ABE 6933 SML, Fall 2020
A Matrix Algebra Approach to Linear Regression

Instructor: Dr. Nikolay Bliznyuk

29 September, 2020

`Slides under "[4].lectures\probability.and.stats.slides"`

Matrix as a Rectangular Array

A matrix with r rows and c columns is a rectangular array. It will be represented either in full form

$$\underset{r \times c}{\mathbf{A}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1c} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2c} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{ic} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rj} & \cdots & a_{rc} \end{bmatrix},$$

or in abbreviated form

$$\mathbf{A} = [a_{ij}] \quad i = 1, \ldots, r; \; j = 1, \ldots, c;$$

or simply by a boldface symbol, such as $\mathbf{A}$.

Transpose of a Matrix: an Illustration

The transpose of a matrix $\mathbf{A}$ is another matrix, denoted by changing corresponding columns and rows of the matrix A. For example, if

$$\underset{3\times 2}{\mathbf{A}} = \begin{bmatrix} 2 & 5 \\ 7 & 10 \\ 3 & 4 \end{bmatrix},$$

then the transpose of $\mathbf{A}$ is

$$\underset{2\times 3}{\mathbf{A}'} = \begin{bmatrix} 2 & 7 & 3 \\ 5 & 10 & 4 \end{bmatrix}.$$

Transpose of a Matrix

$$\underset{r\times c}{\mathbf{A}} = \begin{bmatrix} a_{11} & \cdots & a_{1c} \\ \vdots & & \vdots \\ a_{r1} & \cdots & a_{rc} \end{bmatrix} = [a_{ij}], \quad i = 1, \ldots, r; \quad j = 1, \ldots, c.$$

$$\underset{c\times r}{\mathbf{A}'} = \begin{bmatrix} a_{11} & \cdots & a_{r1} \\ \vdots & & \vdots \\ a_{1c} & \cdots & a_{rc} \end{bmatrix} = [a_{ji}], \quad j = 1, \ldots, c; \quad i = 1, \ldots, r.$$

In `R`, the call to compute $\mathbf{A}'$ is `t(A)`.

Addition and Subtraction of Matrices

In general, if

$$\underset{r \times c}{\mathbf{A}} = [a_{ij}] \quad \text{and} \quad \underset{r \times c}{\mathbf{B}} = [b_{ij}], \quad i = 1, \ldots, r; \quad j = 1, \ldots c,$$

then

$$\underset{r \times c}{\mathbf{A} + \mathbf{B}} = [a_{ij} + b_{ij}] \quad \text{and} \quad \underset{r \times c}{\mathbf{A} - \mathbf{B}} = [a_{ij} - b_{ij}].$$

In `R`, standard $+/$- operations apply, e.g., `(A - B)` so long as the dimensions are conformable.

Matrix Multiplication - Example

$$
\begin{aligned}
\mathbf{AB} &= \left[\begin{array}{ccc} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{array}\right] \left[\begin{array}{cc} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{array}\right] \\
&= \left[\begin{array}{cc} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{array}\right].
\end{aligned}
$$

Matrix Multiplication in General

In general, if $\mathbf{A}$ has dimension $r \times c$ and $\mathbf{B}$ has dimension $c \times s$, the product $\mathbf{AB}$ is a matrix of dimension $r \times s$ whose element in the ith row and jth column is

$$\sum_{k=1}^{c} a_{ik} b_{kj},$$

so that

$$\underset{r \times s}{\mathbf{AB}} = \left[\sum_{k=1}^{c} a_{ik} b_{kj} \right] \quad i = 1, \ldots, r; \quad j = 1, \ldots s$$

In R, `AB = A %*% B`.

Matrix Multiplication: Alternative Views

$$\mathbf{AB} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}$$
$$= \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{bmatrix}.$$

Let $\mathbf{C} = \mathbf{AB}$. Recall that $\mathbf{C}_{ij} = \sum_{k=1}^{c} a_{ik} b_{kj}$.

Column form representation: We can represent columns of $\mathbf{C}$ as linear combinations of the columns of $\mathbf{A}$: $\mathbf{C}_j = \mathbf{A} \cdot \mathbf{B}_j$, where $\mathbf{B}_j$ is the jth column of $\mathbf{B}$ so that $\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_s]$.

We can represent $\mathbf{C}$ in the *outer product form*, i.e., $\mathbf{C} = \sum_{k=1}^{c} \mathbf{A}[,k] \cdot \mathbf{B}[k,]$ using `R` notation, where $\mathbf{A}[,k]$ is the kth column of $\mathbf{A}$ and $\mathbf{B}[k,]$ is the kth row of $\mathbf{B}$.

The Identity Matrix

The identity matrix, denoted by $\mathbf{I}$, is a diagonal matrix whose elements on the main diagonal are all 1s. Premultiplying or postmultiplying any $r \times c$ matrix $\mathbf{A}$ by the identity matrix (of conformable dimensions) leaves $\mathbf{A}$ unchanged. For example,

$$\mathbf{IA} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

Similarly,

$$\mathbf{AI} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

In general, for any $r \times r$ matrix $\mathbf{A}$ we have $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$.
In `R`, the $r \times r$ identity matrix can be created as `diag(r)`.

Linear Dependence/Multicollinearity

Let $\mathbf{A}$ be a matrix with columns $\mathbf{A}_1, \dots, \mathbf{A}_c$.

If one can find scalars $\lambda_1, \dots, \lambda_c$, not all zero, such that

$$\mathbf{A} \cdot \boldsymbol{\lambda} = \lambda_1 \mathbf{A}_1 + \lambda_2 \mathbf{A}_2 + \cdots + \lambda_c \mathbf{A}_c = \mathbf{0},$$

where $\mathbf{0}$ denotes the zero column vector, the column vectors are *linearly dependent*.

If the only set of scalars for which the equality holds is $\lambda_1 = 0, \dots, \lambda_c = 0$, the columns are *linearly independent*.

Let $\mathbf{X}$ be the design matrix for a multiple linear regression problem; i.e., columns of $\mathbf{X}$ are predictors/features. *Collinearity/multicollinearity* occurs when the columns matrix $\mathbf{X}$ are linearly dependent (loosely, contain redundant information).

Linear Dependence: an Illustration

Consider the following matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 5 & 1 \\ 2 & 2 & 10 & 6 \\ 3 & 4 & 15 & 1 \end{bmatrix}.$$

If $\lambda_1 = 5, \lambda_2 = 0, \lambda_3 = -1, \lambda_4 = 0$, then

$$5 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + 0 \begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix} - 1 \begin{bmatrix} 5 \\ 10 \\ 15 \end{bmatrix} + 0 \begin{bmatrix} 1 \\ 6 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

In general, linear dependence is not restricted to the situations where one column is a multiple of another column.

Rank of a Matrix

The rank of a matrix is defined to be the maximum number of linearly independent columns in the matrix.

We know that the rank of $\mathbf{A}$ in our earlier example cannot be 4, since the four columns are linearly dependent.

We can, however, find three columns $(1, 2,$ and $4)$ which are linearly independent. There are no scalars $\lambda_1, \lambda_2, \lambda_4$ such that $\lambda_1 \mathbf{A}_1 + \lambda_2 \mathbf{A}_2 + \lambda_4 \mathbf{A}_4 = \mathbf{0}$ other than $\lambda_1 = \lambda_2 = \lambda_4 = 0$. Thus, the rank of $\mathbf{A}$ in our example is 3.

The rank of a matrix is unique and can equivalently be defined as the maximum number of linearly independent rows. It follows that the rank of an $r \times c$ matrix cannot exceed $\min(r, c)$, the minimum of the two values r and c

Inverse of a Matrix

For a square matrix $\mathbf{A}$ of full rank, the inverse of $\mathbf{A}$ is a matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.
For example, the inverse of the matrix

$$\underset{2\times 2}{\mathbf{A}} = \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix}$$

is

$$\underset{2\times 2}{\mathbf{A}^{-1}} = \begin{bmatrix} -.1 & .4 \\ .3 & -.2 \end{bmatrix}$$

since

$$\mathbf{A}^{-1}\mathbf{A} = \begin{bmatrix} -.1 & .4 \\ .3 & -.2 \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{I}$$

or

$$\mathbf{A}\mathbf{A}^{-1} = \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} -.1 & .4 \\ .3 & -.2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{I}.$$

Example of an Inverse of a 3x3 Matrix

If

$$\underset{3\times 3}{\mathbf{B}} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix},$$

then

$$\mathbf{B}^{-1} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix}^{-1} = \begin{bmatrix} A & B & C \\ D & E & F \\ G & H & K \end{bmatrix},$$

where

$$\begin{array}{lll} A = (ek - fh)/Z & B = -(bk - ch)/Z & C = (bf - ce)/Z \\ D = -(dk - fg)/Z & E = (ak - cg)/Z & F = -(af - cd)/Z \\ G = (dh - eg)/Z & H = -(ah - bg)/Z & K = (ae - bd)/Z \end{array}$$

and

$$Z = a(ek - fh) - b(dk - fg) + c(dh - eg).$$

Z is called the determinant of the matrix $\mathbf{B}$.

Solving Systems of Linear Equations

A solution to the system of linear equations $\mathbf{A}\boldsymbol{x} = \mathbf{b}$ is a vector $\boldsymbol{x}^*$ for which the identity $\mathbf{A}\boldsymbol{x}^* = \mathbf{b}$ is satisfied.

One is generally interested in solving the systems of equations where $\mathbf{A}$ is a square $r \times r$ matrix. Such equations have a unique solution for a general right-hand side $\mathbf{b}$ if and only if $\mathbf{A}$ is of full rank (i.e., invertible), in which case the solution $\boldsymbol{x}^* = \mathbf{A}^{-1}\mathbf{b}$.

Although $\boldsymbol{x}^* = \mathbf{A}^{-1}\mathbf{b}$ is the "mathematical/theoretical" solution, in practice finding $\mathbf{A}^{-1}$ to solve the linear system is generally a bad idea (from the standpoint of numerical accuracy).

In `R`, the linear system $\mathbf{A}\boldsymbol{x} = \mathbf{b}$ may be solved as `solve(A,b)`. In the background, this linear system is solved using efficient matrix factorizations of $\mathbf{A}$, e.g., the "LU" factorization.

In the rare cases where $\mathbf{A}^{-1}$ itself is needed (e.g., when we need a covariance matrix of our least squares estimator $\widehat{\boldsymbol{\beta}}$), it can be found as `solve(A)`.

Common Matrix Algebra Identities

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$$
$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$$
$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$$
$$\mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB}$$
$$\lambda(\mathbf{A} + \mathbf{B}) = \lambda\mathbf{A} + \lambda\mathbf{B}$$
$$(\mathbf{A}')' = \mathbf{A}$$
$$(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$$
$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$$
$$(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$$
$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$
$$(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$$
$$\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$$
$$(\mathbf{A}')^{-1} = \left(\mathbf{A}^{-1}\right)'$$

Expectation of a Random Vector or a Matrix

In general, for a random vector $\mathbf{Y}$ the expectation is

$$\mathbf{E}\{\mathbf{Y}\} = [E\{Y_i\}] \quad i = 1, \ldots, n.$$

For a random matrix $\mathbf{Y}$ with dimension $n \times p$, the expectation is

$$\underset{n \times p}{\mathbf{E}\{\mathbf{Y}\}} = [E\{Y_{ij}\}] \quad i = 1, \ldots, n; \quad j = 1, \ldots, p.$$

Expectation of a Random Vector: SLR/MLR Example

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times 1}{\mathbf{E}\{\mathbf{Y}\}} \qquad\qquad + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} E\{Y_1\} \\ E\{Y_2\} \\ \vdots \\ E\{Y_n\} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} E\{Y_1\} + \varepsilon_1 \\ E\{Y_2\} + \varepsilon_2 \\ \vdots \\ E\{Y_n\} + \varepsilon_n \end{bmatrix}$$

Example: Covariance Matrix I

Let $\mathbf{Y}$ be a random vector consisting of three rvs Y_1, Y_2, Y_3.

The variances of the three rvs, $\sigma^2\{Y_i\}$, and the covariances between any two of the rvs, $\sigma\{Y_i, Y_j\}$, are assembled in the *covariance matrix* of $\mathbf{Y}$, denoted by $\boldsymbol{\sigma}^2\{\mathbf{Y},\}$ as follows:

$$\boldsymbol{\sigma}^2\{\mathbf{Y}\} = \begin{bmatrix} \sigma^2\{Y_1\} & \sigma\{Y_1, Y_2\} & \sigma\{Y_1, Y_3\} \\ \sigma\{Y_2, Y_1\} & \sigma^2\{Y_2\} & \sigma\{Y_2, Y_3\} \\ \sigma\{Y_3, Y_1\} & \sigma\{Y_3, Y_2\} & \sigma^2\{Y_3\} \end{bmatrix}.$$

Notice that $\sigma\{Y_2, Y_1\} = \sigma\{Y_1, Y_2\}$, since $\sigma\{Y_i, Y_j\} = \sigma\{Y_j, Y_i\}$ for all $i \neq j$, $\sigma^2\{\mathbf{Y}\}$ is a symmetric matrix.

In this course, the terms "covariance matrix" and "variance-covariance matrix" are used interchangeably.

Example: Covariance Matrix II

It follows readily that:

$$\boldsymbol{\sigma}^2\{\mathbf{Y}\} = \mathbf{E}\left\{(\mathbf{Y}-\mathbf{E}\{\mathbf{Y}\})\cdot(\mathbf{Y}-\mathbf{E}\{\mathbf{Y}\})'\right\}.$$

For our illustration, we have

$$\boldsymbol{\sigma}^2\{\mathbf{Y}\} = \mathbf{E}\left\{\begin{bmatrix} Y_1 - E\{Y_1\} \\ Y_2 - E\{Y_2\} \\ Y_3 - E\{Y_3\} \end{bmatrix} \cdot [Y_1 - E\{Y_1\}, Y_2 - E\{Y_2\}, Y_3 - E\{Y_3\}]\right\}.$$

If we define $\mathbf{Z} = (\mathbf{Y} - \mathbf{E}\{\mathbf{Y}\})$, then $\boldsymbol{\sigma}^2\{\mathbf{Y}\} = \mathbf{E}(\mathbf{Z}\cdot\mathbf{Z}')$ and $\left[\boldsymbol{\sigma}^2\{\mathbf{Y}\}\right]_{ij} = Cov(Y_i, Y_j) = Cov(Y_j, Y_i) = \left[\boldsymbol{\sigma}^2\{\mathbf{Y}\}\right]_{ji}$.

Cov(Yi,Yj)=E((Yi-E(Yi))*(Yj-E(Yj))

Covariance Matrix - General Case

The covariance matrix for a general $n \times 1$ random vector $\mathbf{Y}$ is

$$\boldsymbol{\sigma}^2\{\mathbf{Y}\} = \begin{bmatrix} \sigma^2\{Y_1\} & \sigma\{Y_1, Y_2\} & \cdots & \sigma\{Y_1, Y_n\} \\ \sigma\{Y_2, Y_1\} & \sigma^2\{Y_2\} & \cdots & \sigma\{Y_2, Y_n\} \\ \vdots & \vdots & & \vdots \\ \sigma\{Y_n, Y_1\} & \sigma\{Y_n, Y_2\} & \cdots & \sigma^2\{Y_n\} \end{bmatrix}.$$

Notice again that $\boldsymbol{\sigma}^2\{\mathbf{Y}\}$ is a symmetric matrix, i.e., $\left[\boldsymbol{\sigma}^2\{\mathbf{Y}\}\right]_{ij} = \left[\boldsymbol{\sigma}^2\{\mathbf{Y}\}\right]_{ji}$.

For notational transparency, the covariance matrix of $\mathbf{Y}$ is denoted here as $\boldsymbol{\sigma}^2\{\mathbf{Y}\}$. A more common notation is $Var(\mathbf{Y}) = \mathbf{\Sigma}_{\mathbf{Y}}$ whenever there are multiple random vectors under consideration, or $Var(\mathbf{Y}) = \mathbf{\Sigma}$ if there is no ambiguity.

Expectation and Covariance for a Linear Transformation

Frequently, we shall encounter a random vector $\mathbf{W}$ which is obtained by premultiplying the random vector $\mathbf{Y}$ by a constant matrix $\mathbf{A}$ (a matrix whose elements are fixed):

$$\mathbf{W} = \mathbf{AY}.$$

Here, $\mathbf{W}$ is called a linear transformation of $\mathbf{Y}$.

Some basic results for this case are

$$\begin{aligned} \mathbf{E}\{\mathbf{A}\} &= \mathbf{A} \\ \mathbf{E}\{\mathbf{W}\} = \mathbf{E}\{\mathbf{AY}\} &= \mathbf{AE}\{\mathbf{Y}\} \\ \boldsymbol{\sigma}^2\{\mathbf{W}\} = \boldsymbol{\sigma}^2\{\mathbf{AY}\} &= \mathbf{A}\boldsymbol{\sigma}^2\{\mathbf{Y}\}\mathbf{A}', \end{aligned}$$

where $\boldsymbol{\sigma}^2\{\mathbf{Y}\}$ is the variance-covariance matrix of $\mathbf{Y}$.

Expectation for a Linear Transformation

$$\underset{2\times 1}{\mathbf{E}\{\mathbf{W}\}} = \left[\begin{array}{rr} 1 & -1 \\ 1 & 1 \end{array}\right] \left[\begin{array}{c} E\left\{Y_1\right\} \\ E\left\{Y_2\right\} \end{array}\right] = \left[\begin{array}{c} E\left\{Y_1\right\} - E\left\{Y_2\right\} \\ E\left\{Y_1\right\} + E\left\{Y_2\right\} \end{array}\right]$$

Covariance Matrix for a Linear Transformation

$$\underset{2\times 2}{\boldsymbol{\sigma}^2}\{\mathbf{W}\} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \sigma^2\{Y_1\} & \sigma\{Y_1, Y_2\} \\ \sigma\{Y_2, Y_1\} & \sigma^2\{Y_2\} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2\{Y_1\} + \sigma^2\{Y_2\} - 2\sigma\{Y_1, Y_2\} & \sigma^2\{Y_1\} - \sigma^2\{Y_2\} \\ \sigma^2\{Y_1\} - \sigma^2\{Y_2\} & \sigma^2\{Y_1\} + \sigma^2\{Y_2\} + 2\sigma\{Y_1, Y_2\} \end{bmatrix}$$

Simple Linear Regression Model Using Equations

Our model for the individual observations Y_i is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n.$$

This implies

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1, \\ Y_2 &= \beta_0 + \beta_1 X_2 + \varepsilon_2, \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_n + \varepsilon_n. \end{aligned}$$

Simple Linear Regression using Matrix Algebra

$$
\begin{aligned}
\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} &= \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \\
&= \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 X_n + \varepsilon_n \end{bmatrix}
\end{aligned}
$$

Multiple Linear Regression Model Using Matrix Algebra I

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \underset{n\times p}{\mathbf{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Multiple Linear Regression Model Using Matrix Algebra II

In matrix terms, a multiple linear regression (MLR) model is

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}} \cdot \overset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}, \quad \text{where}$$

$\mathbf{Y}$ is a vector of responses,
$\boldsymbol{\beta}$ is a vector of parameters/coefficients,
$\mathbf{X}$ is a matrix of constants (the design matrix), and
ε is a vector of uncorrelated errors with expectation $\mathbf{E}\{\varepsilon\} = \mathbf{0}$ and covariance matrix

$$\underset{n\times n}{\boldsymbol{\sigma}^2}\{\varepsilon\} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}.$$

Least Squares (LS) Estimation for the MLR

$$Q(\mathbf{b}) = \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_{i1} - \cdots - b_{p-1} X_{i,p-1}\right)^2 .$$

The least squares (LS) solution/estimators are those values of $b_0, b_1, \ldots, b_{p-1}$ that minimize the SSE, here denoted by Q. Define

$$\underset{p\times 1}{\mathbf{b}} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} .$$

In matrix notation,

$$Q(\mathbf{b}) = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b}) = \|\mathbf{Y} - \mathbf{X}\mathbf{b}\|_2^2.$$

Expanding, we obtain

$$Q(\mathbf{b}) = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}.$$

Normal Equations and the LS Estimator

To find the minimizer of $Q(\mathbf{b})$, differentiate $Q(\mathbf{b})$ wrt $\mathbf{b}$:

$$\frac{\partial Q(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\mathbf{b},$$

set the derivative to $\mathbf{0}$ and solve.

Notice the solution must satisfy $\mathbf{X}' \cdot (\mathbf{Y} - \mathbf{X}\mathbf{b}) = \mathbf{0}$.

The least squares *normal equations* are $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$.

To solve, premultiply both sides by $(\mathbf{X}'\mathbf{X})^{-1}$ (assume this exists):

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}.$$

Since $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$ and $\mathbf{I}\mathbf{b} = \mathbf{b}$, we then find the solution

$$\underset{m\times 1}{\mathbf{b}^*} = \underset{m\times m}{(\mathbf{X}'\mathbf{X})^{-1}}\underset{m\times 1}{\mathbf{X}'\mathbf{Y}}, \quad \text{the LS estimator that minimizes } Q(\mathbf{b}).$$

Statistical Model for MLR

How to solve this problem "statistically"? Assume

$$Y_i = \underbrace{\boldsymbol{x}_i'\boldsymbol{\beta}}_{\sum_{j=1}^{p} x_{ij}\beta_j} + \varepsilon_i,$$

where ε_i's are independent Normal$(0, \sigma^2)$ rvs. In vector-matrix form,

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}} \overset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}.$$

Here, $Y_i \sim$ Normal$(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$; the Y_i's are independent but not identically distributed.

Since we know the joint pdf of the Y_i's, we can estimate $\boldsymbol{\beta}$ and σ^2 using the MLE.

Statistical Estimation by MLE I

Step 1: write down the likelihood:

$$
\begin{aligned}
\mathrm{L}\left(y_1, \ldots, y_n \mid \boldsymbol{\beta}, \sigma^2\right) &= \prod_{i=1}^{n} f_i(y_i \mid \boldsymbol{\beta}, \sigma^2) \\
&= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}_{(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})}\right).
\end{aligned}
$$

Step 2: obtain the log-likelihood:

$$
\begin{aligned}
\ell\left(\boldsymbol{\beta}, \sigma^2\right) &= \ln\left(\mathrm{L}\left(y_1, \ldots, y_n \mid \boldsymbol{\beta}, \sigma^2\right)\right) \\
&= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}).
\end{aligned}
$$

Statistical Estimation by MLE II

Step 3. Find the gradient of $\ell\left(\boldsymbol{\beta}, \sigma^2\right)$ with respect to $\boldsymbol{\beta}$ and σ^2, set the gradient to 0, solve for $\boldsymbol{\beta}$ and σ^2.

$$\frac{\partial \ell\left(\boldsymbol{\beta}, \sigma^2\right)}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2} 2\mathbf{X}'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\frac{\partial \ell\left(\boldsymbol{\beta}, \sigma^2\right)}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} - \frac{1}{2}\frac{(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})}{(\sigma^2)^2}$$

The solution is $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{y}$ and $\widehat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$.

Step 4. Make sure we found the maximizers of $\ell\left(\boldsymbol{\beta}, \sigma^2\right)$ by showing that all eigenvalues of the matrix of second derivatives of $\ell\left(\boldsymbol{\beta}, \sigma^2\right)$—known as the Hessian—are negative. (Equivalently, $(-1) \cdot$ Hessian is positive definite.)

The Vector of Fitted Values and the Hat Matrix

Notice that the expressions for the LS and ML estimators coincide. Let's use $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ for notational transparency.

Let's express the vector of fitted values $\widehat{\mathbf{Y}}$ using the formula for $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y},$$

where

$$\underset{n\times n}{\mathbf{H}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'.$$

Notice that $\mathbf{H}$ is symmetric $(\mathbf{H} = \mathbf{H}')$ and idempotent, i.e.,

$$\mathbf{H} = \mathbf{H}\cdot\mathbf{H}.$$

The Hat Matrix and the Vector of Residuals

We can express the vector of residuals $\mathbf{e}$ as

$$\underset{n\times 1}{\mathbf{e}} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\overset{n\times n}{\mathbf{I} - \mathbf{H}}) \cdot \underset{n\times 1}{\mathbf{Y}},$$

where $\mathbf{H}$ is the hat matrix. The matrix $(\mathbf{I} - \mathbf{H})$, like the matrix $\mathbf{H}$, is symmetric and idempotent.

The variance-covariance matrix of the vector of residuals $\mathbf{e}$ is

$$\boldsymbol{\sigma}^2\{\mathbf{e}\} = \sigma^2(\mathbf{I} - \mathbf{H})$$

and is estimated by

$$\mathbf{s}^2\{\mathbf{e}\} = \widehat{\sigma}^2(\mathbf{I} - \mathbf{H}),$$

where $\widehat{\sigma}^2 = SSE/(n-p) = \mathbf{e}'\mathbf{e}/(n-p)$ is referred to as the MSE in ANOVA tables.

Distributional Results when $\boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I})$

The ML estimatOR is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{A}\mathbf{Y}$. Hence

$$\mathbf{E}(\widehat{\boldsymbol{\beta}}) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' \cdot \mathbf{E}(\mathbf{Y}) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$
$$\boldsymbol{\sigma}^2\{\widehat{\boldsymbol{\beta}}\} = \mathbf{A} \cdot \boldsymbol{\sigma}^2\{\mathbf{Y}\} \cdot \mathbf{A}' = \ldots = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}.$$

Additionally,

$$\widehat{\boldsymbol{\beta}} \sim \text{Multivariate.Normal}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2\{\widehat{\boldsymbol{\beta}}\}), \quad \text{and}$$
$$\boldsymbol{a}'\widehat{\boldsymbol{\beta}} \sim \text{Normal}(\boldsymbol{a}'\boldsymbol{\beta}, \boldsymbol{a}'\boldsymbol{\sigma}^2\{\widehat{\boldsymbol{\beta}}\}\boldsymbol{a}),$$

where $\boldsymbol{a}$ is a column vector of constants.

Lastly,

$$\frac{\boldsymbol{a}'\widehat{\boldsymbol{\beta}} - \boldsymbol{a}'\boldsymbol{\beta}}{\widehat{\sigma}\sqrt{\boldsymbol{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\boldsymbol{a}}} \sim t(n-p).$$

Estimation/Prediction of the Mean Response

For given values of $X_1, \ldots, X_{p-1}$, denoted by $x_{h,1}, \ldots, x_{h,p-1}$, the mean response is denoted by $\mathbf{E}\{Y_h\}$. We define the vector $\boldsymbol{x}_h$

$$\underset{p\times 1}{\boldsymbol{x}_h} = \begin{bmatrix} 1 \\ x_{h,1} \\ \vdots \\ x_{h,p-1} \end{bmatrix},$$

so that the mean response to be estimated is

$$\mathbf{E}\{Y_h\} = \boldsymbol{x}_h' \boldsymbol{\beta}.$$

The estimated mean response corresponding to $\boldsymbol{x}_h$ is

$$\widehat{Y}_h = \boldsymbol{x}_h' \widehat{\boldsymbol{\beta}}.$$

Estimation/Prediction of the Mean Response

This estimator $\widehat{Y}_h = \boldsymbol{x}_h' \widehat{\boldsymbol{\beta}}$ is unbiased:

$$\mathbf{E}\left\{\widehat{Y}_h\right\} = \boldsymbol{x}_h' \boldsymbol{\beta} = \mathbf{E}\left\{Y_h\right\}$$

and its variance is

$$\boldsymbol{\sigma}^2\left\{\widehat{Y}_h\right\} = \sigma^2 \boldsymbol{x}_h' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \boldsymbol{x}_h.$$

This variance can be expressed as a function of the variance-covariance matrix of the estimated regression coefficients

$$\boldsymbol{\sigma}^2\left\{\widehat{Y}_h\right\} = \boldsymbol{x}_h' \boldsymbol{\sigma}^2\{\widehat{\boldsymbol{\beta}}\} \boldsymbol{x}_h.$$

Confidence Interval for the Mean of the Response Y_h

Notice that the variance $\boldsymbol{\sigma}^2\left\{\widehat{Y}_h\right\}$ is a function of the covariance matrix $\boldsymbol{\sigma}^2\{\widehat{\boldsymbol{\beta}}\} = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$.

The estimated variance $s^2\left\{\widehat{Y}_h\right\}$ is given by

$$s^2\left\{\widehat{Y}_h\right\} = MSE \cdot \left(\boldsymbol{x}_h'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\boldsymbol{x}_h\right) = \boldsymbol{x}_h'\mathbf{s}^2\{\mathbf{b}\}\boldsymbol{x}_h.$$

The $(1-\alpha)$ confidence limits for $E\left\{Y_h\right\}$ are

$$\widehat{Y}_h \pm t(1-\alpha/2; n-p)s\left\{\widehat{Y}_h\right\}.$$

Here, $MSE = \frac{SSE}{(n-p)} = \widehat{\sigma}^2$ is the square of the "residual standard error" (RSE) reported by `R` in the summary of an `lm` object.

Prediction Interval for the Response Y_h

The $(1-\alpha)$ prediction limits for a new observation Y_h corresponding to $\boldsymbol{x}_h$, the specified values of the covariates, are

$$\widehat{Y}_h \pm t(1-\alpha/2; n-p) \cdot s\{\text{pred}\},$$

where

$$s^2\{\text{pred}\} = MSE + s^2\left\{\widehat{Y}_h\right\} = MSE \cdot \left(1 + \boldsymbol{x}_h' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \boldsymbol{x}_h\right)$$

and $s^2\left\{\widehat{Y}_h\right\}$ is given above.

In `R`, point-level predictions $\widehat{Y}_h$, confidence intervals and prediction intervals can be obtained using the function `predict.lm`.