left.\vphantom{\begin{bmatrix} x \\ x \\ \vdots \\ x \end{bmatrix}}\right\} \text{rows}$$

- If X and Y are matrices of **same size**, then $X + Y$ is the matrix whose (i, j) entry is $x_{ij} + y_{ij}$
- For any real number c, cX is the matrix whose (i, j)th element entry is cx_{ij}

Transpose

When transposing a matrix, columns become rows and rows become columns.

$$\mathbf{A} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

is a *n* by *m* matrix.

Transpose

When transposing a matrix, columns become rows and rows become columns.

$$\mathbf{A} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

is a n by m matrix.

$$\mathbf{A}^T = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{mn} \end{bmatrix}$$

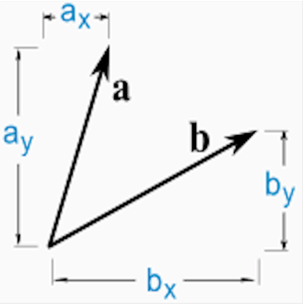
is a m by n matrix.

Transpose

- $(\mathbf{X}^T)^T = \mathbf{X}$.
- A matrix $\mathbf{X}$ is *symmetric* if and only if $\mathbf{X}^T = \mathbf{X}$.

Vectors and dot product

Matrices with only 1 row is called a row vector. Matrices with only 1 column is called a column vector.



$$\mathbf{a} = (a_x, a_y), \quad \mathbf{b} = (b_x, b_y)$$
$$\mathbf{a} \cdot \mathbf{b} = a_x \times b_x + a_y \times b_y.$$

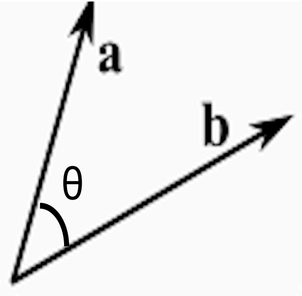
Same rule applies when **a** and **b** are column vectors.

Vector norms

The length of a vector $\mathbf{a}$ is called its *norm* and is denoted by $\|\mathbf{a}\|$. Let $\mathbf{a} = (a_1, \ldots, a_p)^T$. Then,

$$\|\mathbf{a}\| = \sqrt{\sum_{j=1}^{p} a_j^2}.$$

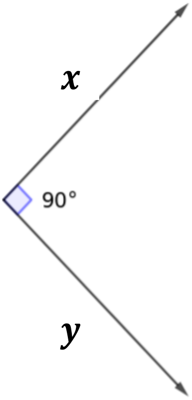
Let $\mathbf{a}$ and $\mathbf{b}$ be two vectors of the same size. Let θ denote the angle between the vectors. Then, θ and the vectors are related as such:



$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}.$$

Vector norms

Two vectors $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal* if and only if $\mathbf{x} \cdot \mathbf{y} = 0$.



$$\boldsymbol{x}.\boldsymbol{y} = \|\boldsymbol{x}\|\|\boldsymbol{y}\| \cos 90^{o} = 0$$

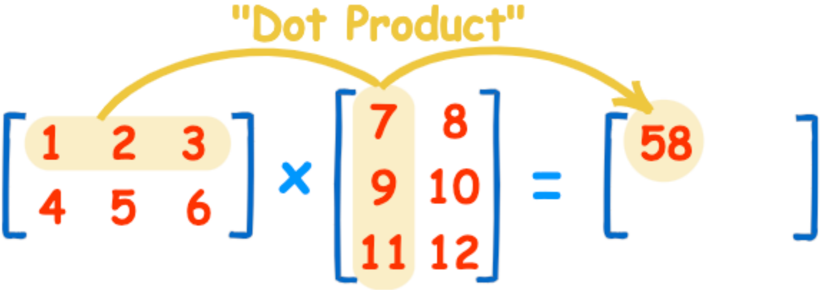
Matrix multiplication

To multiply two matrices, they must be *conformable*: The number of columns of the first matrix must be the same as number of rows of the second.

Let $\mathbf{X}$ be a $n \times k$ matrix and $\mathbf{Y}$ be a $k \times m$ matrix. The matrix $\mathbf{C} = \mathbf{XY}$ is a $n \times m$ matrix.

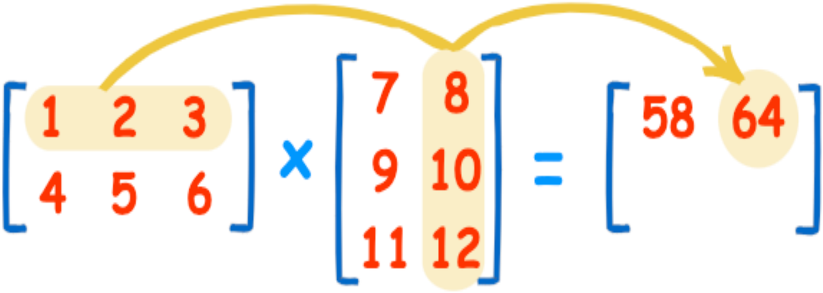
The $(i, j)$the element of C is the dot product of the ith row of $\mathbf{X}$ and the jth column of $\mathbf{Y}$.

Matrix multiplication



$$1 \times 7 + 2 \times 9 + 3 \times 11 = 58$$

Matrix multiplication



$$1 \times 7 + 2 \times 9 + 3 \times 11 = 58$$

Results of matrix multiplication

- $\mathbf{X}$ and $\mathbf{Y}$ are two n by n matrices. In general, $\mathbf{XY} \neq \mathbf{YX}$.
- $(\mathbf{XY})^T = \mathbf{Y}^T\mathbf{X}^T \neq \mathbf{X}^T\mathbf{Y}^T$.
- A matrix $\mathbf{X}$ is *symmetric* if and only if $\mathbf{X}^T = \mathbf{X}$.
- For two column vectors $\mathbf{a}$ and $\mathbf{b}$ of same size, $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T\mathbf{b} = \mathbf{b}^T\mathbf{a}$.
  For two row vectors $\mathbf{a}$ and $\mathbf{b}$ of same size, $\mathbf{a} \cdot \mathbf{b} = \mathbf{ab}^T = \mathbf{ba}^T$.

Linear independence

Suppose that we have a set of vectors $x_1, x_2, \dots, x_k$.

Linear independence

Suppose that we have a set of vectors $x_1, x_2, \ldots, x_k$.

We say that this set is *linearly dependent* if and only if there exists some numbers $a_1, a_2, \ldots, a_k$, which are not all zero, such that

$$a_1 x_1 + a_2 x_2 + \ldots + a_k x_k = 0.$$

Linear independence

Suppose that we have a set of vectors $x_1, x_2, \ldots, x_k$.

We say that this set is *linearly dependent* if and only if there exists some numbers $a_1, a_2, \ldots, a_k$, which are not all zero, such that

$$a_1 x_1 + a_2 x_2 + \ldots + a_k x_k = 0.$$

If the only way in which this equation is satisfied is for all a's to be zero, then we say that the x's are *linearly independent*.

Linear independence

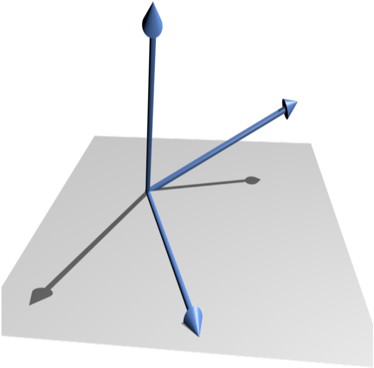
If a set of vectors is linearly dependent, then at least one of the vectors can be written as a linear combination of some or all of the rest.

Linear independence

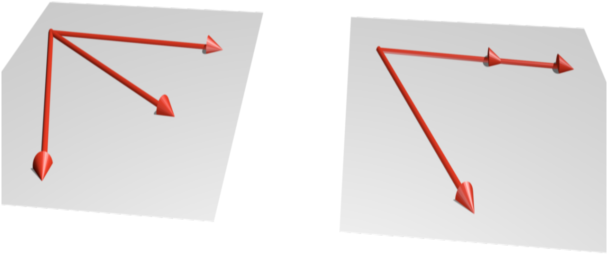
If a set of vectors is linearly dependent, then at least one of the vectors can be written as a linear combination of some or all of the rest.

In particular, a set of two vectors is linearly dependent if and only if one of the vectors is a constant multiple of the other.

Linear independence: independent vectors



Linear independence: dependent vectors



Determining linear independence

Example. Are the vectors

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

linearly independent?

Determining linear independence

Example. Are the vectors

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

linearly independent?

We substitute them into the equation:

$$a_1 x_1 + a_2 x_2 + a_3 x_3 = 0$$

Determining linear independence

Example. Are the vectors

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

linearly independent?

We substitute them into the equation:

$$\begin{aligned} a_1 x_1 + a_2 x_2 + a_3 x_3 = 0 \\ \Rightarrow \quad \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \end{aligned}$$

Determining linear independence

Example. Are the vectors

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

linearly independent?

We substitute them into the equation:

$$\begin{aligned} a_1 x_1 + a_2 x_2 + a_3 x_3 &= 0 \\ \Rightarrow \quad \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \end{aligned}$$

so they are linearly independent.

Determining linear independence

Example. How about the vectors?

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 2 \\ -1 \\ -2 \end{bmatrix}$$

Determining linear independence

Example. How about the vectors?

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 2 \\ -1 \\ -2 \end{bmatrix}$$

We substitute them into the equation:

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + a_3\mathbf{x}_3 = 0$$

Example. How about the vectors?

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 2 \\ -1 \\ -2 \end{bmatrix}$$

We substitute them into the equation:

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + a_3\mathbf{x}_3 = 0$$

$$\Rightarrow \quad \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, \text{ is a solution.}$$

Determining linear independence

Example. How about the vectors?

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 2 \\ -1 \\ -2 \end{bmatrix}$$

We substitute them into the equation:

$$\begin{aligned} & a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + a_3\mathbf{x}_3 = 0 \\ \Rightarrow \quad & \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, \text{ is a solution.} \end{aligned}$$

so $V = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is a set of linearly dependent vectors.

Determining linear independence

Example. Create linearly independent subsets of $V = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, where

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 2 \\ -1 \\ -2 \end{bmatrix}$$

Determining linear independence

Example. Create linearly independent subsets of $V = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, where

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 2 \\ -1 \\ -2 \end{bmatrix}$$

Many solutions:

$$\{\mathbf{x}_1\}, \{\mathbf{x}_2\}, \{\mathbf{x}_3\}, \quad \underbrace{\{\mathbf{x}_1, \mathbf{x}_2\}, \{\mathbf{x}_1, \mathbf{x}_3\}, \{\mathbf{x}_2, \mathbf{x}_3\}}_{\text{Solutions with largest number of vectors}}$$

Determining linear independence

Example. Create linearly independent subsets of $V = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, where

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 2 \\ -1 \\ -2 \end{bmatrix}$$

Many solutions:

$$\{\mathbf{x}_1\}, \{\mathbf{x}_2\}, \{\mathbf{x}_3\}, \quad \underbrace{\{\mathbf{x}_1, \mathbf{x}_2\}, \{\mathbf{x}_1, \mathbf{x}_3\}, \{\mathbf{x}_2, \mathbf{x}_3\}}_{\text{Solutions with largest number of vectors}}$$

Size of largest linearly independent subset $= 2$.

Rank

Consider the columns of the matrix:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_k \end{bmatrix}$$

where $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ are $n \times 1$ vectors.

Rank

Consider the columns of the matrix:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_k \end{bmatrix}$$

where $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ are $n \times 1$ vectors.

Put the columns into a set $V_{col} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$.

Rank

Consider the columns of the matrix:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_k \end{bmatrix}$$

where $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ are $n \times 1$ vectors.

Put the columns into a set $V_{col} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$.

The column rank of X, denoted by $r(X)$, is the largest linearly independent subset of V_{col}.

Rank

Consider the columns of the matrix:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_k \end{bmatrix}$$

where $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ are $n \times 1$ vectors.

Put the columns into a set $V_{col} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$.

The column rank of X, denoted by $r(X)$, is the largest linearly independent subset of V_{col}.

It is obvious that $r(X) \leq k$. If $n \geq k$ and $r(X) = k$, we say that X is of *full rank*.

Rank example

```
> A <- diag(3)
## [,1] [,2] [,3]
## [1,] 1 0 0
## [2,] 0 1 0
## [3,] 0 0 1

> library(Matrix)
> rankMatrix(A)[1]
## [1] 3
```

Rank example

```
> B <- matrix(c(1, 1, -1, -1, 2, 1, 2,
    -1, -2), 3, 3)
## [,1] [,2] [,3]
## [1,] 1 -1 2
## [2,] 1 2 -1
## [3,] -1 1 -2

> rankMatrix(B)[1]
## [1] 2
```

Rank propeties

- For any matrix X we have $r(X) = r(\mathbf{X}^T) = r(\mathbf{X}^T\mathbf{X})$.

Rank propeties

- For any matrix X we have $r(X) = r(\mathbf{X}^T) = r(\mathbf{X}^T\mathbf{X})$.
- Let $\mathbf{B}$ denote a diagonal p by p matrix, i.e.,

$$\mathbf{B} \begin{pmatrix} b_1 & 0 & \dots & 0 & 0 \\ 0 & b_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & b_p \end{pmatrix}$$

  Then, $r(\mathbf{B}) = p$.
- $r(XY) \leq \min\{r(X), r(Y)\}$.

Identity matrix

The *identity matrix* $\mathbf{I}$ is a square matrix of arbitrary size with 1's on the diagonal and 0's off the diagonal:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

Identity matrix

The *identity matrix* $\mathbf{I}$ is a square matrix of arbitrary size with 1's on the diagonal and 0's off the diagonal:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

It has the property that for any $\mathbf{X}$ of size $m \times n$,

$$\mathbf{X}\mathbf{I}_n = \mathbf{I}_{\mathbf{X}} = \mathbf{X},$$

where $\mathbf{I}_k$ is the $k \times k$ identity matrix. For a p by p identity matrix $\mathbf{I}_p$, we have $r(\mathbf{I}_p) = p$.

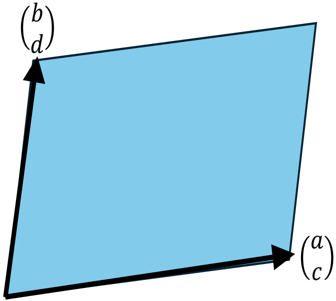
Consider a 2×2 matrix

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

. Its determinant is

$$\det(\mathbf{A}) = ad - bc$$

Sometimes, we express determinant using the notation $|\mathbf{A}|$.



Determinant equals area of parallelogram

Determinants

For a 3×3 matrix $\mathbf{A}$,

$$\mathbf{A} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$

Then

$$\det(\mathbf{A}) = a \left| \begin{pmatrix} e & f \\ h & i \end{pmatrix} \right| - b \left| \begin{pmatrix} d & f \\ g & i \end{pmatrix} \right| + c \left| \begin{pmatrix} d & e \\ g & h \end{pmatrix} \right|$$

$\det(\mathbf{A})$ equals volume of the enclosed parallelepiped.

Determinants

For a 3×3 matrix $\mathbf{A}$,

$$\mathbf{A} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$

Then

$$\det(\mathbf{A}) = a \left| \begin{pmatrix} e & f \\ h & i \end{pmatrix} \right| - b \left| \begin{pmatrix} d & f \\ g & i \end{pmatrix} \right| + c \left| \begin{pmatrix} d & e \\ g & h \end{pmatrix} \right|$$

$\det(\mathbf{A})$ equals volume of the enclosed parallelepiped.

In general, for a p by p square matrix $\mathbf{A}$, the determinant of $\mathbf{A}$ equals enclosed volume. Hand-computation of determinants for square matrice with $p > 3$ is beyond the scope of the course.

Inverses

If $\mathbf{X}$ is a square matrix, its *inverse* is the matrix $\mathbf{X}^{-1}$ of the same size which satisfies

$$\mathbf{X}\mathbf{X}^{-1} = \mathbf{X}^{-1}\mathbf{X} = \mathbf{I}.$$

Inverses

If $\mathbf{X}$ is a square matrix, its *inverse* is the matrix $\mathbf{X}^{-1}$ of the same size which satisfies

$$\mathbf{X}\mathbf{X}^{-1} = \mathbf{X}^{-1}\mathbf{X} = \mathbf{I}.$$

Condition for existence of inverse: $\mathbf{X}^{-1}$ exists if and only if $\mathbf{X}$ is a square matrix with full rank

Inverses

If $\mathbf{X}$ is a square matrix, its *inverse* is the matrix $\mathbf{X}^{-1}$ of the same size which satisfies

$$\mathbf{X}\mathbf{X}^{-1} = \mathbf{X}^{-1}\mathbf{X} = \mathbf{I}.$$

Condition for existence of inverse: $\mathbf{X}^{-1}$ exists if and only if $\mathbf{X}$ is a square matrix with full rank

If $\mathbf{X}^{-1}$ exists, then we say that $\mathbf{X}$ is nonsingular. Otherwise, we say that $\mathbf{X}$ is singular.

Inverse of a 2 by 2 matrix

Consider

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

The inverse is

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Inverse of a 2 by 2 matrix

Consider

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

The inverse is

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Hand-computation of inverses for square matrice with size $p \geq 3$ is beyond the scope of the course.

Inverse properties

$\mathbf{X}$ is nonsingular if and only if $r(\mathbf{X}) = k$

Inverse properties

$\mathbf{X}$ is nonsingular if and only if $r(\mathbf{X}) = k$

Consider three nonsingular matrices: $\mathbf{X}$ is $n \times k$, $\mathbf{P}$ are $n \times n$, and $\mathbf{Q}$ is $k \times k$. Then, $r(\mathbf{X}) = r(\mathbf{PX}) = r(\mathbf{XQ})$.

Inverse properties

$\mathbf{X}$ is nonsingular if and only if $r(\mathbf{X}) = k$

Consider three nonsingular matrices: $\mathbf{X}$ is $n \times k$, $\mathbf{P}$ are $n \times n$, and $\mathbf{Q}$ is $k \times k$. Then, $r(\mathbf{X}) = r(\mathbf{PX}) = r(\mathbf{XQ})$.

If $\mathbf{A}$ and $\mathbf{B}$ are nonsingular square matrices, then

- $\mathbf{A}^{-1}$ is nonsingular and $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$. Also, $\det(\mathbf{A}^{-1}) = 1/\det(\mathbf{A})$.
- $\mathbf{AB}$ is nonsingular and $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
- $\mathbf{A}^T$ is nonsingular and $(\mathbf{X}^T)^{-1} = (\mathbf{X}^T)^{-1}$. For notation convenience, we denote $\mathbf{X}^{-T} = (\mathbf{X}^T)^{-1}$.

Partitioned matrices

Matrices can be *partitioned* into smaller (rectangular) *submatrices*:

$$X = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 3 & -1 \\ 0 & 1 & -1 & 1 \\ 2 & -1 & 0 & 2 \end{bmatrix}$$

Partitioned matrices

Matrices can be *partitioned* into smaller (rectangular) *submatrices*:

$$X = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 3 & -1 \\ 0 & 1 & -1 & 1 \\ 2 & -1 & 0 & 2 \end{bmatrix}$$
$$= \left[\begin{array}{cc|c|c} 1 & 0 & 1 & 0 \\ 0 & 1 & 3 & -1 \\ \hline 0 & 1 & -1 & 1 \\ 2 & -1 & 0 & 2 \end{array}\right].$$

Partitioning

Partitioned matrices can be manipulated as if the submatrices were single elements (using matrix multiplication instead of scalar multiplication). However, the dimensions of the submatrices must be compatible!

Partitioning

Partitioned matrices can be manipulated as if the submatrices were single elements (using matrix multiplication instead of scalar multiplication). However, the dimensions of the submatrices must be compatible!

For example, let

$$X = \left[\begin{array}{cc|c} 2 & 1 & 0 \\ \hline 3 & 4 & 1 \end{array}\right] = \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right]$$

Partitioning

Partitioned matrices can be manipulated as if the submatrices were single elements (using matrix multiplication instead of scalar multiplication). However, the dimensions of the submatrices must be compatible!

For example, let

$$X = \left[\begin{array}{cc|c} 2 & 1 & 0 \\ \hline 3 & 4 & 1 \end{array}\right] = \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right]$$

$$Y = \left[\begin{array}{cc} 1 & 0 \\ 2 & 4 \\ \hline 3 & -1 \end{array}\right] = \left[\begin{array}{c} Y_{11} \\ \hline Y_{21} \end{array}\right]$$

Partitioning

Then

$$XY = \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right] \left[\begin{array}{c} Y_{11} \\ \hline Y_{21} \end{array}\right]$$

Partitioning

Then

$$XY = \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right] \left[\begin{array}{c} Y_{11} \\ \hline Y_{21} \end{array}\right]$$

$$= \left[\begin{array}{c} X_{11}Y_{11} + X_{12}Y_{21} \\ \hline X_{21}Y_{11} + X_{22}Y_{21} \end{array}\right]$$

Partitioning

Then

$$
\begin{aligned}
XY &= \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right] \left[\begin{array}{c} Y_{11} \\ \hline Y_{21} \end{array}\right] \\
&= \left[\begin{array}{c} X_{11}Y_{11} + X_{12}Y_{21} \\ \hline X_{21}Y_{11} + X_{22}Y_{21} \end{array}\right] \\
&= \left[\begin{array}{c} \begin{bmatrix} 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2 & 4 \end{bmatrix} + [0] \begin{bmatrix} 3 & -1 \end{bmatrix} \\ \hline \begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2 & 4 \end{bmatrix} + [1] \begin{bmatrix} 3 & -1 \end{bmatrix} \end{array}\right]
\end{aligned}
$$

using matrix multiplication for the submatrices.

Partitioning

$$X = \left[\begin{array}{cc|c} 2 & 1 & 0 \\ \hline 3 & 4 & 1 \end{array}\right] = \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right]$$

However, if we partition Y into

$$Y = \left[\begin{array}{cc} 1 & 0 \\ \hline 2 & 4 \\ 3 & -1 \end{array}\right] = \left[\begin{array}{c} Y_{11} \\ \hline Y_{21} \end{array}\right]$$

then we cannot do the multiplication through the partitioning because the components do not have compatible dimensions (for example, X_{11}, Y_{11} are not compatible because X_{11} is 1×2 and Y_{11} is also 1×2) !

Partitioning

Consider

$$\mathbf{X} = \left[\begin{array}{c|c} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \hline \mathbf{X}_{21} & \mathbf{X}_{22} \end{array}\right]$$

Partitioning

Consider

$$\mathbf{X} = \left[\begin{array}{c|c} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \hline \mathbf{X}_{21} & \mathbf{X}_{22} \end{array}\right]$$

Then

$$\mathbf{X}^T = \left[\begin{array}{c|c} \mathbf{X}_{11}^T & \mathbf{X}_{12}^T \\ \hline \mathbf{X}_{21}^T & \mathbf{X}_{22}^T \end{array}\right]$$

Partitioning

Consider

$$\mathbf{X} = \left[\begin{array}{c|c} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \hline \mathbf{X}_{21} & \mathbf{X}_{22} \end{array}\right]$$

Then

$$\mathbf{X}^T = \left[\begin{array}{c|c} \mathbf{X}_{11}^T & \mathbf{X}_{12}^T \\ \hline \mathbf{X}_{21}^T & \mathbf{X}_{22}^T \end{array}\right]$$

and

$$\mathbf{X}^{-1} = \left[\begin{array}{c|c} \widetilde{\mathbf{X}}_{11} & \widetilde{\mathbf{X}}_{21} \\ \hline \widetilde{\mathbf{X}}_{12} & \widetilde{\mathbf{X}}_{22} \end{array}\right]$$

$$\begin{aligned}
\widetilde{\mathbf{X}}_{11} &= \mathbf{X}_{11}^{-1} + \mathbf{X}_{11}^{-1}\mathbf{X}_{12}(\mathbf{X}_{22} - \mathbf{X}_{21}\mathbf{X}_{11}^{-1}\mathbf{X}_{12})^{-1}\mathbf{X}_{21}\mathbf{X}_{11}^{-1} \\
\widetilde{\mathbf{X}}_{12} &= -\mathbf{X}_{11}^{-1}\mathbf{X}_{12}(\mathbf{X}_{22} - \mathbf{X}_{21}\mathbf{X}_{11}^{-1}\mathbf{X}_{12})^{-1} \\
\widetilde{\mathbf{X}}_{21} &= -(\mathbf{X}_{22} - \mathbf{X}_{21}\mathbf{X}_{11}^{-1}\mathbf{X}_{12})^{-1}\mathbf{X}_{21}\mathbf{X}_{11}^{-1} \\
\widetilde{\mathbf{X}}_{22} &= (\mathbf{X}_{22} - \mathbf{X}_{21}\mathbf{X}_{11}^{-1}\mathbf{X}_{12})^{-1}
\end{aligned}$$

Back to linear models: least squares estimation

We can express the general linear model in matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Back to linear models: least squares estimation

We can express the general linear model in matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

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

Back to linear models: least squares estimation

The least squares estimate

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \|\mathbf{Y} - \mathbf{X}\beta\|^2,$$

where $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_k)^T$. To minimise the objective $\|\mathbf{Y} - \mathbf{X}\beta\|^2$, we need tools from matrix calculus!