MAST90104 - Lecture 2

Weichang Yu

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

29 Jul, 2024

Linear models: least squares estimation

The least squares regression estimate is the minimiser

$$(\widehat{\beta}_0, \ldots, \widehat{\beta}_k)^T = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{k+1}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2.$$

Our newly acquired linear algebra knowledge allows us to express our RHS as

$$\begin{aligned}(\widehat{\beta}_0, \ldots, \widehat{\beta}_k)^T &= \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{k+1}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{k+1}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2\end{aligned}$$

To calculate the minimiser of $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$, we need tools from matrix calculus!

Suppose $\mathbf{x}$ and $\mathbf{m}$ are m-dimensional column vectors. Then

$$\frac{\partial \mathbf{m}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{m}^T$$

Suppose $\mathbf{x}$ is a m-dimensional column vector and $\mathbf{A}$ is a n by m matrix. Denote $\mathbf{a}_{j\cdot}$ as the j-th row of $\mathbf{A}$. Then,

$$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}}\begin{pmatrix} \mathbf{a}_{1\cdot}\mathbf{x} \\ \mathbf{a}_{2\cdot}\mathbf{x} \\ \vdots \\ \mathbf{a}_{n\cdot}\mathbf{x} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial \mathbf{x}}\mathbf{a}_{1\cdot}\mathbf{x} \\ \frac{\partial}{\partial \mathbf{x}}\mathbf{a}_{2\cdot}\mathbf{x} \\ \vdots \\ \frac{\partial}{\partial \mathbf{x}}\mathbf{a}_{n\cdot}\mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{a}_{1\cdot} \\ \mathbf{a}_{2\cdot} \\ \vdots \\ \mathbf{a}_{n\cdot} \end{pmatrix} = \mathbf{A}$$

Also $\frac{\partial \mathbf{x}^T\mathbf{A}^T}{\partial \mathbf{x}^T} = \mathbf{A}^T$

Suppose $\mathbf{x}$ is a m-dimensional column vector and $\mathbf{B}$ is a square matrix of order m. Then,

$$\frac{\partial \mathbf{x}^T \mathbf{B} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T (\mathbf{B} + \mathbf{B}^T)$$

Linear models: computation of least squares estimate

$$(\widehat{\beta}_0, \ldots, \widehat{\beta}_k)^T = \underset{\boldsymbol{\beta} \in \mathbb{R}^{k+1}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

Now,

$$\begin{aligned}
\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \left(\mathbf{Y}^T - \boldsymbol{\beta}^T\mathbf{X}^T\right)(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} - \underbrace{\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y}}_{=\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta}} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}
\end{aligned}$$

Linear models: computation of least squares estimate

Then,

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2 &= \tfrac{\partial \mathbf{Y}^T\mathbf{Y}}{\partial \boldsymbol{\beta}} - 2\tfrac{\partial \mathbf{Y}^T\mathbf{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} + \tfrac{\partial \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} \\ &= \mathbf{0} - 2\mathbf{Y}^T\mathbf{X} + \boldsymbol{\beta}^T\{\mathbf{X}^T\mathbf{X} + (\mathbf{X}^T\mathbf{X})^T\} \\ &= \mathbf{0} - 2\mathbf{Y}^T\mathbf{X} + 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\end{aligned}$$

Then, setting LHS equals to $\mathbf{0}$, $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$, and applying transpose, we have the *normal equations*

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$$

If $\mathbf{X}$ is full rank, then

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

Linear models: prerequisite knowledge for inference theory

Recap of one-sample test of mean: $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, then we can test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. Using your prior knowledge from MAST90105, our test statistic's null distribution is

$$T = \frac{\overline{X}}{S/\sqrt{n}} \sim t_{n-1}.$$

Similarly, in linear regression, we need to test $H_0 : \beta_1 = \beta_2 = 0$ against $H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$. We derive a test statistic and its null distribution.

We need more knowledge about linear algebra to derive the test statistic and null distribution!

Orthonormal vectors

A set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is called an orthogonal set if every pair of vectors are orthogonal, that is, $\mathbf{x}_j \cdot \mathbf{x}_{j'} = 0$ for all $j \neq j'$.

If $V = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is a set of nonzero orthogonal vectors, then V is a linearly independent set. The converse is not always true.

An orthogonal set is called an orthonormal set if $\|\mathbf{x}_j\| = 1$ for every $j = 1, \ldots, k$.

Orthonormal vectors

```
> ( x <- c(1,2,3)/sqrt(14) )
[1] 1 2 3
> ( y <- c(1,1,-1)/sqrt(3) )
> x%*%y
[1,] 0
> t(x)%*%y
[1,] 0
> sqrt(sum(x^2))
[1,] 1
> sqrt(sum(y^2))
[1,] 1
```

Orthogonal matrices

A square matrix $\mathbf{X}$ is *orthogonal* if and only if

$$\mathbf{X}^T\mathbf{X} = I.$$

If $\mathbf{X}$ is orthogonal, then

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

Orthogonal matrices

```
> X <- matrix(c(c(1,2,3)/sqrt(14),c(1,1,-1)/sqrt(3),
+ c(5,-4,1)/sqrt(42)),3,3)
> X
[,1] [,2] [,3]
[1,] 0.2672612 0.5773503 0.7715167
[2,] 0.5345225 0.5773503 -0.6172134
[3,] 0.8017837 -0.5773503 0.1543033
> round(t(X)%*%X,5)
[,1] [,2] [,3]
[1,] 1 0 0
[2,] 0 1 0
[3,] 0 0 1
```

Orthogonal matrices

X is an orthogonal matrix if and only if the columns (or rows) of **X** form an orthonormal set.

```
> X[,1]%*%X[,2]
[1,] 0
> X[1,]%*%X[3,]
[1,] -8.326673e-17
> sqrt(sum(X[,1]^2))
[1] 1
```

Eigenvalues and eigenvectors

Suppose $\mathbf{A}$ is a $n \times n$ matrix and $\mathbf{x}$ is a $n \times 1$ **nonzero** vector which satisfies the equation

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

where $\boldsymbol{\lambda}$ is a scalar. Then we say that λ is an *eigenvalue* of $\mathbf{A}$, with associated *eigenvector* $\mathbf{x}$.

Eigenvalues and eigenvectors

Suppose $\mathbf{A}$ is a $n \times n$ matrix and $\mathbf{x}$ is a $n \times 1$ **nonzero** vector which satisfies the equation

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

where $\boldsymbol{\lambda}$ is a scalar. Then we say that λ is an *eigenvalue* of $\mathbf{A}$, with associated *eigenvector* $\mathbf{x}$.

Rearranging the definition, we get

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}.$$

Now if $\mathbf{A} - \lambda I$ is invertible, this produces

$$\mathbf{x} = (\mathbf{A} - \lambda I)^{-1}\mathbf{0} = \mathbf{0}.$$

But $\mathbf{x}$ is nonzero by definition, so $\mathbf{A} - \lambda I$ must be singular. In particular, its determinant must be 0. Therefore we can find the eigenvalues of a matrix by solving the *characteristic equation* (this is a polynomial in λ)

$$|\mathbf{A} - \lambda I| = \mathbf{0}.$$

Eigenvalues and eigenvectors

Let

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ -2 & 4 \end{bmatrix}.$$

To find the eigenvalues of $\mathbf{A}$, we solve the equation

$$\begin{vmatrix} 1-\lambda & 1 \\ -2 & 4-\lambda \end{vmatrix} = (1-\lambda)(4-\lambda) - (-2) = 0.$$

This becomes

$$\lambda^2 - 5\lambda + 6 = (\lambda - 2)(\lambda - 3) = 0.$$

Therefore $\mathbf{A}$ has two eigenvalues, 2 and 3.

Eigenvalues and eigenvectors

To find the eigenvector(s) of $\mathbf{A}$ associated with eigenvalue 2, we solve the system of equations

$$\mathbf{Ax} = \begin{bmatrix} 1 & 1 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2\mathbf{x} = 2 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

This is a linear system which has two equations and two unknowns; however, the equations are redundant. Therefore the system has an infinite number of solutions, which always happens for an eigenvector system. One solution is

$$\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Eigenvalue example

```
> A
[,1] [,2] [,3]
[1,] 1 2 0
[2,] 2 3 -1
[3,] 0 -1 8
> e <- eigen(A)
> e$values
[1] 8.2145852 4.0555651 -0.2701503
> e$vectors
# the columns contain the eigenvectors of A,
# normalized to have unit length
[,1] [,2] [,3]
[1,] -0.05806435 0.5357376 0.84238574
[2,] -0.20945510 0.8184906 -0.53497826
[3,] 0.97609277 0.2075052 -0.06468785
```

Eigenvalue example

```
> det(A - e$values[1]*I)
[1] -2.799516e-14
> A %*% e$vectors[,1]
[,1]
[1,] -0.4769745
[2,] -1.7205868
[3,] 8.0181972
> e$values[1]*e$vectors[,1]
[1] -0.4769745 -1.7205868 8.0181972
```

Eigenvalue properties

- If $\mathbf{A}$ is (real and) symmetric, then its eigenvalues are all real, and its eigenvectors associated with distinct eigenvalues are orthogonal.
- If $\mathbf{P}$ is an orthogonal matrix of the same size as A, then the eigenvalues of $\mathbf{P}^T\mathbf{AP}$ are the same as the eigenvalues of $\mathbf{A}$.
- The eigenvalues of a diagonal matrix $\mathbf{S}$ are the elements on the diagonal.
- The determinant of a matrix is the product of its eigenvalues

Eigenvalue properties

If a square matrix $\mathbf{A}$ is singular, its determinant is 0

- At least one eigenvalue of $\mathbf{A}$ is 0

The eigenvectors of the data's covariance matrix are also called the "principal components".
Can be used to find collinear combinations of the predictor variables.

Diagonalization: The Spectral Theorem

Theorem 2.1

Let $\mathbf{A}$ be a $k \times k$ matrix. Then an orthogonal matrix $\mathbf{P}$ exists such that

$$\mathbf{P}^T\mathbf{A}\mathbf{P} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_k \end{bmatrix},$$

where $\lambda_i, i = 1, 2, \dots, k$, are the eigenvalues of $\mathbf{A}$ if and only if $\mathbf{A}$ is symmetric

The proof is beyond the scope of this course.
CONTEST: Free \$10 Coles voucher to the first person who can write and explain the proof to my satisfaction during office hours.

Diagonalization

If $\mathbf{P}$ is an orthogonal matrix such that

$$\mathbf{P}^T\mathbf{AP} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_k \end{bmatrix},$$

then we say that P *diagonalises* A.
It can be shown that the columns of $\mathbf{P}$ must be eigenvectors of A associated with the respective eigenvalues.
Hence (from above) columns of $\mathbf{P}$ form an orthonormal set because $\mathbf{P}$ is an orthogonal matrix.
Note that

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_k \end{bmatrix},$$

is a diagonal matrix.

Exercise: Suppose $\mathbf{A}$ is a p by p matrix that is diagonalisable by a matrix $\mathbf{P}$ and corresponding diagonal matrix $\boldsymbol{\Lambda}$, i.e., $\mathbf{P}^T\mathbf{A}\mathbf{P} = \boldsymbol{\Lambda}$. Use the fact that the determinant of a diagonal matrix equals to the product of the diagonal entries to show that

$$\det(\mathbf{A}) = \prod_{j=1}^{p} \lambda_j,$$

where $\{\lambda_1, \ldots, \lambda_p\}$ are the diagonal entries of $\boldsymbol{\Lambda}$.

Diagonalization example

```
> A
[,1] [,2] [,3]
[1,] 1 2 0
[2,] 2 3 -1
[3,] 0 -1 8
> e$values
[1] 8.2145852 4.0555651 -0.2701503
> P <- e$vectors
> round(t(P)%*%A%*%P,5)
[,1] [,2] [,3]
[1,] 8.21459 0.00000 0.00000
[2,] 0.00000 4.05557 0.00000
[3,] 0.00000 0.00000 -0.27015
```

Idempotence

We say that a square matrix $\mathbf{A}$ is *idempotent* if and only if

$$\mathbf{A}^2 = \mathbf{A}.$$

Example. The identity matrix $\mathbf{I}$ is idempotent.

Exercise. Let $\mathbf{X}$ be an $n \times k$ matrix of full rank, $n \geq k$. Show that

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

exists and is idempotent.

Trace

The *trace* of a square $k \times k$ matrix X, denoted by $tr(X)$, is the sum of its diagonal entries:

$$tr(X) = \sum_{i=1}^{k} x_{ii}.$$

Example.

$$tr\left(\begin{bmatrix} 2 & 0 & 1 \\ 1 & 1 & 0 \\ 3 & 2 & -1 \end{bmatrix}\right) = 2 + 1 - 1 = 2.$$

Trace properties

- If c is a scalar, $tr(cX) = c\ tr(X)$.
- $tr(X \pm Y) = tr(X) \pm tr(Y)$.
- If XY and YX both exist, $tr(XY) = tr(YX)$.

Trace properties

Example. Let

$$X = \begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 3 & 2 \end{bmatrix}, Y = \begin{bmatrix} -1 & 1 & 0 \\ 2 & 4 & 0 \end{bmatrix}.$$

Then

$$tr(XY) = tr\left(\begin{bmatrix} -2 & 2 & 0 \\ 1 & 5 & 0 \\ 1 & 11 & 0 \end{bmatrix}\right) = 3$$

$$tr(YX) = tr\left(\begin{bmatrix} -1 & 1 \\ 8 & 4 \end{bmatrix}\right) = 3$$

so even though $XY \neq YX$, their traces are equal.

Some linear algebra theorems

Theorem 2.2

The eigenvalues of idempotent matrices are always either 0 or 1.

Proof. Let $\mathbf{A}$ be an idempotent matrix with eigenvalue λ and

associated eigenvector $\mathbf{x}$. By definition,

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}.$$

Multiplying by $\mathbf{A}$,

$$\mathbf{A}^2\mathbf{x} = \mathbf{A}\lambda\mathbf{x} = \lambda\mathbf{A}\mathbf{x} = \lambda^2\mathbf{x}.$$

But $\mathbf{A}$ is idempotent, so

$$\begin{aligned} \lambda^2\mathbf{x} = \mathbf{A}^2\mathbf{x} = \mathbf{A}\mathbf{x} &= \lambda\mathbf{x} \\ (\lambda^2 - \lambda)\mathbf{x} &= \mathbf{0}. \end{aligned}$$

By definition, $\mathbf{x} \neq \mathbf{0}$, so $\lambda = \lambda^2$. Therefore $\lambda = 0$ or 1.

Some linear algebra theorems

Theorem 2.3

If A is a symmetric and idempotent matrix, $r(A) = tr(A)$.

Proof. We take A to be $k \times k$. First we diagonalize A, i.e. find P such that

$$P^T AP = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_k \end{bmatrix},$$

where $\lambda_1, \lambda_2, \dots, \lambda_k$ are the eigenvalues of A.
Since P is orthogonal, both P and P^T are nonsingular. Therefore, using results from Tutorial sheet 1 Q5,

$$r(P^T AP) = r(P^T A) \underbrace{=}_{\text{Sylvester's Ineq.}} r(A).$$

Because $P^T AP$ is diagonal, $r(P^T AP)$ is the number of nonzero eigenvalues of A.

Some linear algebra theorems

Theorem 2.4

If $\mathbf{A}$ *is a symmetric and idempotent matrix,* $r(\mathbf{A}) = tr(\mathbf{A})$.

Proof. But $\mathbf{A}$ is idempotent, so its eigenvalues are either 0 or 1.

To count the number of nonzero eigenvalues, we just need to sum them. But since they are the diagonal elements of $\mathbf{P}^T\mathbf{A}\mathbf{P}$, we can just take its trace.

Therefore

$$r(\mathbf{A}) = r(\mathbf{P}^T\mathbf{A}\mathbf{P}) = tr(\mathbf{P}^T\mathbf{A}\mathbf{P}) = tr(\mathbf{P}\mathbf{P}^T\mathbf{A}) = tr(\mathbf{A})$$

since P is orthogonal.

Some linear algebra theorems

Theorem 2.5

Let $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m$ be a collection of symmetric $k \times k$ matrices. Then the following are equivalent:

- *There exists an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T \mathbf{A}_i \mathbf{P}$ is diagonal for all $i = 1, 2, \ldots, m$;*
- *$\mathbf{A}_i \mathbf{A}_j = \mathbf{A}_j \mathbf{A}_i$ for every pair $i, j = 1, 2, \ldots, m$.*

Some linear algebra theorems

Theorem 2.6

Let $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m$ be a collection of symmetric $k \times k$ matrices. Then any two of the following conditions implies the third:

- *All $\mathbf{A}_i$, $i = 1, 2, \ldots, m$ are idempotent;*
- *$\sum_{i=1}^{m} \mathbf{A}_i$ is idempotent;*
- *$\mathbf{A}_i \mathbf{A}_j = 0$ for $i \neq j$.*

Some linear algebra theorems

Theorem 2.7

Let $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m$ be a collection of symmetric $k \times k$ matrices. If the conditions in Theorem 2.6 are true, then

$$r\left(\sum_{i=1}^{m} \mathbf{A}_i\right) = \sum_{i=1}^{m} r\left(\mathbf{A}_i\right).$$

Proof.

Consider $\sum_{i=1}^{m} \mathbf{A}_i$. By assumption, this matrix is idempotent. As a sum of symmetric matrices it is also symmetric.
Thus by Theorem 2.4,

$$\begin{aligned} r\left(\sum_{i=1}^{m} \mathbf{A}_i\right) &= tr\left(\sum_{i=1}^{m} \mathbf{A}_i\right) \\ &= \sum_{i=1}^{m} tr(\mathbf{A}_i) = \sum_{i=1}^{m} r(\mathbf{A}_i). \end{aligned}$$

Theorem examples

```
> X <- matrix(c(1/2,1/2,0,1/2,1/2,0,0,0,1),3,3)
> X %*% X
[,1] [,2] [,3]
[1,] 0.5 0.5 0
[2,] 0.5 0.5 0
[3,] 0.0 0.0 1
> sum(diag(X))
[1] 2
```

Theorem examples

```
> eigen(X)$values
[1] 1.000000e+00 1.000000e+00 5.551115e-16
> rankMatrix(X)[1]
[1] 2
```

Theorem examples

```
> A1 <- matrix(c(1/2,-1/2,-1/2,1/2),2,2)
> A1 %*% A1
[,1] [,2]
[1,] 0.5 -0.5
[2,] -0.5 0.5
> A2 <- matrix(c(1/2,1/2,1/2,1/2),2,2)
> A2 %*% A2
[,1] [,2]
[1,] 0.5 0.5
[2,] 0.5 0.5
```

Theorem examples

```
> A1 + A2
[,1] [,2]
[1,] 1 0
[2,] 0 1
> (A1 + A2) %*% (A1 + A2)
[,1] [,2]
[1,] 1 0
[2,] 0 1
> A1 %*% A2
[,1] [,2]
[1,] 0 0
[2,] 0 0
```

Theorem examples

```
> A2 %*% A1
[,1] [,2]
[1,] 0 0
[2,] 0 0
> rankMatrix(A1 + A2)[1]
[1] 2
> rankMatrix(A1)[1] + rankMatrix(A2)[1]
[1] 2
```

Quadratic forms

Previously, we saw that the least squares solution $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_k)^T$ minimises

$$\|\mathbf{Y} - \mathbf{X}\beta\|^2 = \underbrace{\mathbf{Y}^T\mathbf{Y}}_{\text{quadratic form in } \mathbf{Y}} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \underbrace{\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}}_{\text{quadratic form in } \boldsymbol{\beta}}$$

In general, for a p by p matrix $\mathbf{A}$ and a p-dimensional column vector $\mathbf{x}$, the quantity

$$\mathbf{x}^T\mathbf{A}\mathbf{x}$$

is called a *quadratic form* in $\mathbf{x}$, and $\mathbf{A}$ is the matrix multiplier of the quadratic form. Note that a quadratic form is a scalar.

Quadratic forms

Note that

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}x_ix_j,$$

where a_{ij} is the (i,j)-th entry of $\mathbf{A}$ and x_i is the i-th entry of the vector $\mathbf{x}$.

- $\mathbf{Y}^T\mathbf{Y}$ is a quadratic form in $\mathbf{Y}$ with matrix multiplier $\mathbf{I}$.
- $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$ is a quadratic form in $\boldsymbol{\beta}$ with matrix multiplier $\mathbf{X}^T\mathbf{X}$.

Quadratic forms

Example. Let

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad A = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 0 \\ 4 & 6 & 3 \end{bmatrix}.$$

Then

$$\begin{aligned} \mathbf{x}^T A\mathbf{x} &= 2x_1^2 + 3x_1x_2 + x_1x_3 + x_2x_1 + 2x_2^2 + 4x_3x_1 + 6x_3x_2 + 3x_3^2 \\ &= 2x_1^2 + 2x_2^2 + 3x_3^2 + 4x_1x_2 + 5x_1x_3 + 6x_2x_3. \end{aligned}$$

This can be found from either the summation formula or multiplying out the matrices.

Positive definite matrices

If $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$, then we say that the quadratic form $\mathbf{x}^T A\mathbf{x}$ is *positive definite*; we also say that the matrix $\mathbf{A}$ is positive definite.

If $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x}$, then we say that the quadratic form $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is *positive semi-definite*; we also say that the matrix $\mathbf{A}$ is positive semi-definite.

Positive definite matrices

Example. Let

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$

Then

$$\mathbf{x}^T A\mathbf{x} = 2x_1^2 + 2x_2^2 - 2x_1x_2 = x_1^2 + x_2^2 + (x_1 - x_2)^2.$$

The quadratic form will never be negative, and the onlx wax that it can be 0 is if all the squares are 0, i.e. $x_1 = x_2 = 0$. Therefore, $\mathbf{x}^T A\mathbf{x}$ is positive definite.

Positive definiteness theorems

Theorem 2.8

A symmetric matrix $\mathbf{A}$ *is positive definite if and only if its eigenvalues are all (strictly) positive.*

Theorem 2.9

A symmetric matrix $\mathbf{A}$ *is positive semi-definite if and only if its eigenvalues are all non-negative.*

Positive definite matrices

Example. Consider the matrix in the previous example:

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$

The eigenvalues of $\mathbf{A}$ solve the quadratic equation

$$(2-\lambda)^2 - 1 = \lambda^2 - 4\lambda + 3 = (\lambda - 3)(\lambda - 1) = 0.$$

Therefore its eigenvalues are 1 and 3, which are both positive, and it is positive definite.