17.874 Lecture Notes
Part 2: Matrix Algebra

2. Matrix Algebra

2.1. Introduction: Design Matrices and Data Matrices

Matrices are arrays of numbers. We encounter them in statistics in at least three different ways. First, they are a convenient device for systematically organizing our study designs. Second, data sets typically take the form rectangular arrays of numbers and characters. Third, algebraic and computational manipulation of data is facilitated by matrix algebra. This is particularly important for multivariate statistics.

Over the next 2 weeks or so we will learn the basic matrix algebra used in multivariate statistics. In particular, we will learn how to construct the basic statistical concepts – means, variances, and covariances – in matrix form, and to solve estimation problems, which involve solving systems of equations.

I begin this section of the course by discussing Design Matrices and Data Matrices.

Design Matrices are arrays that describe the structure of a study. Many studies involve little actual design. The design matrix is the data matrix. Experiments, quasi-experiments, and surveys often involve some degree of complexity and care in their design. For any study, a design matrix can be very useful at the start for thinking about exactly what data needs to be collected to answer a question.

There are several components to a design matrix. The rows correspond to units of analysis and unique factors in a study; the columns are variables that one would like to measure (even if you can't). At the very least one needs to set aside columns for the X variable and the Y variable. You might also consider an assignment probability (in an experiment or survey) or variable (this might be an instrumental variable in an observational study).

For example, in the early and mid-1990s I did a series of about two dozen experiments involving campaign advertising. Early on in those studies I wrote down a simple matrix for keeping the factors in the studies straight. Units were individuals recruited to participate

in the study. The factors were that some people saw an ad from a Democrat, some people saw an ad from a Republican, and some saw no ad. In addition, some ads were negative and some were positive. I wanted to know how these factors affected various outcomes – especially, turnout and vote intention. Do candidates do better with positive or negative ads? Which candidates? Does turnout drop?

Simple Design Matrix

Source	Tone	Assignment	Turnout	Vote
D	+	.2	T_{D+}	V_{D+}
D	-	.2	T_{D-}	V_{D-}
R	+	.2	T_{R+}	V_{R+}
R	-	.2	T_{R-}	V_{R-}
0	0	.2	T_0	V_0

I found that writing down a matrix like this helped me to think about what the experiment itself could measure. I can measure the effect of negative versus positive ads on turnout by taking the following difference: $[T_{D+} + T_{R+}] - [T_{D-} + T_{R-}]$.

It also raised some interesting questions. Do I need the control group to measure all of the effects of interest?

Data Matrices implement our design matrices and organize our data. Data matrices begin with information about the units in our study – people, countries, years, etc. They then contain information about the variables of interest. It really doesn't matter how you represent data in a database, except that most databases reserve blanks and "." for missing data.

Below are hypothetical data for 10 subject for our experiments. The content of each variable is recorded so that it is most meaningful. Under the variable Source "D" means Democratic candidate.

Simple Design Matrix					
Subject	Source	Tone	Gender	Turnout	Vote
1	D	+	F	Yes	D
2	D	-	F	Yes	R
3	C	0	M	No	0
4	R	-	F	Yes	R
5	C	0	M	No	0
6	R	+	M	Yes	D
7	D	+	M	Yes	R
8	C	0	M	Yes	D
9	R	-	F	Yes	R
10	D	-	M	No	0

We can't easily analyze these data. We need numerical representations of the outcome variables Turnout and Vote and the control variables, if any – in this case Gender. For such categorical data we would need to assign indicator variables to identify each group. Exactly how we do that may depend on the sort of analysis we wish to perform. Source, for example, may lead to two indicator variables: Democratic Source and Republican Source. Each equals 1 if the statement is true and 0 otherwise. Tone might also lead to two different indicator variables.

Usually variables that are to be used in analysis are coded numerically. We might code D as 1 and R and 2 and Control group as 0. Tone might be 1 for +, 0 for none, and -1 for negative. And so forth. This coding can be done during data analysis or during data entry. A word of caution always err on the side of too rich of a coding, rather than skimp on information available when you assemble your data.

Matrices help us to analyze the information in a data matrix and help us think about the information contained (and not) in a study design. We will learn, for example, when the effects and parameters of interest can infact be estimated from a given design.

2.2. Definitions

Vectors are the building blocks for data analysis and matrix algebra.

A vector is a column of numbers. We usually denote a vector as a lower case bold letter

($\mathbf{x}$, $\mathbf{y}$, $\mathbf{a}$). When we do not have a specific set of numbers at hand but are considering a vector in the abstract. For example,

$$\mathbf{x} = \begin{pmatrix} x_1, \\ x_2, \\ x_3, \\ ., \\ ., \\ ., \\ x_n \end{pmatrix}$$

In statistics, a vector is usually a variable. Geometrically, a vector corresponds to a point in space. Each element in the vector can be graphed as the "observation space." If data were, say, from a survey, we could graph how person 1 answered the survey, person 2, etc., as a vector.

We could take another sort of vector in data analysis–the row corresponding to the values of the variables for any unit or subject. That is, a row vector is a set of numbers arranged in a single row:

$$\mathbf{a} = (\, a_1, a_2, a_3, ., ., ., a_k \,)$$

The *transpose* of a vector is a rewritting of the vector from a column to a row or from a row to a column. We denote the transpose with an elongated apostrophe $'$.

$$\mathbf{x}' = (\, x_1, x_2, x_3, ., ., ., x_n \,)$$

A *matrix* is a rectangular array of numbers, or a collection of vectors. We write matrices with capital letters. The elements of a matrix are numbers.

$$\mathbf{X} = \begin{pmatrix} x_{11}, x_{21}, ...x_{k1} \\ x_{12}, x_{22}, ...x_{k2} \\ x_{13}, x_{23}, ...x_{k3} \\ ... \\ x_{1n}, x_{2n}, ...x_{kn} \end{pmatrix}$$

or

$$\mathbf{X} = (\, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ., ., ., \mathbf{x}_\mathrm{k} \,)$$

The dimensions of a matrix are the number of rows and columns. Above, $\mathbf{X}$ has n rows and k columns so we say that it is an $n \times k$ dimension matrix or just "n by k."

It is important to keep indexes straight because operations such as addition and multiplication work on individual elements.

The transpose of a matrix represents the reorganization of a matrix such that its rows become its columns.

$$\mathbf{X}' = \begin{pmatrix} \mathbf{x}_1', \\ \mathbf{x}_2', \\ \mathbf{x}_3', \\ ., \\ ., \\ ., \\ \mathbf{x}_\mathrm{k}' \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} x_{11}, x_{12}, ...x_{1n} \\ x_{21}, x_{22}, ...x_{2n} \\ x_{31}, x_{32}, ...x_{3n} \\ ... \\ x_{k1}, x_{k2}, ...x_{kn} \end{pmatrix}$$

Note: The dimension changes from n by k to k by n.

A special type of matrix that is symmetric – the numbers above the diagonal mirror the numbers below the diagonal. This means that $\mathbf{X} = \mathbf{X}'$.

As with simple algebra, we will want to have the "numbers" 0 and 1 so that we can define division, subtraction, and other operators. A $\mathbf{0}$ vector or matrix contains all zeros. The analogue of the number 1 is called the Identity matrix. It has 1's on the diagonal and 0's elsewhere.

$$\mathbf{I} = \begin{pmatrix} 1, 0, 0, 0, ..., 0 \\ 0, 1, 0, 0, ..., 0 \\ 0, 0, 1, 0, ..., 0 \\ ... \\ 0, 0, 0, 0, ..., 1 \end{pmatrix}$$

2.3. Addition and Multiplication

2.3.1. Addition

To add vectors and matrices we sum all elements with the same index.

$$\mathbf{x}+\mathbf{y} = \begin{pmatrix} x_1 + y_1, \\ x_2 + y_2, \\ x_3 + y_3, \\ ., \\ ., \\ ., \\ x_n + y_n \end{pmatrix}$$

Matrix addition is constructed from vector addition, because a matrix is a vector of vectors. For matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A}+\mathbf{B}$ consists of first adding the vector of vectors that is A to the vector of vectors that is B and then performing vector addition for each of the vectors.

More simply, keep the indexes straight and add each element in A to the element with the same index in B. This will produce a new matrix C.

$$\mathbf{C} = \mathbf{A}+\mathbf{B} = \begin{pmatrix} a_{11} + b_{11}, a_{21} + b_{21}, ...a_{k1} + b_{k1} \\ a_{12} + b_{12}, a_{22} + b_{22}, ...a_{k2} + b_{k2} \\ ... \\ a_{1n} + b_{1n}, a_{2n} + b_{2n}, ...a_{kn} + b_{kn} \end{pmatrix}$$

NOTE: A and B must have the same dimensions in order to add them together.

2.3.2. Multiplication

Multiplication takes three forms: scalar multiplication, inner product and outer product. We will define the multiplication rules for vectors first, and then matrices.

A scalar product is a number times a vector or matrix. Let α be a number:

$$\alpha\mathbf{x} = \begin{pmatrix} \alpha x_1, \\ \alpha x_2, \\ \alpha x_3, \\ ., \\ ., \\ ., \\ \alpha x_k \end{pmatrix}$$

The inner product of vectors is the sum of the multiple of elements of two vectors with the same index. That is, $\sum_{i=1}^{n} x_i a_i$. This operation is a row vector times a column vector: $\mathbf{x}'\mathbf{a}$. The first element of the row is multiplied times the first element of the column. The second element of the row is multiplied times the second element of the column, and that product is added to the product of the first elements. And so on. This operation returns a number. Note: the columns of the row vector must equal the rows of the column vector.

In statistics this operation is of particular importance. Inner products return sums of squares and sums of cross- products. These are used to calculate variances and covariances.

The outer product of two vectors is the multiplication of a column vector (dimension n) times a row vector (dimension k). The result is a matrix of dimension n by k. The element in the first row of the column vector is multiplied by the element in the first column of the row vector. This produces the element in row 1 and column 1 in the new matrix. Generally, the ith element of the column vector is multiplied times the jth element of the row vector to get the ijth element of the new matrix.

$$\mathbf{ab}' = \begin{pmatrix} a_1b_1, a_1b_2, ...a_1b_k \\ a_2b_1, a_2b_2, ...a_2b_k \\ ... \\ a_nb_1, a_nb_2, ...a_nb_k \end{pmatrix}$$

In statistics you will frequently encounter the outer product of a vector and itself when we consider the variance of a vector.

For example, ee' yields a matrix of the residuals squared and cross-products.

$$\mathbf{ee}' = \begin{pmatrix} e_1^2, e_1e_2, ...e_1e_n \\ e_2e_1, e_2^2, ...e_2e_n \\ ... \\ e_ne_1, e_ne_2, ...e_n^2 \end{pmatrix}$$

2.3.3. Determinants

The *length* of a vector $\mathbf{x}$ equals $\mathbf{x}'\mathbf{x}$. This follows immediately from the Pythagorean Theorem. Consider a vector of dimension 2. The square of the hypotenuse is the sum of the

square of the two sides. The square of the first side is the distance traveled on dimension 1 (x_1^2) and the second side is the square of the distance traveled on dimension 2 (x_2^2).

The magnitude of a matrix is the absolute value of the determinant of the matrix. The determinant is the (signed) area inscribed by the sum of the column vectors of the matrix. The determinant is only defined for a square matrix.

Consider a 2x2 matrix with elements on the diagonal only.

$$\mathbf{X} = \begin{pmatrix} x_{11}, 0 \\ 0, x_{22} \end{pmatrix}$$

The object defined by these vectors is a rectangle, whose area is the base times the height: $x_{11}x_{22}$.

For a 2x2 matrix the determinant is $x_{11}x_{22} - x_{12}x_21$. This can be derived as follows. The sum of the vectors (x_{11}, x_{12}) and (x_{21}, x_{22}) defines a parallelogram. And the determinant is the formula for the area of the parallelogram signed by the orientation of the object. To calculate the area of the parallelogram find the area of the rectangle with sides $x_{11} + x_{21}$ and $x_{12} + x_{22}$. The parallelogram lies within this rectangle. The area inside the rectangle but not in the parallelogram can be further divided into two identical rectangles and two pairs of identical triangles. One pair of triangles is defined by the first vector and the second pair is defined by the second vector. The rectangles are defined by the first dimension of the second vector and the second dimension of the first vector. Subtracting off the areas of the triangles and rectangles leaves the formula for the determinant (up to the sign).

Note: If the first vector equalled a scalar times the second vector, the determinant would equal 0. This is the problem of *multicollinearity* – the two vectors (or variables) have the same values and they cannot be distinguished with the observed values at hand. Or, perhaps, they are really just the same variables.

Let us generalize this by considering, first, a "diagonal matrix."

$$\mathbf{X} = \begin{pmatrix} x_{11}, 0, 0, 0, ..., 0 \\ 0, x_{22}, 0, 0, ..., 0 \\ 0, 0, x_{33}, 0, ..., 0 \\ ... \\ 0, 0, 0, 0, ..., x_{nn} \end{pmatrix}$$

This is an n-dimensional box. Its area equals the product of its sides: $x_{11}x_{22}...x_{nn}$.

For any matrix, the determinant can be computed by "expanding" the cofactors along a given row (or column), say i.

$$|X| = \sum_{j=1}^{n} x_{ij}(-1)^{i+j}|X_{(ij)}|,$$

where $X_{(ij)}$, called the ijth cofactor, is the matrix left upon deleting the ith row and jth column and $|X_{(ij)}|$ is the determinant of the ijth cofactor.

To verify that this works, consider the 2x2 case. Expand along the first row:

$$|X| = x_{11}(-1)^{1+1}|x_{22}| + x_{12}(-1)^{1+2}|x_{21}| = x_{11}x_{22} - x_{12}x_{21}$$

2.4. Equations and Functions

2.4.1. Functions.

It is important to keep in mind what the inpute of a function is and what it returns (e.g., a number, a vector, a matrix). Generally, we denote functions as taking a vector or matrix as input and returning a number or matrix: $f(\mathbf{y})$.

Linear and Quadratic Forms:

Linear form: $\mathbf{Ax}$

Quadratic form: $\mathbf{x'Ax}$

2.4.2. Equations in Matrix Form. As in simple algebra an equation is such that the left side of the equation equals the right side. In vectors or matrices, the equality must hold element by element, and thus the two sides of an equation must have the same dimension.

A simple linear equation is:

$$\mathbf{Ax} = \mathbf{c}$$

or, for matrices:

$$\mathbf{AX} = \mathbf{C}$$

This defines an n-dimensional plane.

A quadratic form defines an ellipsoid.

$$\mathbf{x}'\mathbf{Ax} = \mathbf{c}$$

Given a specific number c, the values of x_i's that solve this equation map out an elliptical surface.

2.5. Inverses.

We use division to rescale variables, as an operation (e.g., when creating a new variable that is a ratio of variables), and to solve equations.

The inverse of the matrix is the division operation. The inverse is defined as a matrix, B, such that

$$\mathbf{AB} = \mathbf{I}.$$

The general formula for solving this equation is

$$b^{ij} = \frac{|\mathbf{A}_{(\mathrm{ij})}|'}{|A|},$$

where $|\mathbf{C}_{(\mathrm{ij})}|$ is the determinant of the ijth cofactor of $\mathbf{A}$.

Consider the 2x2 case

$$\begin{pmatrix} a_{11}, a_{12} \\ a_{21}, a_{22} \end{pmatrix} \begin{pmatrix} b_{11}, b_{12} \\ b_{21}, b_{22} \end{pmatrix} = \begin{pmatrix} 1, 0 \\ 0, 1 \end{pmatrix}$$

This operation produces 4 equations in 4 unknowns (the b's)

$$a_{11}b_{11} + a_{12}b_{21} = 1$$

$$a_{21}b_{11} + a_{22}b_{21} = 0$$

$$a_{11}b_{12} + a_{12}b_{22} = 0$$

$$a_{21}b_{12} + a_{22}b_{22} = 1$$

Solving these equations: $b_{11} = \frac{a_{22}}{a_{11}a_{22}-a_{12}a_{21}}$, $b_{12} = \frac{-a_{12}}{a_{11}a_{22}-a_{12}a_{21}}$, $b_{21} = \frac{-a_{21}}{a_{11}a_{22}-a_{12}a_{21}}$, and $b_{22} = \frac{a_{11}}{a_{11}a_{22}-a_{12}a_{21}}$. You can verify that the general formula above leads to the same solutions.

2.6. Statistical Concepts in Matrix Form

2.6.1. Probability Concepts

We generalize the concept of a random variable t o a *random vector*, $\mathbf{x}$, each of whose elements is a random variable, x_i. The joint density function of $\mathbf{x}$ is $f(\mathbf{x})$. The cumulative density function of $\mathbf{x}$ is the area under the density function up to a point $a_1, a_2, ...a_n$. This is an n-fold integral:

$$F(\mathbf{a}) = \int_{-\infty}^{a_n} ... \int_{-\infty}^{a_1} \mathbf{f}(\mathbf{x}) \mathbf{dx_1} ... \mathbf{dx_n}$$

For example, suppose that $\mathbf{x}$ is uniform on the intervals (0,1). This density function is: $f(\mathbf{x}) = \mathbf{1}$ if $0 < x_1 < 1, 0 < x_2 < 1, 0 < x_3 < 1, ...0 < x_n < 1$. An n-fold box, or hypercube. The cumulative density is the volume of the box with height 1 and sides $a_1, a_2, ...a_n$, so the density is $a_1 a_2 ... a_n$

2.6.1.1. Mean and Variance.

We may also characterize the frequency of random variables in terms of their means and variances. The expectations operator is, as in univariate probability, the weighted average of the values of the variable, where the weights are the frequencies or probabilities. In taking the expected value of a vector or matrix we evaluate the expected value of each element of the vector or matrix. Let $\mathbf{x}$ be a random vector.

$$E[\mathbf{x}] = \begin{pmatrix} E[x_1], \\ E[x_2], \\ E[x_3], \\ . \\ . \\ . \\ E[x_n] \end{pmatrix} \begin{pmatrix} E[x_1], \\ \mu_2, \\ \mu_3, \\ . \\ . \\ . \\ \mu_n \end{pmatrix}$$

The variance of a random vector is the set of variances and covariances of the variables.

This is defined as the expected value of the outer product of the vector:

$$E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)'] = \begin{pmatrix} E[(x_1-\mu_1)^2], E[(x_1-\mu_1)(x_2-\mu_2)], ..., E[(x_1-\mu_1)(x_n-\mu_n)] \\ E[(x_2-\mu_2)(x_1-\mu_1)], E[(x_2-\mu_2)^2], ..., E[(x_2-\mu_2)(x_n-\mu_n)] \\ \cdots \\ E[(x_n-\mu_n)(x_1-\mu_1)], E[(x_n-\mu_n)(x_2-\mu_2)], ...E[(x_n-\mu_n)^2] \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2, \sigma_{12}, \sigma_{13}, ..., \sigma_{1n} \\ \sigma_{12}, \sigma_2^2, \sigma_{23}, ..., \sigma_{2n} \\ .... \\ \sigma_{1n}, \sigma_{2n}, \sigma_{3n}, ..., \sigma_n^2 \end{pmatrix}$$

2.6.1.2. The Normal and Related Distributions

The first two moments are sufficient to characterize a wide range of distribution functions. Most importantly, the mean vector and the variance covariance matrix characterize the normal distribution.

$$f(\mathbf{x}) = (\mathbf{2}\pi)^{-\mathrm{n}/2}|\mathbf{\Sigma}|^{-1/2}\mathbf{e}^{-\frac{1}{2}(\mathrm{x}-\mu)'\,\S\,^{-1}(\mathrm{x}-\mu)}$$

If the x_i are independent (0 covariance) and identical (same variances and means), then Σ simplifies $\sigma^2\mathbf{I}$, and the normal becomes:

$$f(\mathbf{x}) = (\mathbf{2}\pi)^{-\mathrm{n}/2}\sigma^{-\mathrm{n}}\mathbf{e}^{-\frac{1}{2\sigma^2}(\mathrm{x}-\mu)'(\mathrm{x}-\mu)}$$

A linear transformation of a normal random variable is also normal. Let $\mathbf{A}$ be a matrix of constants and $\mathbf{c}$ be a vector of constants and $\mathbf{x}$ be a random vector, then

$$\mathbf{Ax} + \mathbf{c} \sim \mathbf{N}(\mathbf{A}\mu + \mathbf{c}, \mathbf{A\Sigma A}')$$

We may define the standard normal as follows. Let $\mathbf{A}$ be a matrix such that $\mathbf{AA}' = \mathbf{\Sigma}$. Let $\mathbf{c} = \mu$. Let $\mathbf{z}$ be a random vector such that $\mathbf{x} = \mathbf{Az} + \mu$. Then, $\mathbf{z} = \mathbf{A}^{-1}(\mathbf{x} - \mu)$. Because the new variable is a linear transformation of $\mathbf{x}$, we know that the result will be normal. What are the mean and variance:

$$E[\mathbf{z}] = \mathbf{E}[\mathbf{A}^{-1}(\mathbf{x} - \mu)] = \mathbf{A}^{-1}\mathbf{E}(\mathbf{x} - \mu) = \mathbf{0}$$

$$V(\mathbf{z}) = \mathbf{E}[(\mathbf{A}^{-1}(\mathbf{x}-\mu))(\mathbf{A}^{-1}(\mathbf{x}-\mu))']$$

$$= E[(\mathbf{A}^{-1}(\mathbf{x}-\mu))((\mathbf{x}-\mu)'\mathbf{A}'^{-1}]$$

$$= \mathbf{A}^{-1}\mathbf{E}[(\mathbf{x}-\mu))((\mathbf{x}-\mu))']\mathbf{A}'^{-1} = \mathbf{A}^{-1}\mathbf{\Sigma}'\mathbf{A}'^{-1} = \mathbf{I}.$$

The last equality holds because $\mathbf{AA}' = \mathbf{\Sigma}$.

From the normal we derive the other distributions of use in statistical inference. The two key distributions are the χ^2 and the F.

The quadratic form of a normally distributed random vector will follow the χ^2 distribution. Assume a standard normal vector $\mathbf{z}$.

$$\mathbf{z}'\mathbf{z} \sim \chi^2_{\mathrm{n}}$$

because this is the sum of the squares of n normal random variables with mean 0 and variance 1. Alternatively, if $\mathbf{x}$ is a normal random variable with mean μ and variance $\mathbf{\Sigma}$, then $(\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)$ follows a χ^2 with n degrees of freedom.

The ratio of two quadratic forms will follow the F distribution. Let $\mathbf{B}$ and $\mathbf{C}$ be two matrices such that $\mathbf{BC} = \mathbf{0}$. Let $\mathbf{A}$ and $\mathbf{B}$ have r_a and r_b independent columns, respectively (i.e., ranks). Let $\mathbf{x}$ be a random vector with variance $\sigma^2\mathbf{I}$, then

$$\frac{(\mathbf{x}'\mathbf{AA}/\sigma^2)/\mathbf{r}_{\mathrm{a}})}{(\mathbf{x}'\mathbf{Bx}/\sigma^2)/\mathbf{r}_{\mathrm{b}})} \sim F[r_a, r_b]$$

In statistical inference about regressions we will treat the vector of coefficients as a random vector, $\mathbf{b}$. We will construct hypotheses about sets of coefficients from a regression, and express those in the form of the matrices $\mathbf{B}$ and $\mathbf{C}$. The χ^2-statistic amounts to comparing squared deviations of the estimated parameters from the hypothesized paramenters. The F-statistic compares two models – two different sets of restrictions.

2.6.2.3. Conditional Distributions.

For completeness, I present the conditional distributions. These are frequently used in Bayesian statistical analysis; they are also used as a theoretical framework within which to think about regression. They won't be used much in what we do in this course.

To define conditional distributions we will need to introduce an additional concept, partitioning. A matrix can be subdivided into smaller matrices. For example, an nxn matrix $\mathbf{A}$ may be written as, say, four submatrices $\mathbf{A}_{11}, \mathbf{A}_{12}, \mathbf{A}_{21}, \mathbf{A}_{22}$:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11}, \mathbf{A}_{12} \\ \mathbf{A}_{21}, \mathbf{A}_{22} \end{pmatrix}$$

We can write the inverse of A as the inverse of the partitions as follows:

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}_{11}^{-1}(\mathbf{I} + \mathbf{A}_{12}\mathbf{F}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}), -\mathbf{A}_{11}\mathbf{A}_{12}\mathbf{F} \\ -\mathbf{F}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}, \quad \mathbf{F} \end{pmatrix},$$

where $\mathbf{F} = (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}$. We can use the partitioning results to characterize the conditional distributions.

Consider two normally distributed random vectors $\mathbf{x}_1$, $\mathbf{x}_2$: The density function for $\mathbf{x}_1$ is

$$f(\mathbf{x}_1) \sim \mathbf{N}(\mu_1, \mathbf{\Sigma}_{11}).$$

The density function for $\mathbf{x}_1$ is

$$f(\mathbf{x}_2) \sim \mathbf{N}(\mu_2, \mathbf{\Sigma}_{22}).$$

Their joint distribution is

$$f(\mathbf{x}_1, \mathbf{x}_2) \sim \mathbf{N}(\mu, \mathbf{\Sigma}),$$

The variance matrix can be written as follows:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11}, \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21}, \mathbf{\Sigma}_{22} \end{pmatrix}$$

where $\mathbf{\Sigma}_{\mathrm{jk}} = \mathbf{E}[(\mathbf{x}_{\mathrm{j}} - \mu_{\mathrm{j}}\mathbf{x}_{\mathrm{k}} - \mu_{\mathrm{k}})]$ and $j = 1, 2$ and $k = 1, 2$.

The conditional distribution is arrived at by dividing the joint density by the marginal density of $\mathbf{x}_1$, yielding

$$\mathbf{x}_2|\mathbf{x}_1 \sim \mathbf{N}(\mu_{2.1}, \mathbf{\Sigma}_{22.1})$$

where $\mu_{2.1} = \mu_2 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}(\mathbf{x}_2 - \mu_2)$ and $\mathbf{\Sigma}_{22.1} = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{21}$.

2.6.2. Regression model in matrix form.

1. Linear Model

The linear regression model consists of n equations. Each equation expresses how the value of y for a given observation depends on the values of $\mathbf{x_i}'$ and ϵ_i.

$$y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} ... + \beta_k X_{k1} + \epsilon_1$$

$$y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} ... + \beta_k X_{k2} + \epsilon_2$$

$$. \qquad ..$$

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} ... + \beta_k X_{ki} + \epsilon_i$$

$$. \qquad ..$$

$$y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} ... + \beta_k X_{kn} + \epsilon_n$$

This can be rewritten using matrices. Let $\mathbf{y}$ be the random vector of the dependent variable and ϵ be the random vector of the errors. The matrix $\mathbf{X}$ contains the values of the independent variables. The rows are the observations and the columns are the variables. To capture the intercept, the first column is a column of 1's.

$$\mathbf{X} = \begin{pmatrix} 1, x_{11}, ...x_{k1} \\ 1, x_{12}, ...x_{k2} \\ 1, x_{13}, ...x_{k3} \\ ... \\ 1, x_{1n}, ...x_{kn} \end{pmatrix}$$

Let β be a vector of $k+1$ constants: $\beta' = (\beta_0, \beta_1, \beta_2, ..\beta_k)$. Then, $\mathbf{X}\beta$ produces a vector in which each entry is a weighted average of the values of $X_{1i}, ...X_{ki}$ for a given observation i where the weights are the β's.

$$\mathbf{X}\beta = \begin{pmatrix} \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} ... + \beta_k X_{k1} + \epsilon_1 \\ \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} ... + \beta_k X_{k2} + \epsilon_2 \\ ... \\ \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} ... + \beta_k X_{ki} + \epsilon_i \\ ... \\ \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} ... + \beta_k X_{kn} + \epsilon_n \end{pmatrix}$$

With this notation, we can express the linear regression model more succinctly as:

$$y = X\beta + \epsilon$$

2. Assumptions about Errors'

A1. Errors have mean 0.

$$E[\epsilon] = 0$$

A2. Spherical distribution of errors (no autocorrelation, homoskedastic).

$$E[\epsilon\epsilon'] = \sigma_\epsilon^2 \mathbf{I}$$

A3. Independence of X and ϵ.

$$E[X'\epsilon] = 0$$

3. Method of Moments Estimator

The empirical moments implied by Assumptions 1 and 3 are that

$$i'e = 0$$

$$X'e = 0$$

where $e = y - Xb$. We can use these to derive the method of moments estimator of b, i.e., the formula we use to make our best guess about the value of β based on the observed y and X.

Substitute the definition of e into the second equation. This yields

$$X'(y - Xb) = 0$$

Hence, $X'y - X'Xb = 0$, which we can rewrite as $(X'X)b = X'y$. This is a system of linear equations. $X'X$ is a kxk matrix. We can invert it to solve the system:

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

Conceptually, this is the covariance(s) between y and X divided by the variance of X, which is the same idea as the formula for bivariate regression. The exact formula for the coefficient on any given variable, say X_j, will not be the same as the formula for the coefficient from the bivariate regression of Y on X_j. We must adjust for the other factors affecting Y that are correlated with X_j (i.e., the other X variables).

Consider a case with 2 independent variables and a constant. For simplicity, let us deviate all variables from their means: This eliminates the constant term, but does not change the solution for the slope parameters. So the vector $\mathbf{x_j}$ is really $\mathbf{x_j} - \bar{\mathbf{x}}_\mathbf{j}$. Using the inner products vectors to denote the sums in each element, we can express the relevant matrices as follows.

$$\mathbf{X'X} = \begin{pmatrix} x_1'x_1, x_1'x_2 \\ x_2'x_1, x_2'x_2 \end{pmatrix}$$

$$(\mathbf{X'X})^{-1} = \frac{\mathbf{1}}{\mathbf{x_1'x_1x_2'x_2} - (\mathbf{x_1'x_2})^2} \begin{pmatrix} \mathbf{x_2'x_2}, -\mathbf{x_1'x_2} \\ -\mathbf{x_2'x_1}, \mathbf{x_1'x_1} \end{pmatrix}$$

$$(\mathbf{X'y}) = \begin{pmatrix} \mathbf{x_1'y} \\ \mathbf{x_2'y} \end{pmatrix}$$

Calculation of the formula for $\mathbf{b}$ consists of multiplying the $X'X$ matrix (x-prime-x matrix) times the vector $X'y$. This is analogous to dividing the covariance of X and Y by the variance of X.

$$\mathbf{b} = \frac{\mathbf{1}}{\mathbf{x_1'x_1x_2'x_2} - (\mathbf{x_1'x_2})^2} \begin{pmatrix} \mathbf{x_2'x_2}, -\mathbf{x_1'x_2} \\ -\mathbf{x_2'x_1}, \mathbf{x_1'x_1} \end{pmatrix} \begin{pmatrix} \mathbf{x_1'y} \\ \mathbf{x_2'y} \end{pmatrix}$$

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} \frac{(\mathrm{x_2'x_2})(\mathrm{x_1'y}) - (\mathrm{x_1'x_2})(\mathrm{x_2'y})}{(\mathrm{x_1'x_1})(\mathrm{x_2'x_2}) - (\mathrm{x_1'x_2})^2} \\ \\ \frac{(\mathrm{x_1'x_1})(\mathrm{x_2'y}) - (\mathrm{x_1'x_2})(\mathrm{x_1'y})}{(\mathrm{x_1'x_1})(\mathrm{x_2'x_2}) - (\mathrm{x_1'x_2})^2} \end{pmatrix}$$

2.7. Differentiation and Optimization

We will encounter differentiation in statistics in estimation, where we wish to minimize squared errors or maximize likelihood, and in developing approximations, especially Taylor's Theorem.

The primary optimization problems take one of the following two forms.

First, least squares consists of choosing the values of $\mathbf{b}$ that minimize the sum of squared errors with respect to β:

$$S = \epsilon'\epsilon = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

This is a quadratic form in β.

Second, maximum likelihood estimation consists of choosing the values of the parameters that maximize the probability of observing the data. If the errors are assumed to come from the normal distribution, the likelihood of the data is expressed as the joint density of the errors:

$$L(\epsilon; \beta, \sigma_\epsilon^2) = (\mathbf{2}\pi)^{-\mathrm{n}/2}\sigma_\epsilon^{-\mathrm{n}}\mathbf{e}^{-\frac{1}{2\sigma_\epsilon^2}\epsilon'\epsilon}$$
$$= (2\pi)^{-n/2}\sigma_\epsilon^{-n}e^{-\frac{1}{2\sigma_\epsilon^2}(\mathrm{y}-\mathrm{X}\beta)'(\mathrm{y}-\mathrm{X}\beta)}$$

This function is very non-linear, but it can be expressed as a quadratic form after taking logarithms:

$$ln(L) = -\frac{n}{2}ln(2\pi) - nln(\sigma_\epsilon) - \frac{1}{2\sigma_\epsilon^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

The objective in both problems is to find a vector $\mathbf{b}$ that optimizes the objective functions with respect to β. In the least squares case we are minimizing and the likelihood case we are maximizing. In both cases, the functions involved are continuous, so the first and second order conditions for a maximum will allow us to derive the optimal $\mathbf{b}$.

Although the likelihood problem assumes normality, the approach can be adapted to many different probability densities, and provides a fairly flexible framework for deriving estimates.

Regardless of the problem, though, maximizing likelihood or minimizing error involves differentiation of the objective function. I will teach you the basic rules used in statistical applications. They are straightforward but involve some care in accounting the indexes and dimensionality.

2.7.1. Partial Differentiation: Gradients and Hessians. Let $y = f(\mathbf{x})$. The outcome variable y changes along each of the x_i dimensions. We may express the set of first partial derivatives as a vector, called the gradient.

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f(\mathrm{x})}{\partial x_1} \\ \frac{\partial f(\mathrm{x})}{\partial x_2} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial f(\mathrm{x})}{\partial x_n} \end{pmatrix}$$

In statistics we will encounter linear forms and quadratic forms frequently. For example, the regression formula is a linear form and the sums of squared errors is expressed as a quadratic form. The gradients for the linear and quadratic forms are as follows:

$$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}'$$

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}')\mathbf{x}$$

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{A}} = \mathbf{x}\mathbf{x}'$$

Consider the linear form. We build up from the special case where $\mathbf{A}$ is a vector $\mathbf{a}$. Then, $y = \mathbf{a}'\mathbf{x} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{a}_{\mathrm{i}}\mathbf{x}_{\mathrm{i}}$. The gradient returns a vector of coefficients:

$$\frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial \sum_{i=1}^{n} a_i x_i}{\partial x_1} \\ \frac{\partial \sum_{i=1}^{n} a_i x_i}{\partial x_2} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial \sum_{i=1}^{n} a_i x_i}{\partial x_n} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_n \end{pmatrix} = \mathbf{a}$$

Note: $\mathbf{a}$ is transposed in the linear form, but not in the gradient.

More generally, the gradient of $\mathbf{y} = \mathbf{Ax}$ with respect to $\mathbf{x}$ returns the transpose of the matrix $\mathbf{A}$. Any element, y_i, equals a row vector of $\mathbf{A} = \mathbf{a}^{\mathrm{i}}$ times the column vector $\mathbf{x}$. Differentiation of the ith element of $\mathbf{y}$ returns a column vector that is the transpose of the ith row vector of $\mathbf{A}$. The gradient of each element of $\mathbf{y}$ returns a new column vector. As a result, the rows of $\mathbf{A}$ become the columns of the matrix of gradients of the linear form.

The Hessian is the matrix of second partial derivatives:

$$\mathbf{H} = \begin{pmatrix} \frac{\partial f^2(\mathrm{x})}{\partial x_1 \partial x_1}, \frac{\partial f^2(\mathrm{x})}{\partial x_1 \partial x_2}, \cdots \frac{\partial f^2(\mathrm{x})}{\partial x_1 \partial x_n} \\ \frac{\partial f^2(\mathrm{x})}{\partial x_2 \partial x_1}, \frac{\partial f^2(\mathrm{x})}{\partial x_2 \partial x_2}, \cdots \frac{\partial f^2(\mathrm{x})}{\partial x_2 \partial x_n} \\ \cdots \\ \frac{\partial f^2(\mathrm{x})}{\partial x_n \partial x_1}, \frac{\partial f^2(\mathrm{x})}{\partial x_n \partial x_2}, \cdots \frac{\partial f^2(\mathrm{x})}{\partial x_n \partial x_n} \end{pmatrix}$$

Note: This is a symmetric matrix. Each column is the derivative of the function f with respect to the transpose of $\mathbf{X}$. So the Hessian is often written:

$$\mathbf{H} = [\mathbf{f}_{\mathrm{ij}}] = \frac{\partial^2 \mathbf{y}}{\partial \mathbf{x} \mathbf{x}'}.$$

Hessian's are commonly used in Taylor Series approximations, such as used in the Delta Technique to approximate variances. The second order Taylor Series approximation can be expressed as follows:

$$y \approx \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_i^0)(x_j - x_j^0) f_{ij}(\mathbf{x}^0) = \frac{\mathbf{1}}{\mathbf{2}} (\mathbf{x} - \mathbf{x}^0)' \mathbf{H}(\mathbf{x}^0)(\mathbf{x} - \mathbf{x}^0)$$

They are also important in maximum likelihood estimation, where the Hessian is used to construct the variance of the estimator. For the case of a normally distributed random vector with mean μ and variance σ^2, the Hessian is

$$\mathbf{H} = \begin{pmatrix} -n/\sigma^2, & \frac{-\sum_{i=1}^{n}(x_i - \mu)}{\sigma^4} \\ \frac{-\sum_{i=1}^{n}(x_i - \mu)}{\sigma^4}, & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (x_i - \mu)^2 \end{pmatrix}$$

The expected value of the inverse of this matrix is the variance-covariance matrix of the estimated parameters.

2.7.2. Optimization. At an optimum the gradient equals $\mathbf{0}$ – in all directions the rate of change is 0. This condition is called the first order condition. To check whether a given

point is a maximum or a minimum requires checking second order conditions. I will not examine second order conditions, but you are welcome to read about and analyze them in the textbook.

The first order condition amounts to a system of equations. At an optimum, there is a vector that solves the minimization or maximization problem in question. That vector is arrived at by setting the gradient equal to zero and solving the system of equations.

Consider the least squares problem. Find the value of $\beta = b$ that minimizes:

$$S = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

Using the rules of transposition, we can rewrite this as follows:

$$S = (\mathbf{y}' - \beta'\mathbf{X}')(\mathbf{y} - \mathbf{X}\beta) = (\mathbf{y}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta - \mathbf{y}'\mathbf{X}\beta - \beta'\mathbf{X}'\mathbf{y}) = \mathbf{y}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta - \mathbf{2}\mathbf{y}'\mathbf{X}\beta$$

The gradient of S with respect to β consists of derivatives of quadratic and linear forms. Using the rules above:

$$\frac{\partial S}{\partial \beta} = 2\mathbf{X}'\mathbf{X}\beta - \mathbf{2}\mathbf{X}'\mathbf{y}$$

At a minimum, the value $\mathbf{b}$ guarantees that the gradient equals 0:

$$\frac{\partial S}{\partial \beta} = 2\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{2}\mathbf{X}'\mathbf{y} = \mathbf{0}$$

The solution in terms of $\mathbf{b}$ is

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

For the maximum likelihood problem, the first order conditions are:

$$\frac{\partial ln(L)}{\partial \beta} = -\frac{1}{2\hat{\sigma}_\epsilon^2}(2\mathbf{X}'\mathbf{X}\hat{\beta} - \mathbf{2}\mathbf{X}'\mathbf{y}) = \mathbf{0}$$

$$\frac{\partial ln(L)}{\partial \sigma_\epsilon} = -n\hat{\sigma}_\epsilon^{-1} + \sigma_\epsilon^{-3} + ((\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{0}$$

The solution is

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

$$\hat{\sigma}_\epsilon = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) = \frac{\mathbf{1}}{\mathbf{n}}\mathbf{e}'\mathbf{e}$$

where $\mathbf{e}$ is the vector of residuals.