Intermediate Social Statistics

Lecture 1. Matrices and Linear Regression

Tom A.B. Snijders

University of Oxford

January, 2008

This course is taught by Raymond Duch and Tom Snijders.

Computer classes by David Armstrong and Mark Pickup.

Course websites:

`http://www.raymondduch.com/`

(see *teaching*)

`http://www.stats.ox.ac.uk/~snijders/iss.htm`

Today:

A recap of linear regression,

and an introduction to matrix calculus.

Regression analysis

Regression is the basic method for relating *independent* or *explanatory* or *predictor* variables to *dependent* or *outcome* variables:

$$X \Rightarrow Y$$

The formula for simple linear regression reads

$$Y_i = \beta_1 + \beta_2 x_i + E_i \ ,$$

where the expected value of E_i is 0.

(Capital letters for random and abstract variables;
small letters for deterministic variables;
Greek letters for statistical parameters.)

We start counting βs with 1 for later convenience.

The main interest is in situations where there are multiple predictor variables $X_1, X_2, \ldots, X_p$:

$$\begin{aligned} Y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_p x_{ip} + E_i \\ &= \sum_{j=1}^{p} \beta_j x_{ij} + E_i \, . \end{aligned}$$

Motive: control one effect (e.g., of X_p) for the other effects (e.g., X_1 to X_{p-1}).
Here it is very convenient to define the variable X_1 as the constant with all values $x_{i1} = 1$: the *intercept*.

For each case (subject, respondent) i,
Y_i is the observed outcome,
$\sum_{j=1}^{p} \beta_j x_{ij}$ is the expected value of the outcome,
and E_i is the random disturbance, also called error.

The matrix notation for the equation

$$Y_i = \sum_{j=1}^{p} \beta_j x_{ij} + E_i \text{ for } i = 1, \ldots, n$$

is

$$Y = X\beta + E\,.$$

This illustrates the tremendous simplification that is obtained when replacing scalar notation by matrix notation.
(scalar = real number, as compared to a matrix)

Vectors and matrices

Here Y, β and E are column vectors:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{pmatrix}, \qquad E = \begin{pmatrix} E_1 \\ E_2 \\ . \\ . \\ . \\ E_n \end{pmatrix},$$

of lengths n, p, and n, respectively;

and X is a n by p matrix:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

X is called the *design matrix*:
this is the matrix that contains all values
of independent variables for all cases.

An "n by p", or "$n \times p$", matrix has n rows and p columns,
and x_{ij} is the element of X in row i and column j:
the subscripts are first row number, then column number.

These numbers n and p are called
the orders, or dimensions, or sizes, of the matrix.

Vectors are matrices with only one row or only one column.

A a "column vector of length n" is also a "n by 1 matrix";
and a "row vector of length n" is also a "1 by n matrix",

$$\left(a_1, a_2, \ldots, a_n\right) .$$

We then do not write the superfluous index 1.

A scalar (real number) is a 1 by 1 matrix....

Thus, basically, a matrix is a rectangular array of numbers.
Matrix calculus is important for handling multiple variables.
For understanding of statistical methods,
basic knowledge of matrix calculus is very useful.

Thus, basically, a matrix is a rectangular array of numbers.
Matrix calculus is important for handling multiple variables.
For understanding of statistical methods,
basic knowledge of matrix calculus is very useful.

It is most fruitful to regard a matrix as a linear transformation.
In the regression equation

$$Y = X\beta + E \,,$$

X transforms the vector of p parameters β
to a vector of n expected outcomes $X\beta$.

This equation shows *matrix multiplication* $X\beta$
and *matrix addition* $X\beta + E$.

Matrix addition and multiplication

The same matrix equation written out in elements:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{pmatrix} + \begin{pmatrix} E_1 \\ E_2 \\ . \\ . \\ . \\ E_n \end{pmatrix} .$$

How are the multiplication of X by β,
and the addition of $X\beta$ and E defined?

Matrix addition is simple:

it operates on the corresponding elements separately.

E.g.,

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{pmatrix}$$

Matrix multiplication is more complicated.

For multiplying the $n \times p$ matrix X by the $p \times 1$ vector β the equation is

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{pmatrix} = \begin{pmatrix} x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1p}\beta_p \\ x_{21}\beta_1 + x_{22}\beta_2 + \ldots + x_{2p}\beta_p \\ . \\ . \\ . \\ x_{n1}\beta_1 + x_{n2}\beta_2 + \ldots + x_{np}\beta_p \end{pmatrix}.$$

$\Rightarrow$ For each case separately,
the variables are multiplied by the parameters, and added.

$\Rightarrow$ For each row of X separately, the row elements are multiplied by the corresponding elements of β, and added.

General definition of matrix multiplication

The general definition of matrix multiplication $AB = C$
is as follows.

A $n \times p$ matrix A can be multiplied by a $p \times k$ matrix B,
and this yields a $n \times k$ matrix C.
(Earlier with the multiplication $X\beta$ we had $k = 1$.)

Element (i, j) – i.e., in row i and column j – of matrix C
is given by

$$c_{ij} = a_{i1}\, b_{1j} + a_{i2}\, b_{2j} + a_{i3}\, b_{3j} + \ldots + a_{p1}\, b_{pj} \,.$$

In words:
we take row i of A and column j of β;
these have equally many elements!!!;
we multiply corresponding elements and add the products.

This product of two vectors of the same length,
the first a row vector and the second a column vector,
is called the *inner product* or *dot product*.

If we have a $p \times 1$ column vector $a = \begin{pmatrix} a_1 \\ a_2 \\ . \\ a_p \end{pmatrix}$,

then the *transposed vector* a' is defined as
the row vector with the same elements, and

$$a'a = \sum_{i=1}^{p} a_i^2 \ .$$

Thus, a sum of squares is easy to define in matrix notation.

Examples:

$$\begin{pmatrix} -1 & 0 & 1 \\ 1 & 2 & 3 \\ 2 & 2 & 2 \end{pmatrix} \begin{pmatrix} 5 \\ 6 \\ 7 \end{pmatrix} = \begin{pmatrix} ((-1) \times 5) & + & (0 \times 6) & + & (1 \times 7) \\ (1 \times 5) & + & (2 \times 6) & + & (3 \times 7) \\ (2 \times 5) & + & (2 \times 6) & + & (2 \times 7) \end{pmatrix}$$

$$= \begin{pmatrix} 2 \\ 38 \\ 36 \end{pmatrix}$$

Examples:

$$\begin{pmatrix} -1 & 0 & 1 \\ 1 & 2 & 3 \\ 2 & 2 & 2 \end{pmatrix} \begin{pmatrix} 5 \\ 6 \\ 7 \end{pmatrix} = \begin{pmatrix} ((-1)\times 5) & + & (0\times 6) & + & (1\times 7) \\ (1\times 5) & + & (2\times 6) & + & (3\times 7) \\ (2\times 5) & + & (2\times 6) & + & (2\times 7) \end{pmatrix}$$

$$= \begin{pmatrix} 2 \\ 38 \\ 36 \end{pmatrix}$$

and similarly

$$\begin{pmatrix} -1 & 0 & 1 \\ 1 & 2 & 3 \\ 2 & 2 & 2 \end{pmatrix} \begin{pmatrix} 5 & 0 \\ 6 & 1 \\ 7 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 38 & 8 \\ 36 & 6 \end{pmatrix} .$$

A special matrix is the $n \times n$ identity matrix I_n
with diagonal elements 1 and off-diagonal elements 0.

For example:

$$I_3 \begin{pmatrix} 5 \\ 6 \\ 7 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 5 \\ 6 \\ 7 \end{pmatrix} = \begin{pmatrix} 5 \\ 6 \\ 7 \end{pmatrix} .$$

A special matrix is the $n \times n$ identity matrix I_n
with diagonal elements 1 and off-diagonal elements 0.

For example:

$$I_3 \begin{pmatrix} 5 \\ 6 \\ 7 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 5 \\ 6 \\ 7 \end{pmatrix} = \begin{pmatrix} 5 \\ 6 \\ 7 \end{pmatrix} .$$

and, generally,

$$I_n A = A I_k = A$$

for any $n \times k$ matrix A.

Matrix inversion

Since matrix addition operates elementwise,
matrix subtraction is simple: e.g.,

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} - \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11} - b_{11} & a_{12} - b_{12} \\ a_{21} - b_{21} & a_{22} - b_{22} \end{pmatrix}$$

The analogue of division is more complicated,
as matrix multiplication is more complicated.

Mathematically, the definition of division is done by defining the inverse:

$$\frac{1}{a} \text{ is the number for which } a \times \frac{1}{a} = 1 \, .$$

This works only if $a \neq 0$.

We also write

$$\frac{1}{a} = a^{-1} \, .$$

Mathematically, the definition of division is done by defining the inverse:

$$\frac{1}{a} \text{ is the number for which } a \times \frac{1}{a} = 1 \, .$$

This works only if $a \neq 0$.

We also write

$$\frac{1}{a} = a^{-1} \, .$$

Similarly, for matrices, the inverse is defined as follows:

For a $n \times n$ matrix A,

$$A^{-1} \text{ is the matrix for which } A\,A^{-1} = A^{-1}\,A = I_n \, .$$

Only square matrices (equally many rows as columns) are invertible.

For diagonal matrices, inversion is simple:

$$\begin{pmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{a_{11}} & 0 & 0 \\ 0 & \frac{1}{a_{22}} & 0 \\ 0 & 0 & \frac{1}{a_{33}} \end{pmatrix}.$$

This shows that not all matrices have an inverse:
e.g., a diagonal matrix with at least
one zero diagonal element does not.
Non-invertible matrices are called *singular*.

Only square matrices (equally many rows as columns) are invertible.

For diagonal matrices, inversion is simple:

$$\begin{pmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{a_{11}} & 0 & 0 \\ 0 & \frac{1}{a_{22}} & 0 \\ 0 & 0 & \frac{1}{a_{33}} \end{pmatrix}.$$

This shows that not all matrices have an inverse:
e.g., a diagonal matrix with at least
one zero diagonal element does not.
Non-invertible matrices are called *singular*.

For non-diagonal square matrices, there exist algorithms (computational procedures) for calculating inverses;
e.g., see Freedman, Section 3.2.

Matrix transposition

Another basic matrix operation is *transposition*,
denoted by a prime ($'$)
which is simply the interchange between
the roles of the rows and the columns: e.g.,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix}' = \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \\ a_{14} & a_{24} & a_{34} \end{pmatrix}.$$

Matrix transposition

Another basic matrix operation is *transposition*,
denoted by a prime ($'$)
which is simply the interchange between
the roles of the rows and the columns: e.g.,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix}' = \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \\ a_{14} & a_{24} & a_{34} \end{pmatrix} .$$

Generally,

$$\text{Element } (i,j) \text{ of } A' \text{ is } a_{ji} .$$

Next to A', also the notation A^T is in use.

Some of the usual properties for operating with real numbers also hold for matrices:

$$\begin{aligned} (AB)C &= A(BC) \\ A(B+C) &= AB + AC \\ (A+B)C &= AC + AB \end{aligned}$$

Some of the usual properties for operating with real numbers also hold for matrices:

$$\begin{aligned} (AB)C &= A(BC) \\ A(B+C) &= AB + AC \\ (A+B)C &= AC + AB \end{aligned}$$

while others don't:

generally, AB is not equal to BA

and even if AB is defined BA may be undefined.

However, for $p \times 1$ vectors a and b, it does hold that $a'b = b'a$;
This is a special case of

$$(AB)' = B'A' .$$

also note that $a'b$ is a 1×1 matrix, i.e., a scalar,
so that $a'b = (a'b)'$.

However, for $p \times 1$ vectors a and b, it does hold that $a'b = b'a$;
This is a special case of

$$(AB)' = B'A' .$$

also note that $a'b$ is a 1×1 matrix, i.e., a scalar,
so that $a'b = (a'b)'$.

Further, it holds that

$$\left(A^{-1}\right)' = \left(A'\right)^{-1}$$

if the inverses exist.

The linear regression model again

We return to the linear regression model

$$Y = X\beta + E\,.$$

When the i'th row vector of X is denoted X_i,
for case i the equation can be written as

$$Y_i = X_i\beta + E_i\,.$$

Check:
this is a scalar equation;
X_i is a $1 \times p$ vector and β is a $p \times 1$ vector,
which two can be multiplied and yield a scalar.

Ordinary least squares estimators

The ordinary least squares (OLS) estimator for the parameter vector β is the value for which the sum of squared deviations

$$\sum_i (Y_i - X_i \beta)^2 = (Y - X\beta)' (Y - X\beta)$$

is minimal.

The estimator is denoted $\hat{\beta}$ and is also called the OLSE.

Ordinary least squares estimators

The ordinary least squares (OLS) estimator for the parameter vector β is the value for which the sum of squared deviations

$$\sum_i (Y_i - X_i \beta)^2 = (Y - X\beta)'(Y - X\beta)$$

is minimal.

The estimator is denoted $\hat{\beta}$ and is also called the OLSE.

If the disturbances E_i are independent and normally distributed with constant variances, then this is also the maximum likelihood estimator.

Special cases - scalar notation

First, suppose that there is no 'real' regression, and the model contains just the intercept: $p = 1$, $x_{i1} = 1$ for all i.

Then the OLSE is

$$\hat{\beta}_1 = \bar{y}_\cdot = \frac{1}{n}\sum_{i=1}^{n} y_i \ .$$

Second, suppose that there is an intercept and one predictor variable,
$p = 2$, $x_{i1} = 1$ for all i.

Then the OLSE is

$$\begin{aligned} \hat{\beta}_1 &= \bar{y}_. - \hat{\beta}_2 \bar{x}_. \\ \hat{\beta}_2 &= \frac{\sum_{i=1}^n (x_{i2} - x_{.2})(y_i - y_.)}{\sum_{i=1}^n (x_{i2} - x_{.2})^2}, \end{aligned}$$

where $x_{.2} = \frac{1}{n} \sum_{i=1}^n x_{i2}$.

This can be proved, e.g., by differential calculus.

For $p \geq 3$ the equations become horrible.

OLS estimator - matrix notation

In matrix notation, the OLSE is given by

$$\hat{\beta} = (X'X)^{-1} X'Y \, .$$

Exercise 1. Check the orders of the matrices for the multiplications.

Exercise 2. Check this equation for $p = 1$ (intercept only).

A geometric approach to the OLSE

The OLSE can be derived using calculus as in the lecture notes by Stephen Fisher.

A different approach is geometric:
$n \times 1$ column vectors are vectors in n-dimensional space.

The length of a vector a is defined as

$$\|a\| = \sqrt{a'a} = \sqrt{\sum_{i=1}^{n} a_i^2} \, .$$

A geometric approach to the OLSE

The OLSE can be derived using calculus as in the lecture notes by Stephen Fisher.

A different approach is geometric:
$n \times 1$ column vectors are vectors in n-dimensional space.

The length of a vector a is defined as

$$\|a\| = \sqrt{a'a} = \sqrt{\sum_{i=1}^{n} a_i^2} \, .$$

Vectors a and b are orthogonal ($a \perp b$) if $a'b = 0$,
which is the same as $b'a = 0$.
Orthogonality is the sample analogue of non-correlatedness!

Pythagoras' Theorem

If $a \perp b$, then $\|a+b\|^2 = \|a\|^2 + \|b\|^2$.

Pythagoras' Theorem

If $a \perp b$, then $\|a+b\|^2 = \|a\|^2 + \|b\|^2$.

Proof.

$$\begin{aligned} \|a+b\|^2 &= (a+b)'(a+b) = a'a + b'a + a'b + b'b \\ &= a'a + b'b = \|a\|^2 + \|b\|^2 . \end{aligned}$$

Recall that the OLSE is $\hat{\beta} = (X'X)^{-1} X'Y$.
We define the vector of residuals by $\hat{E} = Y - X\hat{\beta}$.

Proposition 1. $\hat{E}'X = 0$
(The residual vector is orthogonal to the variables in the columns of X).

Recall that the OLSE is $\hat{\beta} = (X'X)^{-1} X'Y$.
We define the vector of residuals by $\hat{E} = Y - X\hat{\beta}$.

Proposition 1. $\hat{E}'X = 0$
(The residual vector is orthogonal to the variables in the columns of X).
Proof.

$$\begin{aligned} \hat{E}'X &= (Y - X(X'X)^{-1} X'Y)'X = (Y' - Y'X(X'X)^{-1}X')X \\ &= Y'(X - X(X'X)^{-1}X'X) = Y'(X - X) = 0 . \end{aligned}$$

Proposition 1 is basic to the geometric explanation of the reason why $(X'X)^{-1}X'Y$ is the OLS estimator.

Proposition 2. For any $p \times 1$ vector γ,

$$\|Y - X\gamma\|^2 = \|Y - X\hat{\beta}\|^2 + \|X\hat{\beta} - X\gamma\|^2 \geq \|Y - X\hat{\beta}\|^2 .$$

This means that the OLSE has the smallest possible sum of squared residuals.

Proposition 1 is basic to the geometric explanation of the reason why $(X'X)^{-1}X'Y$ is the OLS estimator.

Proposition 2. For any $p \times 1$ vector γ,

$$\|Y - X\gamma\|^2 = \|Y - X\hat{\beta}\|^2 + \|X\hat{\beta} - X\gamma\|^2 \geq \|Y - X\hat{\beta}\|^2 .$$

This means that the OLSE has the smallest possible sum of squared residuals.

Proof.
Proposition 1 implies that $\hat{E} = Y - X\hat{\beta}$ is orthogonal to $X\gamma - X\hat{\beta}$:
$\hat{E}'(X\hat{\beta} - X\gamma) = \hat{E}'X(\hat{\beta} - \gamma) = 0(\hat{\beta} - \gamma) = 0$.
The rest follows immediately with the aid of Pythagoras' theorem.

Once the estimate $\hat{\beta}$ has been obtained, we can calculate the *fitted values*

$$\hat{Y} = X\hat{\beta} .$$

As we can write this in the form

$$\hat{Y} = X(X'X)^{-1}X'Y ,$$

premultiplying by $X(X'X)^{-1}X'$ means putting a hat on Y; therefore, $H = X(X'X)^{-1}X'$ is called the *hat matrix*.

Ideally, fitting the values means that we average out
as much as possible of the random disturbances;
but if the diagonal value h_{ii} of H is close to 1,
then the outcome Y_i, and the disturbance E_i,
have a dominating influence on the fitted value $\hat{Y}_i$.
This is undesired.
Therefore the diagonal values of the hat matrix
are diagnostic values expressing the *leverage* of case i.

Ideally, fitting the values means that we average out
as much as possible of the random disturbances;
but if the diagonal value h_{ii} of H is close to 1,
then the outcome Y_i, and the disturbance E_i,
have a dominating influence on the fitted value $\hat{Y}_i$.
This is undesired.
Therefore the diagonal values of the hat matrix
are diagnostic values expressing the *leverage* of case i.

It can be proved that the diagonal values of the hat matrix
all are positive and that they sum to p.

Ideally, fitting the values means that we average out
as much as possible of the random disturbances;
but if the diagonal value h_{ii} of H is close to 1,
then the outcome Y_i, and the disturbance E_i,
have a dominating influence on the fitted value $\hat{Y}_i$.
This is undesired.
Therefore the diagonal values of the hat matrix
are diagnostic values expressing the *leverage* of case i.

It can be proved that the diagonal values of the hat matrix
all are positive and that they sum to p.

Note that the hat matrix depends on the design matrix only,
and not on the outcomes Y;
cases with high h_{ii} may be considered outliers in the space of
independent variables.

These topics are treated more extensively in Freedman, Chapters 2-3.

Reading list for next week:
Freedman Chapter 3; Long, Chapter 2.

For students who do not have expertise in Stata, there is a class on Thursday of week 3 at 4-6 pm taught by David Armstrong and Mark Pickup.