Soci709 (formerly 209) – Module 4
Matrix Representation of Regression Model

François Nielsen

March 22, 2006

1 Introduction

Module 4 covers the following topics

- properties of, and operations on, matrices and vectors
- matrix representation of the linear regression model, both simple and multiple
- derivation of the most important aspects of the regression model, including estimation of the parameters and ANOVA decomposition
- derivation of the covariance matrices and standard errors of the estimators needed for statistical inference (calculation of confidence intervals and hypothesis tests)

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 in Module 5. This material is in part foundational: matrices are almost indispensable for advanced methods beyond regression analysis. Readings for Module 4 are ALSM5e ???; ALSM4e Chapter 5 and part of Chapter 6 (pp. 225-236).

2 Matrices

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.

2.1 Matrices

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

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

Dimensions of a matrix are specified as the *number of rows* followed by the *number of columns*: $\mathbf{A}$ is 4×2, $\mathbf{B}$ and $\mathbf{D}$ are 3×3, and $\mathbf{C}$ is 3×2, . An element of a matrix $\mathbf{C}$ 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. $\mathbf{B}$ and $\mathbf{D}$ are square matrices. $\mathbf{B}$ and $\mathbf{D}$ 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 $\mathbf{D}$ 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 $\mathbf{I}$ and called an *identity matrix.* A matrix of all zeroes is called a *null matrix.* Matrices are usually designated by bold capital letters. Examples of an identity matrix and a null matrix are

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

2.2 Transpose of a matrix

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

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

(I find it useful to visualize the transpose operation with a special hand movement which I will demonstrate in class.) An alternative notation found for the transpose $\mathbf{A}'$ is $\mathbf{A}^T$. 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 $\mathbf{A}' = \mathbf{A}$.

2.3 Vector 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. $\mathbf{a}$. *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.

$$\mathbf{a} = \begin{bmatrix} 1.2 \\ 3 \\ 1.7 \end{bmatrix} \qquad \mathbf{b}' = \begin{bmatrix} 1.1 & .5 & 2 \end{bmatrix} \qquad c = 22.6 \qquad \alpha = .05$$

2.4 Equality of matrices

Two matrices $\mathbf{A}$ and $\mathbf{B}$ are equal if they have the same dimensions and all corresponding elements are equal.

2.5 Addition & 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.

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

2.6 Multiplication of vectors and matrices

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

$$\mathbf{a}'\mathbf{b} = \begin{bmatrix} 2 & 4 & 1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 5 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \times 1 = 2 \\ 4 \times 3 = 12 \\ 1 \times 5 = 5 \\ 3 \times 2 = 6 \\ \mathbf{a}'\mathbf{b} = 2 + 12 + 5 + 6 = 25 \end{bmatrix}$$

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 $\mathbf{A}$ by the matrix $\mathbf{B}$ one multiplies each of the row vectors of $\mathbf{A}$ in turn by each of the column vectors of $\mathbf{A}$. Each of the vector multiplications yields a single number that becomes the element (row# of $\mathbf{A}$, column# of $\mathbf{B}$) of the product matrix $\mathbf{AB}$. The result matrix $\mathbf{AB}$ 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.

$$\mathbf{A} = \begin{bmatrix} 0 & 4 \\ 2 & 5 \\ 7 & 1 \end{bmatrix} \mathbf{B} = \begin{bmatrix} 3 & 4 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} \mathbf{AB}_{11} = (0 \times 3) + (4 \times 2) = 8 \\ \mathbf{AB}_{12} = (0 \times 4) + (4 \times 5) = 20 \\ \mathbf{AB}_{21} = (2 \times 3) + (5 \times 2) = 16 \\ \mathbf{AB}_{22} = (2 \times 4) + (5 \times 5) = 33 \\ \mathbf{AB}_{31} = (7 \times 3) + (1 \times 2) = 23 \\ \mathbf{AB}_{32} = (7 \times 4) + (1 \times 5) = 33 \end{bmatrix} \tag{1}$$

$$\mathbf{AB} = \begin{bmatrix} 8 & 20 \\ 16 & 33 \\ 23 & 33 \end{bmatrix} \tag{2}$$

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 $n \times 1$ or $1 \times n$ 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**.

2.7 Special cases of matrix multiplication

Work through your own examples to derive the following rules.

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**, $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$ always exist.

2.8 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.

3 Systems of Equations & Matrix Inverse

3.1 Matrix representation of systems of equations

Matrices are especially useful to represent systems of equations. For example defining $\mathbf{b}' = \begin{bmatrix} b_1 & b_2 \end{bmatrix}$ (two unknown quantities), $\mathbf{c}' = \begin{bmatrix} 20 & 10 \end{bmatrix}$ and

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

one can represent the system of equations

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

as

$$\begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 20 \\ 10 \end{bmatrix}$$

or

$$\mathbf{A}\mathbf{b} = \mathbf{c}$$

Note how compact the matrix representation is: the very same expression $\mathbf{A}\mathbf{b} = \mathbf{c}$ 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 (see next).

3.2 Inverse of a matrix

In ordinary algebra the inverse of a number x is its reciprocal $1/x$ or x^{-1}. Multiplying by the reciprocal is equivalent to dividing, so $a(1/x) = a/x$. For matrices there is no direct equivalent of division but the matrix inverse is the equivalent of the reciprocal; multiplying by the inverse is the matrix equivalent of dividing. The inverse of a matrix $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, is a matrix such that $\mathbf{A}^{-1}\mathbf{A}$ or $\mathbf{A}\mathbf{A}^{-1}$ equals $\mathbf{I}$, an identity matrix (the same way that $(1/a)a = 1$ for scalars). An example is

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

One can verify that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. 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 (see "Linear dependence & rank" below). A matrix that has no inverse is called *singular*. The inverse of a matrix allows solving systems of equations. Given the system of equations $\mathbf{A}\mathbf{b} = \mathbf{c}$, 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:

$$\begin{aligned} \mathbf{A}\mathbf{b} &= \mathbf{c} \\ \mathbf{A}^{-1}\mathbf{A}\mathbf{b} &= \mathbf{A}^{-1}\mathbf{c} \\ \mathbf{I}\mathbf{b} &= \mathbf{A}^{-1}\mathbf{c} \\ \mathbf{b} &= \mathbf{A}^{-1}\mathbf{c} \end{aligned}$$

For example the system of equations $\mathbf{A}\mathbf{b} = \mathbf{c}$ above can be solved for b by forming the product $\mathbf{A}^{-1}\mathbf{c}$, 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 $b_2 = 4$.

3.3 Calculating inverses

Loehlin (2004, p. 242-243) writes: "Obtaining the inverse of a matrix tends in general to be a large computational task. Let the computer do it. You can always check to see whether the results it has given you is correct by carrying out the multiplication $\mathbf{A}\mathbf{A}^{-1}$, which should equal $\mathbf{I}$ within rounding error." 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. $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$, i.e. taking the inverse of an inverse yields the original matrix
3. The inverse of a symmetric matrix is also symmetric

In a few special cases matrix inversion does not require extensive computations:

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

If the original matrix is

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

the inverse is obtained as

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

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 $1/(ad-bc)$. One can verify this procedure by calculating the inverse of the matrix $\mathbf{A}$ introduced above.

3.4 Inverse or transpose of a matrix product

Two important properties of a product of matrices are

1. $(\mathbf{ABCD})' = \mathbf{D'C'B'A'}$, 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. $(\mathbf{ABCD})^{-1} = \mathbf{D^{-1}C^{-1}B^{-1}A^{-1}}$, 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).

3.5 Eigenvalues & eigenvectors

This will be added later.

3.6 Linear dependence & rank

Consider the 5×4 matrix

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 1 & 0 \\ 1 & 4 & 1 & 0 \\ 1 & 3 & 1 & 0 \\ 1 & 7 & 0 & 1 \\ 1 & 4 & 0 & 1 \end{bmatrix}$$

$\mathbf{X}$ can be viewed as composed of 4 columns vectors $\begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_4 \end{bmatrix}$. The columns of a $r \times c$ matrix are said to be *linearly dependent* when c scalars $\lambda_1, \lambda_2, \cdots, \lambda_c$ *not all equal to 0* can be found such that

$$\lambda_1\mathbf{c}_1 + \lambda_2\mathbf{c}_2 + \cdots + \lambda_c\mathbf{c}_c = \mathbf{0}$$

If this relation only holds when all the λ's are zero then the columns are *linearly dependent*. The columns of the matrix $\mathbf{X}$ are not independent.

Q - Find 4 λ's not all zero such that the relation above holds.

A - One possible answer is $1, 0, -1, -1$.

The *rank* of a matrix is the maximum number of linearly independent columns in the matrix.

Q - What is the rank of $\mathbf{X}$?

An important property of matrices is that if the columns are independent, so are the rows.

4 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 $n \times 1$ vector $\boldsymbol{\epsilon}$ containing the errors for each observation, so that $\boldsymbol{\epsilon}' = \begin{bmatrix} \epsilon_1 & \epsilon_2 & \cdots & \epsilon_n \end{bmatrix}$. (Note that the vector is given by its transpose to save vertical space in text.)

4.1 Expectation of a random vector

The *expectation* of a random vector is the the vector of expectations of the random variables constituting the elements. For example in the simple regression model $\mathcal{E}\{\boldsymbol{\epsilon}\} = \begin{bmatrix} \mathcal{E}\{\epsilon_1\} & \mathcal{E}\{\epsilon_2\} \cdots \mathcal{E}\{\epsilon_n\} \end{bmatrix}$ The regression model assumes that the expectation of the distributon of the error for each observation is zero. This is the same as assuming $\mathcal{E}\{\boldsymbol{\epsilon}\} = \mathbf{0}$, a null vector. (For definition of *expectation* or *expected value* of a random variable see ALSM5e Appendix A, Equations A.11 and A.12.)

4.2 Variance-covariance matrix of a random vector

The *variance-covariance matrix* or *covariance matrix* of a random vector $\boldsymbol{\epsilon}$ with expectation is denoted $\boldsymbol{\sigma}^2\{\boldsymbol{\epsilon}\}$ (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 $n \times 1$ random vector $\boldsymbol{\epsilon}$ is the $n \times n$ matrix

$$\boldsymbol{\sigma}^2\{\boldsymbol{\epsilon}\} = \begin{bmatrix} \sigma^2\{\epsilon_1\} & \sigma\{\epsilon_1, \epsilon_2\} & \cdots & \sigma\{\epsilon_1, \epsilon_n\} \\ \sigma\{\epsilon_2, \epsilon_1\} & \sigma^2\{\epsilon_2\} & \cdots & \sigma\{\epsilon_2, \epsilon_n\} \\ \vdots & \vdots & & \vdots \\ \sigma\{\epsilon_n, \epsilon_1\} & \sigma\{\epsilon_n, \epsilon_2\} & \cdots & \sigma^2\{\epsilon_n\} \end{bmatrix}$$

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:

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

These assumptions can be represented compactly as

$$\boldsymbol{\sigma}^2\{\boldsymbol{\epsilon}\} = \mathcal{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\} = \sigma^2\mathbf{I}$$

4.3 Linear transformations of random vectors

To be added later.

5 Matrix Representation of Linear Regression Model

5.1 Simple Linear Regression

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

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, ..., n$$

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

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ &\cdots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned}$$

as many time as there are cases.

Matrices provide a more compact representation of the model. One defines vectors and matrices $\mathbf{y}, \mathbf{X}, \beta$, and ϵ such that

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The matrices $\mathbf{y}, \mathbf{X}$ and $\boldsymbol{\epsilon}$ 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

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

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

5.2 Multiple Linear Regression

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

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i \quad i = 1, ..., n$$

where there are now $p-1$ independent variables x_1 to x_{p-1} plus the constant term x_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 $p-1+1=p$, a nice simple symbol.) Defining $\mathbf{y}$ and ϵ as before, and

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

the regression model for the entire data set can be written

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

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 $\mathbf{X}$ (now $n \times p$) and of β (now $p \times 1$). As before there are two sets of assumptions on the errors ϵ:

- the weaker set assumes that $\mathcal{E}\{\epsilon\} = 0$ (mean of errors is zero) and that the errors are uncorrelated and identiccally distributed with covariance matrix $\boldsymbol{\sigma}^2\{\epsilon\} = \mathcal{E}\{\epsilon\epsilon'\} = \sigma^2\mathbf{I}$
- the stronger set assumes in addition that the errors are normally distributed

It follows from either set of assumptions that the random vector $\mathbf{y}$ has expectation

$$\mathcal{E}\{\mathbf{y}\} = \mathbf{Xb}$$

and the variance-covariance matrix of $\mathbf{y}$ is the same as that of ϵ so that

$$\boldsymbol{\sigma}^2\{\mathbf{y}\} = \mathcal{E}\{(\mathbf{y} - \mathbf{Xb})(\mathbf{y} - \mathbf{Xb})'\} = \mathcal{E}\{\epsilon\epsilon'\} = \sigma^2\mathbf{I}$$

5.3 OLS Estimation of the Regression Coefficients

The OLS estimator of $\boldsymbol{\beta}$ are the values of the regression parameters that minimize

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

It can be shown with calculus (ALSM5e ??; ALSM4e pp. 201–202) that the OLS estimator $\mathbf{b}$ of $\boldsymbol{\beta}$ is the vector $\mathbf{b}' = \begin{bmatrix} b_0 & b_1 & \cdots & b_{p-1} \end{bmatrix}$ that is the solution of the *normal equations*

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$$

Q – What is the dimension of $\mathbf{X}'\mathbf{X}$? Of $\mathbf{X}'\mathbf{y}$? What do these matrices contain?

(The normal equations are obtained by setting the partial derivatives of Q with respect to the regression coefficients equal to zero and rearranging the terms.) Note that the normal equations constitute a system of p equations with p unknowns. The solution $\mathbf{b}$ of the system is obtained by pre-multiplying both sides by the inverse of $\mathbf{X}'\mathbf{X}$ in the following steps

$$\begin{aligned} \mathbf{X}'\mathbf{X}\mathbf{b} &= \mathbf{X}'\mathbf{y} \\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ \mathbf{I}\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ \mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \end{aligned}$$

The relation

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

is the fundamental formula of OLS.

Q - What are the dimensions of $\mathbf{b}$? Of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$? The Gauss-Markov theorem says that $\mathbf{b}$ is BLUE; can you tell from the formula for b why there is an L in BLUE?

5.4 Fitted Values $\hat{y}_i$, Residuals e_i, & the Hat Matrix H

5.4.1 Fitted Values (aka Predictors aka Estimates)

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

$$\hat{y}_i = b_0 + b_1 X_{i,1} + ... + b_{p-1} X_{i,p-1} \quad i = 1, \cdots, n$$

The $n \times 1$ vector $\hat{\mathbf{Y}}$ containing the fitted values, $\hat{\mathbf{Y}}' = \begin{bmatrix} \hat{Y}_1 & \hat{Y}_2 & \cdots & \hat{Y}_n \end{bmatrix}$ is

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$$

Q - Why does Xb represent the fitted (predicted, estimated) values of Y?

5.4.2 The Hat Matrix H

Replacing $\mathbf{b}$ in $\mathbf{X}\mathbf{b}$ by its value in terms of the data yields

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

or equivalently

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

where

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

The matrix $\mathbf{H}$ is called the *hat matrix.* (Q - Why would $\mathbf{H}$ be called the "hat" matrix?) H has the following properties

- $\mathbf{H}$ is square of dimension $n \times n$ and involves only $\mathbf{X}$ (the observations on the independent variables which are assumed to be fixed constants)
- Thus $\hat{\mathbf{y}}$ is a linear combination of the observations $\mathbf{y}$. $\mathbf{H}$ is very important in outlier diagnostics, as discussed in Module 10. (To anticipate, the diagonal elements h_{ii} of $\mathbf{H}$ measure the *leverage* of observation i, measured as the extent to which the value y_i of the dependent variable for observation i affects its *own* fitted value $\hat{y}_i$.)
- $\mathbf{H}$ is *idempotent,* i.e., $\mathbf{HH} = \mathbf{H}$. (There is an underlying geometric interpretation, which is that an idempotent matrix represents a *projection* of a point in space on a subspace. Idempotency reflects the fact that once a point is projected on the subspace, the point stays there if projected again.)

Q - Can you show why $\mathbf{H}$ is idempotent? (Hint: calculate $\mathbf{HH}$, replacing $\mathbf{H}$ by its value in terms of $\mathbf{X}$.

5.4.3 Residuals

The residual e_i for observation i is estimated as

$$e_i = y_i - \hat{y}_i$$

same as in simple regression. The $n \times 1$ vector $\mathbf{e}$ containing the residuals, $\mathbf{e}' = \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix}$, is given by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

One can derive

$$\begin{aligned} \mathbf{e} &= \mathbf{y} - \hat{\mathbf{y}} \\ \mathbf{e} &= \mathbf{y} - \mathbf{H}\mathbf{y} \\ \mathbf{e} &= (\mathbf{I} - \mathbf{H})\mathbf{y} \end{aligned}$$

Q - How does the last expression follow?

Like $\mathbf{H}$, the matrix $\mathbf{I} - \mathbf{H}$ is $n \times n$, symmetric and idempotent. Furthermore $\mathbf{H}(\mathbf{I} - \mathbf{H}) = 0$. (Q–Can you see why?)

5.5 Analysis of Variance (ANOVA)

The ANOVA decomposition of the variation in $\mathbf{y}$ is exactly the same as for the simple regression model, except for the degrees of freedom corresponding to the number of independent variables ($n - 2$ in simple regression becomes $n - p$ in multiple regression).

Table 1: ANOVA Table

Source	Sum of Squares	df	Mean Squares
Regression	$SSR = \sum(\hat{y}_i - \bar{y})^2$	$p-1$	$MSR = SSR/(p-1)$
Error	$SSE = \sum(y_i - \hat{y}_i)^2$	$n-p$	$MSE = SSE/(n-p)$
Total	$SSTO = \sum(y_i - \bar{y})^2$	$n-1$	$Var(Y) = SSTO/(n-1)$

5.5.1 Partitioning of Sums of Squares and Degrees of Freedom

The ANOVA decomposition starts with the identity

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

which expresses, for each observation, the deviation of y_i from the mean $\bar{y}$ as the sum of the deviation of the fitted value from the mean plus the deviation of y_i from the fitted value. (The identity is tautological, as one can see by removing the parentheses on the right-hand side; then $\hat{y}_i$ and $-\hat{y}_i$ cancel out and both sides of the equation are the same.)

One can take the sums of the squared deviations on both sides of the identity to obtain the decomposition

$$SSTO = SSR + SSE$$

The fact that this decomposition holds is not obvious and must be demonstrated (see ALSM5e ???; ALSM4e ???). The degrees of freedom (*df*) associated with the sums of squares are

- for $SSTO$ $df = (n-1)$, same as in simple regression (1 *df* lost estimating $\bar{y}$)
- for SSR $df = (p-1)$, number of independent variables, not including the constant
- for SSE $df = (n-p)$, where p is total number of variables, including the constant (p *df* lost in estimating regression function to calculate $\hat{y}_i$ and e_i)

The *mean squares* MSR, MSE, and S^2 (variance of y) are the sums of squares divided by their respective *df*, same as in simple regression. See Table 5.5.1.

5.5.2 (Optional) Sums of Squares as Quadratic Forms

Sums of squares in matrix form are shown in Table 5.5.2. Matrix notation for SSR and $SSTO$ uses an $n \times 1$ vector $\mathbf{u}$ (for *unity*) with all elements equal to 1. One can verify that $\mathbf{y'uu'y}$ is equal to $(\sum y_i)^2$ by noting that $\mathbf{y'uu'y} = (y'u)(u'y) = (\sum y_i)(\sum y_i) = (\sum y_i)^2$.

Table 5.5.2 (Collumn 3) shows that the sums of squares can all be represented in a form $\mathbf{y'Ay}$ where $\mathbf{A}$ is a symmetric matrix called a *quadratic form*. A

Table 2: ANOVA Table in Matrix Notation

SS	Matrix Notation	Quadratic Form	df
SSR	$[\mathbf{Hy} - (1/n)\mathbf{uu'y}]'[\mathbf{Hy} - (1/n)\mathbf{uu'y}]$	$\mathbf{y'}[\mathbf{H} - (1/n)\mathbf{uu'}]\mathbf{y}$	$p-1$
SSE	$\mathbf{ee'} = (\mathbf{y} - \mathbf{Hy})'(\mathbf{y} - \mathbf{Hy})$	$\mathbf{y'}(\mathbf{I} - \mathbf{H})\mathbf{y}$	$n-p$
$SSTO$	$[\mathbf{y} - (1/n)\mathbf{uu'y}]'[\mathbf{y} - (1/n)\mathbf{uu'y}]$	$\mathbf{y'}[\mathbf{I} - (1/n)\mathbf{uu'}]\mathbf{y}$	$n-1$

quadratic form is an expression of the form $\mathbf{y'Ay}$ where $\mathbf{A}$ is a symmetric matrix. Then

$$\mathbf{y'Ay} = \sum_{i=1}^{n} a_{ij} y_i y_j$$

where $a_{ij} = a_{ji}$ and $\mathbf{y'Ay}$ is 1×1 (a scalar). One can see that $\mathbf{y'Ay}$ is a second-degree polynomial involving the squares and cross products of the observations y_i. For example $5y_{12} + 6y_1y_2 + 4y_{22}$ can be represented as $\mathbf{y'Ay}$ where $\mathbf{y'} = \begin{bmatrix} y_1 & y_2 \end{bmatrix}$ and

$$\mathbf{A} = \begin{bmatrix} 5 & 3 \\ 3 & 4 \end{bmatrix}$$

$\mathbf{A}$ is called the *matrix of the quadratic form.* (In the underlying geometry $n \times n$ quadratic forms represent distances in n-dimensional space.)

5.5.3 (Optional) Sums of Squares, Quadratic Forms, df, and χ^2 Distributions

It is possible to show that

- if $\mathbf{A}$ is the matrix of an idempotent quadratic form, and $\mathbf{y}$ a vector of independent random variables each distributed ~N(0,1), then $\mathbf{y'Ay}$ is the sum of a number rank($\mathbf{A}$) of independent random variables z^2, each distributed as $\chi^2(1)$ (chi-square with 1 *df*), where rank($\mathbf{A}$) denotes the rank of $\mathbf{A}$
- if $\mathbf{A}$ is an idempotent square matrix, the rank of $\mathbf{A}$ is equal to the trace of $\mathbf{A}$, denoted tr($\mathbf{A}$); the *trace* of a square matrix is the sum of the diagonal elements of the matrix, so that tr($\mathbf{A}$) = $\sum_{i=1}^{n} a_{ii}$
- the traces of the quadratic forms corresponding to SSR, SSE, and $SSTO$ are $p-1$, $n-p$, and $n-1$, respectively, the same as their *df*!
- the sum of k independent random variables, each distributed as $\chi^2(1)$ (chi-square with 1 *df*) is distributed as $\chi^2(k)$ (chi-square with k *df*)
- thus if $\mathbf{y}$ is a vectors of independent random variables each distributed ~N(0,1), then SSR, SSE, and $SSTO$ are distributed as $\chi^2(p-1)$, $\chi^2(n-p)$, and $\chi^2(n-1)$, respectively

- the ratio of two sums of squares, each divided by its df, is distributed as $F(df_1, df_2)$ where $df1$ and $df2$ are the df of the numerator and the denominator, respectively; thus for example $F* = MSR/MSE = (SSR/(p-1))/(SSE/(n-p))$ is distrinuted as $F(p-1, n-p)$

5.6 Sampling Distributions of Estimators

5.6.1 Need for the Covariance Matrix of Estimators

As we saw in the context of simple regression, various aspects of the regression model that are substantively interesting (such as regression coefficients, the fitted values $\hat{y}_i$, the residuals e_i) are estimated from (i.e., functions of) the values of the dependent variables ($\mathbf{y}$), which are functions of random errors ($\boldsymbol{\epsilon}$). Thus these estimates are themselves random variables. In order to carry out statistical inference (to do hypothesis tests or calculate confidence intervals) one needs the *standard error* of the estimate. The standard error is the standard deviation of the sampling distribution of the estimate. The following considerations help in understanding how these standard errors of estimates are obtained for different kinds of estimates such as $\mathbf{b}$, $\hat{y}_h$ (predicted value of y for a combination $\mathbf{X}_h$ of the independent variables), $\mathbf{e}$, and others.

- the standard error of estimate is the square root of the variance of the estimate, which is the element in the diagonal of the covariance matrix of the estimate(s)
- all the OLS estimates (such as $\mathbf{b}$, $\hat{y}_h$, $\mathbf{e}$, and others) are *linear functions of the observed* $\mathbf{y}$; thus the estimate(s) are obtained as $\mathbf{Ay}$ where $\mathbf{A}$ in a constant matrix expressing the estimate(s) as function(s) of $\mathbf{y}$
- thus, the covariance matrix of the estimate(s) $\mathbf{Ay}$ can be obtained by applying a theorem that states that $\boldsymbol{\sigma}^2\{\mathbf{Ay}\} = \mathbf{A}\boldsymbol{\sigma}^2\{\mathbf{y}\}\mathbf{A}'$ where $\mathbf{y}$ is a random vector

5.6.2 Covariance Matrix of b

One can derive the theoretical covariance matrix $\boldsymbol{\sigma}^2\{\mathbf{b}\}$ of the regression coefficients $\mathbf{b}$ with the following steps

$$\begin{aligned}
\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{Ay} \\
\boldsymbol{\sigma}^2\{\mathbf{b}\} &= \boldsymbol{\sigma}^2\{\mathbf{Ay}\} \\
&= \mathbf{A}\boldsymbol{\sigma}^2\{\mathbf{y}\}\mathbf{A}' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

so that

$$\boldsymbol{\sigma}^2\{\mathbf{b}\} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Then the *estimated* covariance matrix of $\mathbf{b}$ is obtained by replacing replacing the unknown variance of the errors σ^2 by its estimated value MSE, as

$$\mathbf{s}^2\{\mathbf{b}\} = MSE(\mathbf{X}'\mathbf{X})^{-1}$$

Q–What are the dimensions of this matrix? A–This is a $p \times p$ matrix.

The diagonal contains the estimated variances of the elements of $\mathbf{b}$, so that $s^2\{b_0\}$ is in position (1,1), $s^2\{b_1\}$ in position (2,2), etc. For statistical inference the standard error of a regression coefficient is estimated as the square root of the corresponding diagonal element of this matrix. The off-diagonal elements correspond to the covariances among elements of $\mathbf{b}$.

5.6.3 Variance of Fitted Value $\hat{y}_h$

A combination of specific values of the independent variables can be represented as a row vector $\mathbf{x}_h' = \begin{bmatrix} 1 & x_{h,1} & x_{h,2} & \cdots & x_{h,p-1} \end{bmatrix}$. $\mathbf{x}_h'$ may correspond to one of the observations in the data set on which the regression model is estimated, in which case it represents a row of $\mathbf{X}$, but this is not necessarily the case. $\mathbf{x}_h'$ may also represent an "as if" combination of values of the independent variables that does not exist in the data set. The mean response $\mathcal{E}\{y_h\}$ is estimated as

$$\hat{y}_h = \mathbf{x}_h'\mathbf{b}$$

Note that $\hat{y}_h = \mathbf{x}_h'\mathbf{b}$ is a linear function of $\mathbf{b}$. Furthermore $\hat{y}_h$ is a scalar, so its covariance matrix reduces to a single variance. The theoretical variance $\sigma^2\{\hat{y}_h\}$ of $\hat{y}_h$ is derived with the following steps

$$\begin{aligned} \hat{y}_h &= \mathbf{x}_h'\mathbf{b} \\ \sigma^2\{\hat{y}_h\} &= \mathbf{x}_h'\boldsymbol{\sigma}^2\{\mathbf{b}\}\mathbf{x}_h \\ &= \mathbf{x}_h'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h \\ &= \sigma^2\mathbf{x}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h \end{aligned}$$

Then the *estimated* variance of $\hat{y}_h$ is given by

$$s^2\{\hat{y}_h\} = MSE(\mathbf{x}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h)$$

and the estimated standard error of $\hat{y}_h$ by

$$s\{\hat{y}_h\} = \sqrt{MSE(\mathbf{x}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h)}$$

Q - What kind of matrix expression is $\mathbf{x}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h$? Hint: it is a q— f—.

5.6.4 Covariance Matrix of Residuals e

The theoretical covariance matrix $\boldsymbol{\sigma}^2\{\mathbf{e}\}$ of the residuals $\mathbf{e}$ is derived in the following steps

$$\begin{aligned}\mathbf{e} &= (\mathbf{I}-\mathbf{H})\mathbf{y}\\ \boldsymbol{\sigma}^2\{\mathbf{e}\} &= (\mathbf{I}-\mathbf{H})\boldsymbol{\sigma}^2\{\mathbf{y}\}(\mathbf{I}-\mathbf{H})'\\ &= \sigma^2(\mathbf{I}-\mathbf{H})\mathbf{I}(\mathbf{I}-\mathbf{H})\\ &= \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})\end{aligned}$$

so that

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

Then the estimated covariance matrix of $\mathbf{e}$ is given by

$$\mathbf{s}^2\{\mathbf{e}\} = MSE(\mathbf{I}-\mathbf{H})$$

The diagonal of this $n \times n$ matrix contains the estimated variances of the elements of $\mathbf{e}$. This matrix is important in identifying outliers (i.e., observations with abnormally large residuals) as discussed in Module 10.

5.6.5 MSE, MSR and the Distribution of F*

As discussed later in Module 5, the *screening test* for the significance of the regression as a whole is based on the test statistic

$$F* = MSR/MSE$$

As mentioned above in connection with quadratic forms, $F* = MSR/MSE$ is distributed as $F(p-1, n-p)$ where $(p-1)$ and $(n-p)$ are the *df* of SSR and SSE, respectively, from the ANOVA table. The expectations of MSR and MSE are discussed in ALSM5e p. ?; ALSM4e p. 229.

5.7 Matrix Notation in Practice

The following sections show the use of matrix operations to perform multiple regression with the construction industry data using programs STATA and SYSTAT.

5.7.1 STATA Example – ho4statamat.do

```
use craft, clear
* create a constant term that is 1 for each case
generate const=1
* setup matrices y and X (beware that STATA is case sensitive)
mkmat clerks, matrix(y)
```

```
mkmat const season size, matrix(X)
matrix list y
matrix list X
* calculate y'y, X'X and X'y
matrix yy=y'*y
matrix XX=X'*X
matrix Xy=X'*y
* calculate number of observations and df
matrix nobs=rowsof(X)
matrix df=nobs[1,1]-colsof(X)
* calculate b as (X'X)-1X'y
matrix b=syminv(XX)*Xy
matrix list b
* calculate SSE and MSE
matrix SSE=yy-b'*Xy
matrix MSE=SSE/df[1,1]
*calculate covariance matrix of b, call it V
matrix V=syminv(XX)*MSE
*calculate the t-ratio t* for each coefficient (parentheses are not brackets!)
display b[1,1]/sqrt(V[1,1])
display b[2,1]/sqrt(V[2,2])
display b[3,1]/sqrt(V[3,3])
* calculate the 2-sided P-value for each coefficient using the following formula
* where t* is one of the t-ratios you just calculated; copy and paste the
* value of t* from your output each time (abs() is the absolute value function)
display 2*ttail(df[1,1],abs(t*))
* decide which coefficient(s) is (are) significant at the .05 level
* calculate the hat matrix H
matrix H=X*syminv(XX)*X'
matrix list H
* calculate the trace of H (=sum of diagonal elements)
matrix tr=trace(H)
matrix list tr
* guess a general formula giving the value of the trace of H
* end of STATA commands
```

5.7.2 SYSTAT Example – ho4systatmat.txt

```
matrix
use craft
mat x = craft(;size season)
mat const = m(9,1,1)
mat x = const||x
mat y = craft(; clerks)
sho x y
mat xpxi = inv(trp(x)*x)
```

```
mat xpy = trp(x)*y
mat b = xpxi*xpy
sho xpxi xpy b
mat h = x*xpxi*trp(x)
sho h
mat i9 = i(9)
sho i9
mat sse = trp(y)*(i9 - h)*y
mat mse = sse/6
sho sse mse
mat varb = mse#xpxi
sho varb
mat se = trp(diag(varb))
sho se
mat se = sqr(se)
mat t = b/se
sho t
calc 2*(1 - tcf(1.401501,6))
calc 2*(tcf(-4.799469,6))
```