Matrix Representation of the Regression Model

1. Introduction to Matrices

These notes cover the following:

1. properties of, and operations on, matrices and vectors
2. matrix representation of the linear regression model, both simple and multiple
3. derivation of the most important aspects of the regression model matrix representation

The purpose of all this is to develop the formal statistical machinery of the multiple regression model to clear the way for discussion of substantive and practical aspects of multiple regression. This material is in part foundational: matrices are almost essential for advanced methods beyond regression analysis.

Matrix notation represents a set of conventions to represent tables of numbers and operations on these numbers. It allows representation of the simple and multiple regression model in a compact form.

1. Matrices

A matrix is a rectangular array of numbers or symbols. Examples of matrices are

$$\mathbf{A} = \begin{bmatrix} 2.1 & 1.7 \\ 4.0 & -2.3 \\ 5.0 & 0.3 \\ 4.7 & -1.5 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 1.0 & .32 & .55 \\ .32 & 1.0 & .61 \\ .55 & .61 & 1.0 \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

- Matrices are usually designated by bold capital letters.
- Dimensions of a matrix are specified as the number of rows followed by the number of columns: **A** is 4 × 2, **B** and **D** are 3 × 3, and **C** is 3 × 2.
- An element of a matrix **C** is identified as c_ij with the first subscript referring to the row and the second to the column.
- A square matrix is a matrix with the same number of rows and columns. **B** and **D** are square matrices. **B** and **D** are also symmetric matrices: each row is identical with the same-numbered column, so the matrix is symmetrical around the leading or main diagonal that runs from upper-left to bottom-right corner. (A symmetric matrix has to be square, but a square matrix is not necessarily symmetric.)
- Matrix **D** is a diagonal matrix; all entries are zero except those on the main diagonal. A diagonal matrix with has only 1s on the main diagonal is designated matrix **I** and called an identity matrix. A matrix of all zeroes is called a null matrix. Examples of an identity matrix and a null matrix are

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{0} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

2. Transpose of a Matrix

The transpose of the matrix **A**, denoted $\mathbf{A}'$, is the matrix obtained by interchanging each row by the same-numbered column, so that the first column of **A** becomes the first row of $\mathbf{A}'$, and so on. (Formally, if A = [a_ij] then $\mathbf{A}'$ = [a_ji].) For example the transpose $\mathbf{A}'$ of the matrix **A** above is

$$\mathbf{A}' = \begin{bmatrix} 2.1 & 4.0 & 5.0 & 4.7 \\ 1.7 & -2.3 & 0.3 & -1.5 \end{bmatrix}$$

An alternative notation found for the transpose $\mathbf{A}'$ is $\mathbf{A}^{\mathbf{T}}$. The transpose of a 4 × 2 matrix has dimensions 2 × 4 because rows and columns are interchanged. A symmetric matrix is equal to its transpose so that $\mathbf{A}' = \mathbf{A}$.

3. Vectors and Scalars

A single row or column of numbers is called a *vector*.

- Column vectors are designated by lower case bold-face letters, e.g. **a**.
- Row vectors are viewed as transposes of column vectors and marked with a prime, e.g. $\mathbf{b}'$. (When used alone the term vector refers to a column vector.)

Single numbers, as used in ordinary arithmetic, are called *scalars*. Scalars are usually designated by lower case roman or Greek letters in regular weight, e.g. c or λ (lambda).

The following are examples of vectors and scalars:

$$\mathbf{a} = \begin{bmatrix} 1.2 \\ 3 \\ 1.7 \end{bmatrix} \qquad \mathbf{b}' = [1.1 \quad .5 \quad 2] \qquad c = 22.6$$

4. Equality of Matrices

Two matrices **A** and **B** are equal if they have the same dimensions and all corresponding elements are equal.

5. Addition and subtraction of Matrices

The matrices involved in addition or subtraction must have the same dimensions. Two matrices are added together by adding algebraically the corresponding elements of the two matrices. A matrix may be subtracted from another by reversing the sign of its elements and then adding the corresponding elements. Two columns vectors of the same length, or two row vectors of the same length, may be added or subtracted in the same way. Examples of addition and subtraction are shown below

$$A = \begin{bmatrix} 0 & 4 \\ 2 & 5 \\ 7 & 1 \end{bmatrix} \quad B = \begin{bmatrix} -1 & 3 \\ 3 & -2 \\ 5 & 0 \end{bmatrix} \quad A + B = \begin{bmatrix} -1 & 7 \\ 5 & 3 \\ 12 & 1 \end{bmatrix} \quad A - B = \begin{bmatrix} 1 & 1 \\ -1 & 7 \\ 2 & 1 \end{bmatrix}$$

6. Addition and subtraction of Matrices

A row vector $\mathbf{b}'$ and a column vector **a** of the same length may be multiplied by forming the sum of the products of the corresponding elements, as follows:

$$\mathbf{b}'\mathbf{a} = [2 \quad 4 \quad 1 \quad 3]\begin{bmatrix}1\\3\\5\\2\end{bmatrix} = 2(1) + 4(3) + 1(5) + 3(2)$$

The result of multiplying a row vector by a same-length column vector is a scalar. However, the result of multiplying a column vector by a row vector is entirely different. In multiplying vectors and matrices the order of multiplication matters (see below).

To multiply a matrix **A** by the matrix **B** one multiplies each of the row vectors of **A** in turn by each of the column vectors of **A**. Each of the vector multiplications yields a single number that becomes the element (row# of **A**, column# of **B**) of the product matrix **AB**. The result matrix **AB** has as many rows as the first matrix (**A**) and as many columns as the second matrix (**B**). For multiplication to be possible the rows of the first matrix (**A**) and the columns of the second matrix (**B**) must be of equal length. Thus, the second dimension (number of columns) of **A** must be equal to the first dimension (number of rows) of **B**, in which case the matrices are said to be conformable for multiplication. For example one can multiply a 3 × 2 matrix by a 2 × 5 matrix (the product being of dimensions 3 × 5), or a 1 × 5 vector by a 5 × 1 vector (the product being a 1×1 scalar), or a 5×1 column vector by a 1×5 row vector (the result being a 5 × 5 matrix). In each case for the matrices to be conformable the middle numbers have to be equal. The dimensions of the product are given by the outside dimensions.

$$\mathbf{A} = \begin{bmatrix}0 & 4\\2 & 5\\7 & 1\end{bmatrix} \quad \mathbf{B} = \begin{bmatrix}3 & 4\\2 & 5\end{bmatrix} \quad \mathbf{AB} = \begin{bmatrix}(0 \times 3) + (4 \times 2) & (0 \times 4) + (4 \times 5)\\(2 \times 3) + (5 \times 2) & (2 \times 4) + (5 \times 5)\\(7 \times 3) + (1 \times 2) & (7 \times 4) + (1 \times 5)\end{bmatrix} = \begin{bmatrix}8 & 20\\16 & 33\\23 & 33\end{bmatrix}$$

These principles generalize to longer series of matrix multiplications. If **W**, **X**, **Y**, and **Z** are, respectively, of dimensions 4×2, 2×3, 3×7, and 7×5, then the multiplication **WXYZ** can be carried out and the result has dimensions 4 × 5 (given by outermost dimensions in the series). The principle holds for vectors, viewed as n × 1 or 1 × n matrices. Thus the product of a 1 × 4 row vector by a 4 × 1 column vector is a 1 × 1 scalar; the product of a 1 × 4 column vector by a 4 × 1 row vector is a 4 × 4 matrix. When one wants to specify the order of multiplication of the product **AB** one can say that **A** premultiplies **B** or that **B** postmultiplies **A**.

7. Special cases of matrix multiplication

1. Pre- or postmultiplying a matrix **A** by a null matrix **0** yields a null matrix (i.e., a null matrix acts like zero in ordinary arithmetic)
2. Pre- or postmultiplying a matrix **A** by an identity matrix **I** leaves **A** unchanged (i.e., an identity matrix acts like a 1 in ordinary arithmetic)

3. Premultiplying a matrix **A** by a diagonal matrix **D** rescales the rows of **A** by the corresponding elements of **D**; postmultiplying **A** by **D** rescales the columns of **A** by the corresponding elements of **D**.
4. Pre- or postmultiplying a matrix by its transpose can always be done and yields a symmetric matrix: given a matrix **X**, $\mathbf{X'X}$ and $\mathbf{XX'}$ always exist.

8. Multiplying a vector or matrix by a scalar

To multiply by a matrix or vector by a scalar, multiply each element of the matrix or vector by the scalar. In a series of matrix multiplications the position of a scalar does not matter and can be changed as desired, e.g. if k is a scalar then $\mathbf{A}k\mathbf{I}\mathbf{A'} = k\mathbf{A}\mathbf{I}\mathbf{A'} = \mathbf{A}\mathbf{I}\mathbf{A'}k$. One can factor out a scalar that is a common factor of every matrix element.

2. Systems of Equations and Matrix Inverse

1. Matrix representation of systems of equations

Matrices are especially useful to represent systems of equations. For example, the system

$$\begin{aligned} 2b_1 + 4b_2 &= 20 \\ 3b_1 + \ \ b_2 &= 10 \end{aligned}$$

can be written as

$$\begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 20 \\ 10 \end{bmatrix} \qquad \text{or} \qquad \mathbf{Ab} = \mathbf{c}$$

where

$$\mathbf{A} = \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 20 \\ 10 \end{bmatrix}.$$

Note how compact the matrix representation is: the very same expression **Ab** = **c** can represent two equations with two unknown as well as 1000 equations with 1000 unknowns. To solve a system of equations requires the inverse of a matrix.

2. Inverse of a matrix

The matrix inverse is the equivalent of a reciprocal; multiplying by the inverse is the matrix equivalent of dividing. The inverse of a matrix **A**, denoted $\mathbf{A^{-1}}$, is a matrix such that $\mathbf{A^{-1}A}$ or $\mathbf{AA^{-1}}$equals **I**, an identity matrix. An example is

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

One can verify that $\mathbf{A^{-1}A = I}$. Only square matrices can have inverses, but not all of them do. If a matrix has some rows or columns that are linearly predictable from others, it does not have an inverse. A matrix that has no inverse is called *singular*. The inverse of a matrix allows solving systems of equations. Given the system of equations $\mathbf{Ab = c}$, if $\mathbf{A}$ has an inverse one can solve for $\mathbf{b}$ by premultiplying both sides of the equation by $\mathbf{A^{-1}}$ like this:

$$\mathbf{Ab = c}$$
$$\mathbf{A^{-1}Ab = A^{-1}c}$$
$$\mathbf{Ib = A^{-1}c}$$
$$\mathbf{b = A^{-1}c}$$

For example the system of equations $\mathbf{Ab = c}$ above can be solved for $\mathbf{b}$ by forming the product $\mathbf{A^{-1}c}$, i.e.

$$\begin{bmatrix} -.1 & .4 \\ .3 & -.2 \end{bmatrix} \begin{bmatrix} 20 \\ 10 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

So the solution is $b_1 = 2$ and $b_2 = 4$.

Some useful properties of inverses are:

1. $(\mathbf{A}')^{-1} = (\mathbf{A^{-1}})'$, i.e. the inverse of the transpose is the transpose of the inverse
2. $\mathbf{(A^{-1})^{-1} = A}$, i.e. taking the inverse of an inverse yields the original matrix
3. The inverse of a symmetric matrix is also symmetric

3. Calculating inverses

Calculating the inverse of a matrix tends to be a large computational task so we usually let a computer do it. However, for a few special cases matrix inversion does not require extensive computations:

1. The inverse of an identity matrix is itself, i.e. $\mathbf{I^{-1} = I}$
2. To invert a diagonal matrix, simply replace each diagonal element by its reciprocal
3. The inverse of a 2 by 2 matrix can be obtained as follows:

For the matrix:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

The inverse is

$$1/(ad - bc) \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

In words, to obtain the inverse interchange the diagonal elements *a* and *d*, change the sign of the two off-diagonal elements *b* and *c*, and multiply the result by the scalar $1/(ad - bc)$.

4. Inverse or transpose of a matrix product

Two important properties of a product of matrices are

1. $(\mathbf{ABCD})' = \mathbf{D}'\mathbf{C}'\mathbf{B}'\mathbf{A}'$, i.e. the transpose of a product of matrices is equal to the product of the transposes of the matrices, taken in reverse order
2. $(\mathbf{ABCD})^{-1} = \mathbf{D}^{-1}\mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$, i.e. the inverse of a product of matrices is equal to the product of the inverses of the matrices, taken in reverse order

For the second property to hold all the matrices have to be square, of the same order, and non-singular (otherwise multiplication would not be possible and/or the necessary inverses would not exist).

3. Random Vectors

A random vector is a vector containing elements that are random variables. For example, in the simple regression model one can define the n × 1 vector $\boldsymbol{\varepsilon}$ containing the errors for each observation, so that $\boldsymbol{\varepsilon}' = [\varepsilon_1 \quad \varepsilon_2 \quad \cdots \quad \varepsilon_n]$.

1. Expectation of a random vector

The expectation of a random vector is the vector of expectations of the random variables constituting the elements. For example in the simple regression model

$$\mathrm{E}\{\boldsymbol{\varepsilon}\} = [\mathrm{E}\{\varepsilon_1\} \quad \mathrm{E}\{\varepsilon_2\} \quad \cdots \quad \mathrm{E}\{\varepsilon_n\}]$$

The regression model assumes that the expectation of the distribution of the error for each observation is zero. This is the same as assuming $\mathrm{E}\{\boldsymbol{\varepsilon}\} = 0$, a null vector.

2. Variance-covariance matrix of a random vector

The variance-covariance matrix or covariance matrix of a random vector $\boldsymbol{\varepsilon}$ with variance denoted $\boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\}$ (with bold-face sigma) and contains the variances of the elements on the diagonal and the covariances of the elements off the diagonal. For example the covariance matrix of an n×1 random vector $\boldsymbol{\varepsilon}$ is the n × n matrix

$$\boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\} = \begin{bmatrix} \sigma^2\{\varepsilon_1\} & \sigma\{\varepsilon_1,\varepsilon_2\} & \cdots & \sigma\{\varepsilon_1,\varepsilon_n\} \\ \sigma\{\varepsilon_2,\varepsilon_1\} & \sigma^2\{\varepsilon_2\} & \cdots & \sigma\{\varepsilon_2,\varepsilon_n\} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma\{\varepsilon_n,\varepsilon_1\} & \sigma\{\varepsilon_n,\varepsilon_2\} & \cdots & \sigma^2\{\varepsilon_n\} \end{bmatrix}$$

The covariance matrix is symmetrical.

As an example the simple regression model assumes that the errors are distributed with constant variance and are uncorrelated (i.e., the covariances of the errors are zero). These assumptions correspond to a covariance matrix that has the same variance σ^2 on the diagonal and off-diagonal elements zero:

$$\boldsymbol{\sigma^2\{\varepsilon\}} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

These assumptions can be represented compactly as

$$\boldsymbol{\sigma^2\{\varepsilon\}} = \sigma^2 \mathbf{I}\,.$$

4. Matrix Representation of the Linear Regression Model

1. Simple linear regression

So far we have represented the simple linear regression model as a generic equation

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad\qquad i = 1,\ldots,n$$

that is supposed to hold for each observation i = 1, ..., n. If one wanted to write the model corresponding to each observation in the data set one would have to write

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

as many time as there are cases. Matrices provide a more compact representation of the model. One defines vectors and matrices **Y**, **X**, **β**, **ε** such that

$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_3 \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The matrices **Y**, **X**, and **ε** have n rows corresponding to the n cases in the data set. The regression model for the entire data set (i.e. the equivalent of the n separate equations above) can then be written

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

The intercept β_0 is treated as the coefficient of the constant term, a variable that has the same value 1 for all observations.

2. Multiple linear regression

The multiple linear regression model can be written as the generic equation

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad i = 1,\ldots,n$$

where there are now $p - 1$ independent variables x_1 to x_{p-1} plus the constant term β_0, for a total of p variables (including the constant term) on the right hand side of the equation. (The reason for setting the index of the last independent variable to $p - 1$ is that with this convention the total number of independent variables, including the constant term, becomes $p - 1 + 1 = p$).

Defining **y** and **ε** as before, and

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p-1} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

the regression model for the entire data set can be written as

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

which is precisely the same as for the simple linear regression model. The only differences between simple and multiple regression models are the dimensions of **X** (now n×p) and of **β** (now p×1). As before, there are two sets of assumptions on the errors **ε**:

1. The weaker set assumes that $E\{\boldsymbol{\varepsilon}\} = 0$ (mean of errors is zero) and that the errors are uncorrelated and identically distributed with covariance matrix $\boldsymbol{\sigma^2}\{\boldsymbol{\varepsilon}\} = \sigma^2\mathbf{I}$.
2. The stronger set assumes in addition that the errors are normally distributed.

3. OLS estimation of the regression coefficients

The OLS estimator of β are the values of the regression parameters that minimize

$$Q = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{i,1} - b_2 X_{i,2} - \ldots - b_{p-1} X_{i,p-1})^2$$

The normal equations (from SLR notes)

$$\sum Y_i = nb_0 + b_1 \sum X_i$$
$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$$

In matrix form are (see pg. 200)

$$\mathbf{X'Xb = X'Y}$$

It can be shown with calculus (pg. 201) that the OLS estimator b of β is the vector $\mathbf{b'} = [b_0 \quad b_1 \quad \cdots \quad b_{p-1}]$ is the solution of the normal equations.

The solution **b** of the system is obtained by premultiplying both sides by the inverse of $\mathbf{X'X}$ in the following steps

$$\begin{aligned} \mathbf{X'Xb} &= \mathbf{X'Y} \\ (\mathbf{X'X})^{-1}\mathbf{X'Xb} &= (\mathbf{X'X})^{-1}\mathbf{X'Y} \\ \mathbf{Ib} &= (\mathbf{X'X})^{-1}\mathbf{X'Y} \end{aligned}$$

Therefore,

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

4. Predicted values $\hat{y}_i$, residuals e_i and the Hat matrix H

1. Predicted values $\hat{y}_i$

The predicted (fitted, estimated) value of $\hat{Y}_i$ of Y for observation i is

$$\hat{Y}_i = b_0 + b_1X_{i,1} + b_2X_{i,2} + \ldots + b_{p-1}X_{i,p-1} \qquad i = 1,\ldots,n$$

The n×1 vector $\hat{\mathbf{Y}}$ containing the predicted values, $\hat{\mathbf{Y}}' = [\hat{Y}_1 \quad \hat{Y}_2 \quad \cdots \quad \hat{Y}_n]$ is

$$\hat{\mathbf{Y}} = \mathbf{Xb}$$

2. The Hat matrix H

We know that $\hat{\mathbf{Y}} = \mathbf{Xb}$ and $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ so

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

or

$$\hat{\mathbf{Y}} = \mathbf{HY}$$

where

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

The matrix **H** is called the *hat matrix*.

3. Residuals e_i

The residual e_i for observation i is estimated as

$$e_i = Y_i - \hat{Y}_i$$

same as in simple regression. The n×1 vector **e** containing the residuals, $\mathbf{e}' = [e_1 \quad e_2 \quad \cdots \quad e_n]$, is given by

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = \mathbf{(I - H)Y}$$

Therefore,

$$\mathbf{e} = \mathbf{(I - H)Y}$$