Lecture 2: Linear Algebra Review

Reading

- Weisberg, Appendix A.6-7: "A Brief Introduction to Matrices and Vectors"

- Muller and Fetterman, Appendix A: "Matrix Algebra Useful for Linear Models"

- Namboodiri: "Matrix Algebra: An Introduction" (Optional)

Why Linear Algebra?

Data for linear models can be can be represented as vectors and matrices

Simplification of theory and estimation of linear models, general representation

Easier to address certain problem in estimating such models

Basics of Notation

A *matrix* is a two-dimensional array of elements.

$\mathbf{A} = \{a_{ij}\}$ means $\mathbf{A}$ is the matrix whose i, j^{th} element (or component) is the *scalar* a_{ij}, where i indexes row and j indexes column, $i = 1, \ldots, r$, $j = 1, \ldots, c$. In this context, a scalar s will represent a real number. Capital letters are used to represent matrices, and lowercase letters are used for vectors.

The *dimension* of $\mathbf{A}$ is $(r \times c)$ (say "r by c"), often written $\mathbf{A}_{r \times c}$.

The entire matrix $\mathbf{A}$ may be represented

$$\mathbf{A}_{r \times c} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1c} \\ a_{21} & \ddots & & \ldots \\ \ldots & & \ddots & \ldots \\ a_{r1} & \ldots & \ldots & a_{rc} \end{bmatrix}.$$

A matrix with one column, i.e. an $(r \times 1)$ matrix, is called a *vector* or

column vector. A $(1 \times r)$ matrix is called a *row vector.* We can represent the matrix $\mathbf{A}_{r \times c}$ in terms of its columns $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_c\}$. For example, we might represent the $\mathbf{X}$ matrix from the one of the examples in lecture 1 as

$$\mathbf{X} = \{\mathbf{1}, \mathbf{height}\}$$

$$= \begin{bmatrix} 1 & 61.61 \\ 1 & 59.17 \\ 1 & 54.75 \\ \vdots & \vdots \\ 1 & 62.92 \end{bmatrix}.$$

Types of Matrices

A *square matrix* has the same number of rows as columns, so that $r = c$.

For $\mathbf{A}_{r \times c}$ and $r \geq c$, the *diagonal* of $\mathbf{A}$ is $\{a_{11}, a_{22}, \ldots, a_{cc}\}$.

A *symmetric matrix* is a square matrix with $a_{ij} = a_{ji}$ for all i, j. That is, the entry in row i and column j is the same as the entry in row j and column i. Only square matrices can be symmetric. The matrix

$$
\begin{bmatrix}
a_{11} & a_{12} & a_{13} \\
a_{12} & a_{22} & a_{23} \\
a_{13} & a_{23} & a_{33}
\end{bmatrix}
$$

is symmetric (note that we refer to symmetry about the main diagonal, which runs from the upper left to the lower right).

A square matrix is called a *diagonal matrix* if all elements off the main diagonal are zero; that is, if $i \neq j$, then $a_{ij} = 0$. We write $\text{Diag}(\mathbf{b}) = \text{Dg}(\mathbf{b})$ for a square diagonal matrix created from a vector. For example,

$$
\text{Diag} \left(\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \right) = \begin{bmatrix} b_1 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{bmatrix}.
$$

An *identity matrix*, denoted $\mathbf{I} = \mathbf{I}_n = \mathbf{I}_{n \times n}$, is a diagonal matrix with all 1's on the main diagonal and 0's elsewhere. That is, $a_{ij} = 1$ for $i = j$ and $a_{ij} = 0$ for $i \neq j$. Thus

$$
\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.
$$

Denote an $(n \times 1)$ vector of 1's by

$$\mathbf{1} \;=\; \mathbf{1}_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n \times 1} = \mathbf{J}_n.$$

A *zero matrix* or *null matrix*, denoted $\mathbf{0}_{r \times c}$, has $a_{ij} = 0$ for all i, j. So

$$\mathbf{0}_{\mathbf{2} \times \mathbf{2}} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

An *upper triangular matrix* $\mathbf{A}$ has $a_{ij} = 0$ for $i > j$. That is,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}$$

is an upper triangular matrix. A *lower triangular matrix* has $a_{ij} = 0$ for $i < j$.

A *partitioned* matrix has elements grouped into submatrices by combinations of vertical and horizontal slicing. For example,

$$
\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \vdots & a_{13} \\ a_{21} & a_{22} & \vdots & a_{23} \\ \dots & \dots & \dots & \dots \\ a_{31} & a_{32} & \vdots & a_{33} \end{bmatrix}_{3\times 3} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}_{3\times 3},
$$

where

$$
\mathbf{A}_{11} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}_{2\times 2}, \ \mathbf{A}_{12} = \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix}_{2\times 1},
$$

$$
\mathbf{A}_{21} = \begin{bmatrix} a_{31} & a_{32} \end{bmatrix}_{1\times 2}, \ \text{and } \mathbf{A}_{22} = \begin{bmatrix} a_{33} \end{bmatrix}_{1\times 1}.
$$

A *block diagonal matrix* has square diagonal submatrices with all

off-diagonal submatrices equal to 0. For example,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}$$

is a block diagonal matrix.

Matrix Operations

The *trace* of a square matrix $\mathbf{A}_{n \times n}$ is the sum of the diagonal elements of $\mathbf{A}$. That is, $\text{trace}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$.

$$\text{trace}\left(\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\right) = a_{11} + a_{22} + a_{33}$$

.

The *transpose* of a matrix $\mathbf{A}$, denoted $\mathbf{A}'$, changes the rows of $\mathbf{A}$ into the columns of a new matrix $\mathbf{A}'$. If $\mathbf{A}$ is an $(r \times c)$ matrix, its transpose $\mathbf{A}'$ is a $(c \times r)$ matrix.

The transpose of a column vector is a row vector, i.e.

$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}_{4 \times 1}' = \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}_{1 \times 4}.$$

A symmetric matrix has $\mathbf{A}' = \mathbf{A}$.

Matrix addition of two matrices $\mathbf{A}$ and $\mathbf{B}$ is defined only if $\mathbf{A}$ and $\mathbf{B}$ have the same number of rows and the same number of columns, *matrix addition* yields $\mathbf{A}_{r \times c} + \mathbf{B}_{r \times c} = \{a_{ij} + b_{ij}\}_{r \times c}$.

Exercise:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \qquad .$$

There are several types of *matrix multiplication*. We will discuss the following types:

1. scalar multiplication,

2. matrix multiplication

- Define *scalar* multiplication of any matrix $\mathbf{A}$ by a scalar s as $s\mathbf{A} = \{sa_{ij}\}$.

  Exercise:

$$\sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \qquad\qquad .$$

- Two matrices $\mathbf{A}$ and $\mathbf{B}$ conform for the *matrix multiplication* $\mathbf{AB}$ if the number of columns of $\mathbf{A}$ is equal to the number of rows of $\mathbf{B}$. If the matrices $\mathbf{A}$ and $\mathbf{B}$ conform, *matrix multiplication* is defined as

$$\mathbf{A}_{r\times s}\mathbf{B}_{s\times t} = \left\{ \sum_{k=1}^{s} a_{ik}b_{kj} \right\} = \mathbf{C}_{r\times t}.$$

Multiplying the ith row of $\mathbf{A}$ with the jth column of $\mathbf{B}$ yields the scalar $c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{is}b_{sj}$:

Calculation of c_{11}:

$$
\text{row} \longrightarrow \begin{bmatrix} a_{11} & \rightarrow & \rightarrow & a_{1s} \\ & & & \\ & & & \\ & & & \end{bmatrix}_{r \times s} \begin{bmatrix} b_{11} \\ \downarrow \\ \downarrow \\ b_{s1} \end{bmatrix}_{s \times t} \qquad \downarrow \text{column}
$$

So $c_{11} = a_{11}b_{11} + a_{12}b_{21} + \ldots + a_{1s}b_{s1}$.

Note that $\mathbf{AI} = \mathbf{A}$ and $\mathbf{IA} = \mathbf{A}$. In general, $\mathbf{AB} \neq \mathbf{BA}$.

Exercise: Let $\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix}$.

$\mathbf{X'X} =$

Orthogonal Matrices

An *orthogonal matrix* is a **square matrix** with $\mathbf{A}' = \mathbf{A}^{-1}$. That is, a square matrix $\mathbf{A}$ is orthogonal if $\mathbf{A}'\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}'$. To establish that $\mathbf{A}$ is orthogonal, it is sufficient to show either that $\mathbf{A}'\mathbf{A} = \mathbf{I}$ or that $\mathbf{A}\mathbf{A}' = \mathbf{I}$.

- The vectors $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal vectors* if $\mathbf{x}'\mathbf{y} = 0$.

- The vectors $\mathbf{x}$ and $\mathbf{y}$ are *orthonormal vectors* if $\mathbf{x}$ and $\mathbf{y}$ are orthogonal vectors and are normalized: $\mathbf{x}'\mathbf{y} = 0$, $\mathbf{x}'\mathbf{x} = 1$, and $\mathbf{y}'\mathbf{y} = 1$.

- Some authors, including Muller and Fetterman, call an orthogonal matrix a column orthonormal matrix. While this is more accurate, we will retain standard terminology and refer to a matrix with $\mathbf{A}'\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}'$ as orthogonal. Thus, somewhat paradoxically, an orthogonal matrix has orthonormal columns.

Rules of Matrix Operation

Suppose $\mathbf{A}$ and $\mathbf{B}$ conform for the operation of interest. The following laws apply to $\mathbf{A}$ and $\mathbf{B}$.

Commutative Laws

- $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$

- $a\mathbf{B} = \mathbf{B}a$

In general, $\mathbf{A} f B \neq \mathbf{B}\mathbf{A}$.

Distributive Laws

- $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{A}\mathbf{B} + \mathbf{A}\mathbf{C}$

- $(\mathbf{B} + \mathbf{C})\mathbf{D} = \mathbf{B}\mathbf{D} + \mathbf{C}\mathbf{D}$

- $a(\mathbf{B} + \mathbf{C}) = a\mathbf{B} + a\mathbf{C} = (\mathbf{B} + \mathbf{C})a$

Associative Laws

- $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$

- $(\mathbf{A}\mathbf{B})\mathbf{C} = \mathbf{A}(\mathbf{B}\mathbf{C})$

Transpose Operations

- $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$

- $(\mathbf{A}\mathbf{B})' = \mathbf{B}'\mathbf{A}'$

Linear Dependence and Rank

Consider the system of equations

$$5x_1 + 2x_2 = 2 \tag{1}$$

$$10x_1 + 4x_2 = 4, \tag{2}$$

or

$$\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 = \mathbf{b},$$

where

$$\mathbf{a}_1 = \begin{pmatrix} 5 \\ 10 \end{pmatrix}, \mathbf{a}_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \text{ and } \mathbf{b} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}.$$

- This system of equations has infinite solutions given by $x_1 = \frac{2}{5} - \frac{2}{5}x_2$.

- We do not have one unique solution because the coefficients in

the second equation are a multiple of the first and thus are *linearly dependent*.

- So equation (2) does not provide any additional information about x_1 or x_2.

- Examining $\mathbf{a}_1$ and $\mathbf{a}_2$, we see that $a_{11} = \frac{5}{2}a_{12}$ and $a_{21} = \frac{5}{2}a_{22}$. Thus $\mathbf{a}_1$ and $\mathbf{a}_2$ are linearly dependent vectors.

The n vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ in $\mathcal{R}^n$ are *linearly dependent* if there exist numbers $c_1, c_2, \ldots, c_n$ (not all zero) such that

$$\mathbf{a}_1 c_1 + \mathbf{a}_2 c_2 + \ldots + \mathbf{a}_n c_n = 0.$$

If this equation is true only when $c_1 = c_2 = \ldots = c_n = 0$ then the vectors are *linearly independent*.

More generally, suppose we have the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 2 & 4 \\ 1 & 0 & 2 \\ 0 & 8 & 0 \end{bmatrix}.$$

Each column of $\mathbf{A}$ may be viewed as a vector. The *column space* of $\mathbf{A}$, denoted $C(\mathbf{A})$, is the set of all vectors that may be written as a linear combination of the columns of $\mathbf{A}$. That is, $C(\mathbf{A})$ is the set of all vectors that may be written as

$$\lambda_1 \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2 \\ 0 \\ 8 \end{bmatrix} + \lambda_3 \begin{bmatrix} 4 \\ 2 \\ 0 \end{bmatrix} = \mathbf{A} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \mathbf{A}\boldsymbol{\lambda},$$

for some vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)'$.

For example, the column space of $\mathbf{J}_2$ includes all vectors of the form

$$\lambda \mathbf{J}_2 = \begin{pmatrix} \lambda \\ \lambda \end{pmatrix},$$

including

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} -0.03 \\ -0.03 \end{pmatrix}, \text{ and } \begin{pmatrix} \pi \\ \pi \end{pmatrix}.$$

The columns of $\mathbf{A}$ are *linearly dependent* if they contain redundant information. If we can find two distinct vectors, say $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$, such that $\mathbf{A}\boldsymbol{\lambda} = \mathbf{A}\boldsymbol{\gamma} = \mathbf{x}$, then the columns of $\mathbf{A}$ are linearly dependent.

An equivalent definition, obtained by letting $\boldsymbol{\delta} = \boldsymbol{\lambda} - \boldsymbol{\gamma}$ (prove to yourself!), is that the columns of $\mathbf{A}$ are linearly dependent if there exists a vector $\boldsymbol{\delta} \neq \mathbf{0}$ such that $\mathbf{A}\boldsymbol{\delta} = \mathbf{0}$. If the columns of $\mathbf{A}$ are not linearly dependent, then they are *linearly independent*.

Exercise: Does the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 2 & 4 \\ 1 & 0 & 2 \\ 0 & 8 & 0 \\ 3 & 1 & 6 \end{bmatrix}$$

have linearly independent or linearly dependent columns?

Exercise: Does the matrix

$$\mathbf{B} = \begin{bmatrix} 2 & 2 & 4 \\ 1 & 0 & 2 \\ 0 & 8 & 0 \\ 4 & 1 & 6 \end{bmatrix}$$

have linearly independent or linearly dependent columns?

The *rank* of a matrix $\mathbf{A}$ is the number of linearly independent columns in $\mathbf{A}$. Knowledge of the matrix rank is important in determining the existence and multiplicity of solutions to a system of linear equations.

If $\mathbf{A}$ is an $(r \times c)$ matrix with $r \geq c$, we say $\mathbf{A}$ is *full rank* if rank($\mathbf{A}$)=c. If rank($\mathbf{A}$)$< c$, then we say $\mathbf{A}$ is less than full rank. In linear regression, the matrix of covariates $\mathbf{X}$ must have full rank in order for our parameter estimates, $\widehat{\boldsymbol{\beta}}$, to be unique.

A square matrix that is less than full rank is called *singular*, while a full rank square matrix is called *nonsingular*. One method to find the rank of a matrix is by determining the number of linearly independent columns of a matrix.

Another method to find rank is by using elementary row operations to transform the matrix to a triangular matrix; once the matrix is in triangular form, we can determine the rank visually.

The three elementary row operations are

1. multiplying a row by a nonzero constant,

2. adding one row to another, and

3. exchanging two rows.

Other approaches to finding rank include using matrix decompositions or finding eigenvalues (the rank of a square matrix is equal to the number of nonzero eigenvalues).

Determinants

The *determinant* is a single number summary of a square matrix that gives us information about the rank of the matrix. The determinant of a square matrix $\mathbf{A}$ is denoted $|\mathbf{A}|$ or $\det(\mathbf{A})$.

- The determinant of a diagonal or triangular matrix equals the product of the diagonal values.

    Exercise:

    $$\left| \begin{bmatrix} 1 & 6 & 5 \\ 0 & 2 & 4 \\ 0 & 0 & 3 \end{bmatrix} \right| =$$

- For any 2×2 matrix,

    $$\left| \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right| = ad - bc.$$

- For any 3×3 matrix,

$$\left| \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \right| = a \left| \begin{bmatrix} e & f \\ h & i \end{bmatrix} \right| - d \left| \begin{bmatrix} b & c \\ h & i \end{bmatrix} \right| + g \left| \begin{bmatrix} b & c \\ e & f \end{bmatrix} \right|$$

$$= a(ei - hf) - d(bi - hc) + g(bf - ec).$$

Useful properties of determinants

$$| \mathbf{A}_{n \times n} | = 0 \quad \Leftrightarrow \quad \mathrm{rank}(\mathbf{A}) < n$$

$$\Leftrightarrow \quad \mathbf{A} \text{ is less than full rank}$$

$$\Leftrightarrow \quad \text{the } \textit{inverse} \text{ of } \mathbf{A} \text{ does not exist}$$

$$\Leftrightarrow \quad \text{the columns of } \mathbf{A} \text{ are linearly dependent,}$$

while

$$| \mathbf{A}_{n \times n} | \neq 0 \quad \Leftrightarrow \quad \mathrm{rank}(\mathbf{A}) = n$$

$$\Leftrightarrow \quad \mathbf{A} \text{ is full rank}$$

$$\Leftrightarrow \quad \text{the } \textit{inverse} \text{ of } \mathbf{A} \text{ exists}$$

$$\Leftrightarrow \quad \text{columns of } \mathbf{A} \text{ are linearly independent.}$$

For full rank matrices that conform, $| \mathbf{AB} | = | \mathbf{A} | | \mathbf{B} |$. In addition, $| \mathbf{A}' | = | \mathbf{A} |$.

Positive Definite and Semidefinite Matrices

Let $\mathbf{A}$ be an $n \times n$ symmetric matrix. $\mathbf{A}$ is *positive definite* if and only if

1. $a_{ii} > 0$ for all $i = 1, \ldots, n$

2. the determinant of every square submatrix of upper-left corner of $\mathbf{A}$ is positive. That is,

$$\left| \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \right| > 0,$$

$$\left| \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} \right| > 0,$$

$$\vdots$$

$$\mid \mathbf{A} \mid > 0.$$

The matrix $\mathbf{A}$ is *positive semidefinite* if we replace "> 0" in (1) and (2) with "≥ 0". A matrix is called *nonnegative definite* if it is positive definite or positive semidefinite.

Covariance matrices are nonnegative definite.

Exercise:

Let

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

Is $\mathbf{A}$ positive definite, positive semidefinite, or neither?

Inverses and Generalized Inverses

Suppose we have a system of equations like

$$(\mathbf{X}_{n\times p})'(\mathbf{X}_{n\times p})\widehat{\boldsymbol{\beta}} = (\mathbf{X}_{n\times p})'\mathbf{y}_{n\times 1},$$

where are the *normal equations* for the linear model. In order to solve these equations and obtain our estimate, $\widehat{\boldsymbol{\beta}}$, we would like to divide by $\mathbf{X}'\mathbf{X}$ in some sense. Because $\mathbf{X}'\mathbf{X}$ is a matrix, this presents us some difficulty. Thus we develop the idea of a matrix inverse.

Consider the $n \times n$ matrix $\mathbf{A}$. If $\mathbf{A}$ has full rank, then rank($\mathbf{A}$)=n, and there exists a unique matrix, $\mathbf{A}^{-1}$, called the *inverse* of $\mathbf{A}$, such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}.$$

Some properties of inverses

1. For a scalar, $\mathbf{A}_{1\times 1} = a$, $\mathbf{A}^{-1} = \frac{1}{a}$.

2. The inverse of a diagonal matrix is the diagonal matrix of reciprocals of the diagonal elements.

3. For conforming full rank matrices, $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

4. A symmetric matrix has a symmetric inverse.

5. The inverse of the transpose is the transpose of the inverse. That is, $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$. $(\mathbf{A}')^{-1}$ is also sometimes denoted $\mathbf{A}^{-T}$.

6. The determinant of the inverse is the inverse of the determinant. That is, $\mid \mathbf{A}^{-1} \mid = \frac{1}{|\mathbf{A}|}$.

7. The inverse of the 2×2 matrix $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ is given by

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}.$$

Generalized Inverses For certain linear models (some ANOVA models for example), we will want to obtain inverses of matrices that are not full rank. For any $\mathbf{A}_{r \times c}$, there exists a *generalized inverse*, denoted $\mathbf{A}_{c \times r}^{-}$, such that

$$\mathbf{A}_{r \times c} \mathbf{A}_{c \times r}^{-} \mathbf{A}_{r \times c} = \mathbf{A}_{r \times c}.$$

The *generalized inverse* is not unique.

The *Moore-Penrose generalized inverse (MPGI)*, $\mathbf{A}^{+}$, is a unique type of generalized inverse satisfying the following properties:

- $\mathbf{A}\mathbf{A}^{+}\mathbf{A} = \mathbf{A}$ (definition of generalized inverse)

- $\mathbf{A}^{+}\mathbf{A}\mathbf{A}^{+} = \mathbf{A}^{+}$ ($\mathbf{A}$ is a generalized inverse of $\mathbf{A}^{+}$)

- $(\mathbf{A}^{+}\mathbf{A})' = \mathbf{A}^{+}\mathbf{A}$ ($\mathbf{A}^{+}\mathbf{A}$ is symmetric)

- $(\mathbf{A}\mathbf{A}^{+})' = \mathbf{A}\mathbf{A}^{+}$ ($\mathbf{A}\mathbf{A}^{+}$ is symmetric)

Facts about the MPGI

1. For $\mathbf{A}_{r \times c}$ with $r \leq c$ and $\mathbf{A}$ of full row rank r,

$$\mathbf{A}^{+}_{c \times r} = \mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}.$$

2. For $\mathbf{A}_{r \times c}$ with $c \leq r$ and $\mathbf{A}$ of full column rank c,

$$\mathbf{A}^{+}_{r \times c} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'.$$

3. The Moore-Penrose generalized inverse of a less than full rank diagonal matrix is the diagonal matrix with reciprocals of the nonzero elements on the diagonal in the same locations as the nonzero elements, and zero elsewhere.

Both the Moore-Penrose generalized inverse and other generalized inverses reduce to the regular inverse when the matrix of interest is square and full rank.

Eigenvalues, Eigenvectors, and the Spectral Decomposition

Eigenanalysis is defined only for square matrices. Most interest in decomposing matrices in statistics lies with symmetric matrices, for example, covariance matrices.

Suppose $\mathbf{A}$ is an $n \times n$ matrix (not necessarily symmetric). A *right eigenvector* of $\mathbf{A}$ is any nonzero $n \times 1$ vector $\mathbf{x}$ satisfying

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x},$$

where λ is the *eigenvalue* corresponding to $\mathbf{x}$. Eigenvalues and eigenvectors are also called *characteristic values* and *characteristic vectors* (*eigen* is German for *characteristic*).

Eigenvectors are not unique (prove it?), the convention is to scale the eigenvector $\mathbf{x}$ so that $\mathbf{x}'\mathbf{x} = 1$, normalizing it to unit length.

Finding Eigenvectors and Eigenvalues

Using the definition of eigenvectors, we can find the *characteristic equation* that is used to find eigenvalues.

$$\mathbf{Ax} = \lambda \mathbf{x}$$

$$\mathbf{Ax} - \lambda \mathbf{x} = \mathbf{0}$$

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}$$

$$\mid \mathbf{A} - \lambda \mathbf{I} \mid = \mathbf{0}.$$

The last step follows because $\mathbf{x}$ is a nonzero vector.

The characteristic equation of an $(n \times n)$ matrix equals an nth degree polynomial in λ. An nth degree polynomial has n roots, so a $(n \times n)$ matrix has n eigenvalues. For a (2×2) matrix,

$$| \mathbf{A} - \lambda \mathbf{I} | = \left| \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$\left| \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = \left| \begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} \right| = 0$$

$$(a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0$$

$$\lambda^2 - \lambda(a_{11} + a_{22}) + a_{11}a_{22} - a_{12}a_{21} = 0.$$

Once the eigenvalues are found, eigenvectors corresponding to these eigenvalues may be found using the equation $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$.

Example:

Suppose

$$\mathbf{A} = \begin{bmatrix} 5 & -3 \\ 4 & -2 \end{bmatrix}.$$

The characteristic equation is

$$
\begin{aligned}
\mid \mathbf{A} - \lambda\mathbf{I} \mid \ &= \ \left| \begin{bmatrix} 5 - \lambda & -3 \\ 4 & -2 - \lambda \end{bmatrix} \right| \\
&= \ (5 - \lambda)(-2 - \lambda) - (-3)(4) \\
&= \ \lambda^2 - 3\lambda + 2 \\
&= \ (\lambda - 2)(\lambda - 1) = 0,
\end{aligned}
$$

so the eigenvalues of $\mathbf{A}$ are 2 and 1.

We find eigenvectors corresponding to these eigenvalues by solving $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ for $\lambda = 1$ and $\lambda = 2$.

For $\lambda = 1$, we have

$$
\begin{bmatrix} 5 & -3 \\ 4 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},
$$

which leads to the system of equations

$$
5x_1 - 3x_2 = x_1
$$

$$
4x_1 - 2x_2 = x_2.
$$

Rearranging both equations, we see $x_1 = \frac{3}{4}x_2$. So one eigenvector corresponding to the eigenvalue 1 is $(3, 4)'$.

To normalize this vector, we take

$$
\begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = 9 + 16 = 25
$$

and divide by the square root of this number. So the normalized eigenvector corresponding to the eigenvalue 1 is $(\frac{3}{5}, \frac{4}{5})'$.

Similarly, we can show the normalized eigenvector corresponding to $\lambda = 2$ is $\frac{1}{\sqrt{2}}(1,1)'$.

Some Properties of Eigenvalues and Eigenvectors

- For $\mathbf{A}_{n \times n}$, the number of distinct eigenvalues ranges from 1 to n.

- The trace of a matrix is the sum of the eigenvalues. That is, $\text{trace}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$.

- The determinant of a matrix is the product of its eigenvalues. That is, $\mid \mathbf{A} \mid = \prod_{i=1}^{n} \lambda_i$.

- $\mathbf{A}$ full rank $\Leftrightarrow \mathbf{A}$ has no zero eigenvalues

- $\mid \mathbf{A} \mid = 0 \Leftrightarrow$ at least one eigenvalue is zero $\Leftrightarrow \mathbf{A}$ is not full rank

- The number of nonzero eigenvalues of $\mathbf{A}$ is $\text{rank}(\mathbf{A})$.

- Small eigenvalues imply that there are near-linear dependencies in the columns of a matrix $\mathbf{A}$.

- $\mathbf{A}$ is positive definite if $\min(\lambda_i) > 0$

- $\mathbf{A}$ is positive semidefinite if $\min(\lambda_i) \geq 0$

Spectral Decomposition

It is sometimes easier to deal with matrices if we write them as a product of more simple matrices that have special structures. Matrix decomposition allows us to write matrices as a product of simpler matrices. We will see one such decomposition, the *spectral decomposition*, later in the course.

The *spectral decomposition* allows us to write any symmetric matrix in terms of an orthogonal matrix and a diagonal matrix of eigenvalues.

Spectral Theorem: Suppose $\mathbf{A}$ is an $(n \times n)$ symmetric matrix. Then there exists an orthogonal (column orthonormal) matrix $\mathbf{V}$ such that

$$\mathbf{A} = \mathbf{V \Lambda V'},$$

where $\mathbf{\Lambda} = \text{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is an $(n \times n)$ diagonal matrix of the ordered eigenvalues of $\mathbf{A}$ so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. $\mathbf{V}$ is the orthogonal matrix of eigenvectors corresponding to the eigenvalues of $\mathbf{A}$. The eigenvalues and eigenvectors must be in the same order.

Example: Verify that the spectral decomposition of

$$\mathbf{A} = \begin{bmatrix} 1.5 & -0.5 \\ -0.5 & 1.5 \end{bmatrix}$$

is given by $\mathbf{A} = \mathbf{V \Lambda V'}$, where

$$\mathbf{V} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}, \text{ and } \mathbf{\Lambda} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$

Once a matrix is in spectral decomposition form, it is easy to obtain the inverse.

- For $\mathbf{A}$ full rank, we have $\mathbf{A}^{-1} = \mathbf{V}(\mathbf{\Lambda})^{-1}\mathbf{V}'$.

- For $\mathbf{A}$ less than full rank, we have $\mathbf{A}^{+} = \mathbf{V}(\mathbf{\Lambda})^{+}\mathbf{V}'$.

Thus computing the reciprocals of the eigenvalues allows us to find the inverse or Moore-Penrose generalized inverse of a symmetric matrix in this form.

We will see the spectral decomposition again in Chapter 8.

Singular Value Decomposition (SVD)

The *singular value decomposition* gives us a more accurate way to find the inverse of an ill-conditioned matrix. Both SAS and S-plus use the SVD when fitting linear models. The SVD is valid for *any* matrix, while the spectral decomposition is valid only for symmetric matrices.

For *any* matrix $\mathbf{A}_{m \times n}$ with $m \geq n$, there exist orthogonal matrices $U_{m \times m}$ and $V_{n \times n}$ along with an $(m \times n)$ matrix $\mathbf{S}$ such that

$$\mathbf{A}_{m \times n} = \mathbf{U}_{m \times m} \mathbf{S}_{m \times n} \mathbf{V}'_{n \times n}.$$

We define

$$\mathbf{S}_{m \times n} = \begin{bmatrix} \text{Diag}(\mathbf{s}) \\ \mathbf{0}_{(\mathbf{m}-\mathbf{n}) \times \mathbf{n}} \end{bmatrix},$$

where the vector $\mathbf{s}_{n \times 1}$ contains the n *singular values* of $\mathbf{A}$. The rank r of $\mathbf{A}$ is the number of nonzero singular values of $\mathbf{A}$. Singular values are computed as the positive square roots of the eigenvalues of $\mathbf{A}'\mathbf{A}$.

Facts about the SVD

- $\mathbf{U}'\mathbf{U} = I_m \qquad \mathbf{V}'\mathbf{V} = I_n$

- $\mathbf{A}'\mathbf{A} = \mathbf{V}\text{Diag}(\mathbf{s}^2)\mathbf{V}'$ (note: $\text{Diag}(\mathbf{s}^2)$ here means $\text{Diag}(\{s_i^2\})$)

- $\mathbf{A}\mathbf{A}' = \mathbf{U}\text{Diag}(\mathbf{s}^2, \mathbf{0_{m-n}})\mathbf{U}'$

- The Moore-Penrose generalized inverse of $\mathbf{A}$ may be written

$$
\begin{aligned}
\mathbf{A}^+ &= \mathbf{V}\mathbf{S}^+\mathbf{U}' \\
&= \mathbf{V}\left[\text{Diag}(s)^+ \quad \mathbf{0_{n\times(m-n)}}\right]\mathbf{U}',
\end{aligned}
$$

  where $\text{Diag}(s)^+ = \text{Diag}(\{s_1^{-1}, s_2^{-1}, \ldots, s_r^{-1}, 0, \ldots, 0\})$.

- SVD is also widely used in statistics where it is related to principal component analysis, and in signal processing and pattern recognition.

Random Vectors and Matrices

Suppose

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

is a vector of random variables with $E(Y_i) = \mu_i$, $\mathsf{Var}(Y_i) = \sigma_{ii}$, and $\mathsf{Cov}(Y_i, Y_j) = \sigma_{ij}$.

The expectation of the random vector $\mathbf{Y}$ is defined

$$E(\mathbf{Y}) = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu}.$$

Suppose $\mathbf{Z}$ is an $(n \times p)$ matrix of random variables. Then

$$E(\mathbf{Z}) = \begin{pmatrix} E(Z_{11}) & \dots & E(Z_{1p}) \\ \vdots & \dots & \vdots \\ E(Z_{n1}) & \dots & E(Z_{np}) \end{pmatrix}.$$

Thus the expectation of a random matrix is the matrix of the expectations.

For $\mathbf{Y}$ an $(n \times 1)$ random vector, the *covariance matrix* of $\mathbf{Y}$ is

$$
\begin{aligned}
\mathsf{Cov}(\mathbf{Y}) &= E\left[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})'\right] \\
&= \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ & & \vdots & \vdots & \vdots \\ \sigma_{21} & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \cdots & \cdots & \sigma_{nn} \end{pmatrix},
\end{aligned}
$$

where $\sigma_{ij} = E[(Y_i - \mu_i)(Y_j - \mu_j)']$, $i, j = 1, \ldots, n$.

Suppose $\boldsymbol{\mu} = E(\mathbf{Y})$ and covariance matrix $\boldsymbol{\Sigma} = \mathsf{Cov}(\mathbf{Y})$. In addition, suppose $\mathbf{A}_{r \times n}$ is a matrix of constants and $\mathbf{b}_{r \times 1}$ is a vector of constants. Then

$$
E(\mathbf{A}\mathbf{Y} + \mathbf{b}) = \mathbf{A}E(\mathbf{Y}) + \mathbf{b} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}
$$

$$
\mathsf{Cov}(\mathbf{A}\mathbf{Y} + \mathbf{b}) = \mathbf{A}\mathsf{Cov}(\mathbf{Y})\mathbf{A}' = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'.
$$

Let $\mathbf{W}_{r \times 1}$ be a random vector with $E(\mathbf{W}) = \boldsymbol{\gamma}$. Then

$$\mathrm{Cov}(\mathbf{W}, \mathbf{Y}) = E\left[(\mathbf{W} - \boldsymbol{\gamma})(\mathbf{Y} - \boldsymbol{\mu})'\right],$$

where $\mathrm{Cov}(\mathbf{W}, \mathbf{Y})$ is an $(r \times n)$ matrix of covariances with ij^{th} element equal to $\mathrm{Cov}(W_i, Y_j)$.

Important Distributions for Linear Models

The theory of linear models involves many statistical distributions, including the following distributions:

1. Gaussian (Normal) Distribution

2. Multivariate Normal Distribution

3. Chi-Squared Distribution

4. Non-Central Chi-Squared Distribution

5. t Distribution

6. Non-Central t Distribution

7. F Distribution

8. Non-Central F Distribution.

We will now discuss many of these distributions in detail.
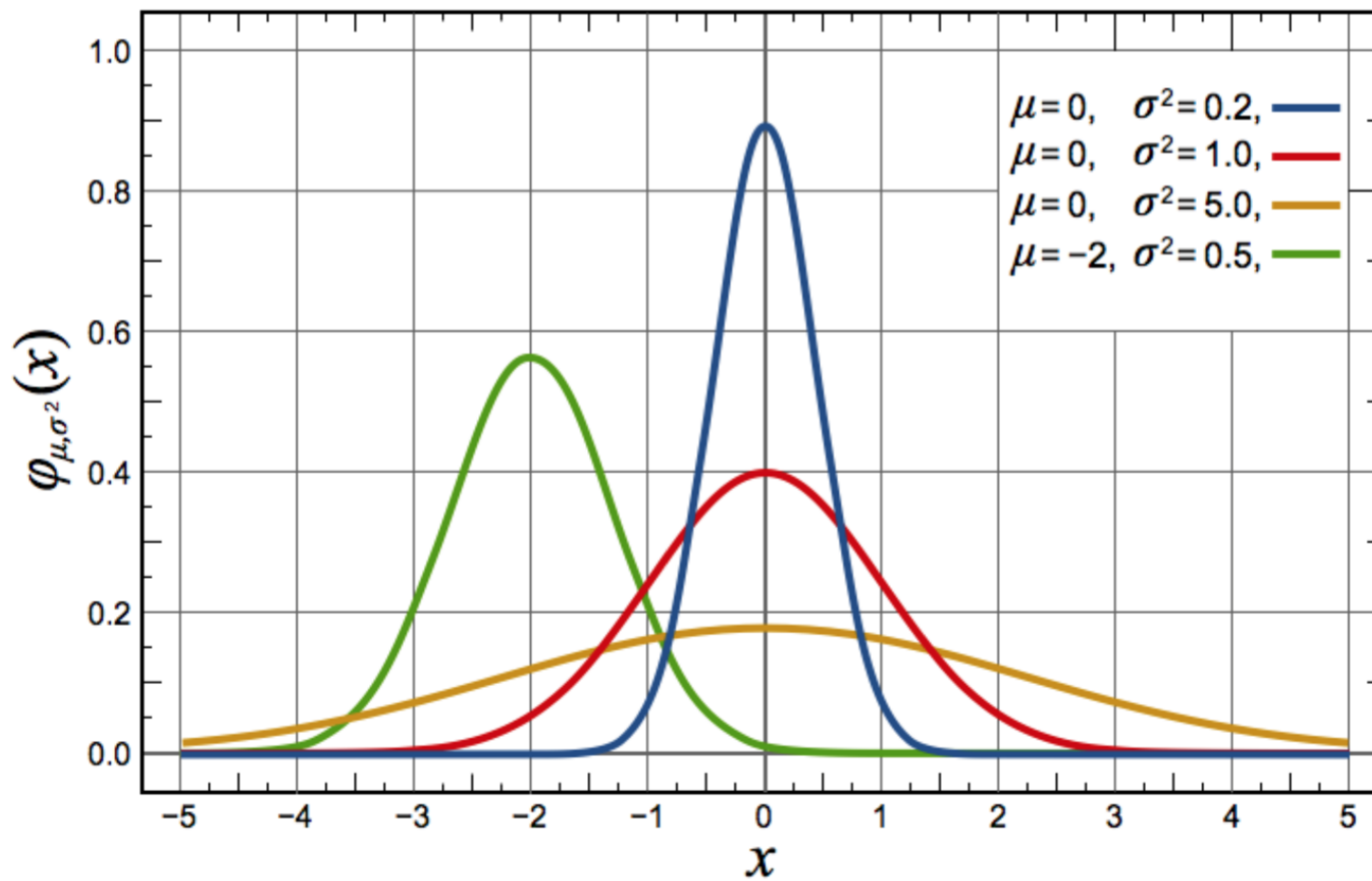
Gaussian (Normal) Distribution

A random variable X has a Gaussian (normal) distribution with mean μ and variance σ^2, written $X \sim N(\mu, \sigma^2)$, if X has density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}.$$

When $\mu = 0$ and $\sigma^2 = 1$, then we say $X \sim N(0, 1)$ has a *standard normal distribution*.

If $X \sim N(\mu, \sigma^2)$, then $\frac{x-\mu}{\sigma} \sim N(0, 1)$. For the standard normal distribution, 95% of the probability mass falls between -1.96 and 1.96 (roughly 2 standard deviations of the mean). Much of hypothesis testing is based on this fact.

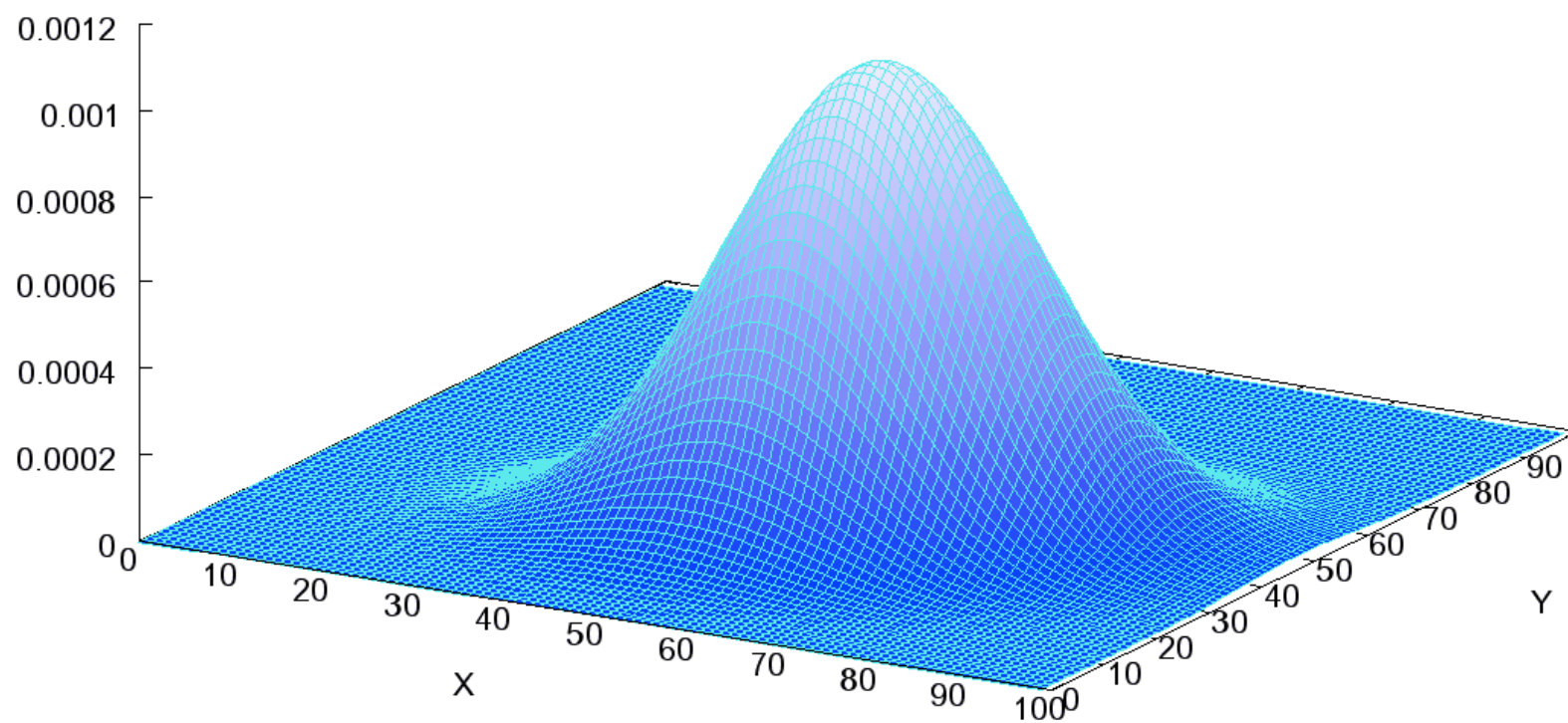The normal distribution is symmetric about its mean.

Multivariate Normal Distribution

Suppose $X = (X_1, \ldots, X_n)'$. Then X has an n dimensional multivariate normal distribution with mean μ and covariance matrix Σ if X has density

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid \Sigma \mid^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(x - \mu)'\Sigma^{-1}(x - \mu) \right\}.$$

We write this $X \sim N_n(\mu, \Sigma)$.

This definition requires Σ to be positive definite.

Facts about the multivariate normal distribution

1. A linear transformation of a multivariate normal distribution yields another multivariate normal distribution. Suppose $X \sim N_n(\mu, \Sigma)$. For $A_{r \times n}$ a matrix of constants and $b_{r \times 1}$ a vector of constants, then $Y = AX + b$ has the multivariate normal distribution given by $Y \sim N_r(A\mu + b, A\Sigma A')$.

2. A linear combination of independent multivariate normal distributions is a multivariate normal distribution. Suppose $X_1, \ldots, X_k$ are independent with $X_i \sim N_n(\mu_i, \Sigma_i)$, $i = 1, \ldots, k$. Suppose $a_1, \ldots, a_k$ are scalars and define

    $$Y = a_1 X_1 + \ldots + a_k X_k.$$

    Then $Y \sim N(\mu^*, \Sigma^*)$, where $\mu^* = \sum_{i=1}^{k} a_i \mu_i$ and $\Sigma^* = \sum_{i=1}^{k} a_i^2 \Sigma_i$.

3. Marginal distributions of a multivariate normal distribution are multivariate normal distributions. Suppose $X \sim N_n(\mu, \Sigma)$.

Partition X into $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ where X_1 is $r \times 1$ and X_2 is

$(n - r) \times 1$. Partition μ as $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ where μ_1 is $r \times 1$ and μ_2 is

$(n - r) \times 1$. Similarly, partition Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where Σ_{11} is $r \times r$, Σ_{12} is $r \times (n - r)$, $\Sigma_{21} = \Sigma'_{12}$ is $(n - r) \times r$, and Σ_{22} is $(n - r) \times (n - r)$. Then the marginal distribution of X_1 is given by $X_1 \sim N_r(\mu_1, \Sigma_{11})$, and the marginal distribution of X_2 is given by $X_2 \sim N_{(n-r)}(\mu_2, \Sigma_{22})$.

4. Conditional distributions of multivariate normal distributions are multivariate normal distributions. Suppose that $X \sim N_n(\mu, \Sigma)$.

Using the same partition as above, then we have

$$
\begin{aligned}
X_1 \mid X_2 \;\; &= \;\; x_2 \\
&\sim \;\; N_r\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma^*\right),
\end{aligned}
$$

where $\Sigma^* = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.
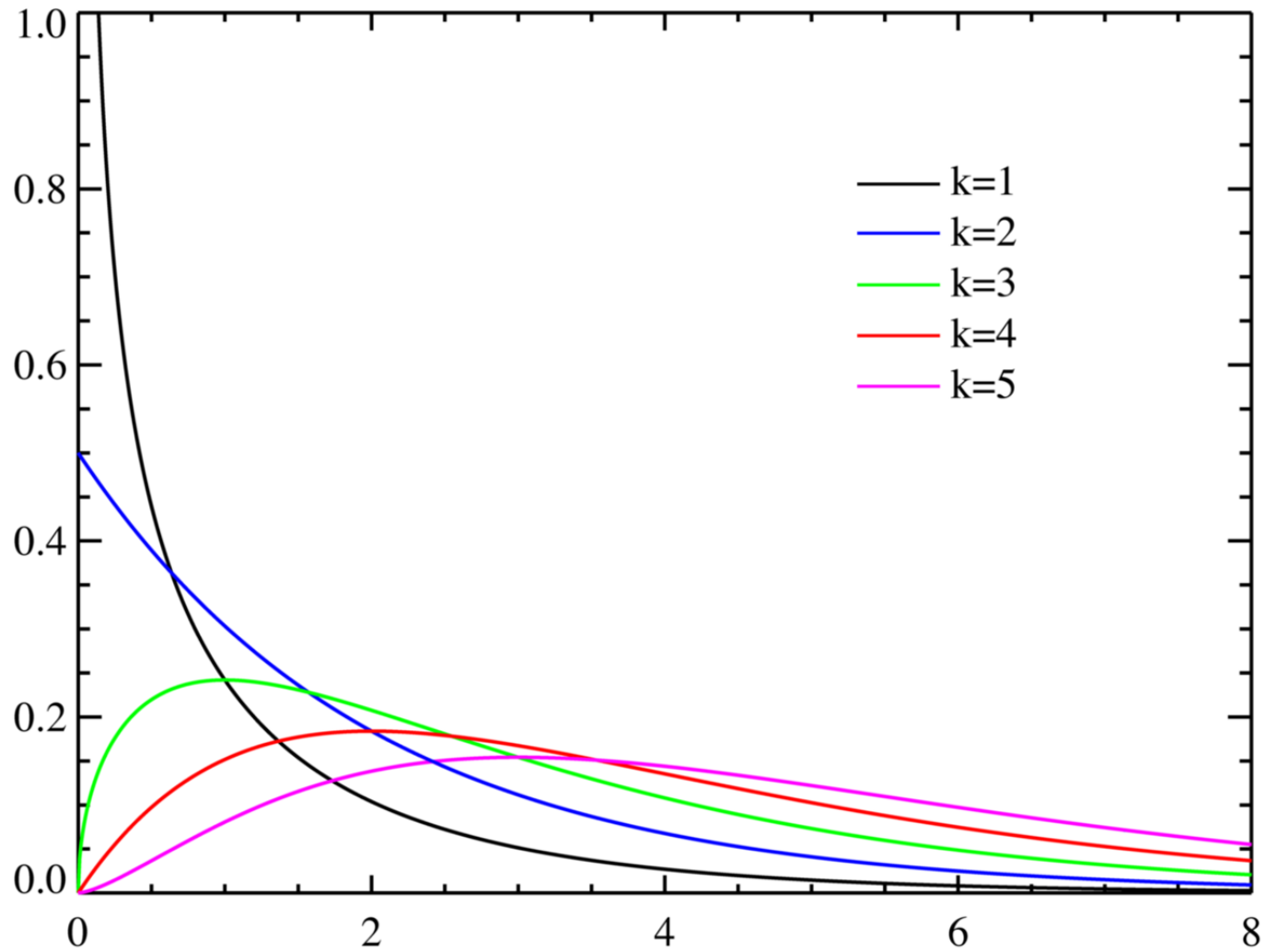
Chi-Squared Distribution

A random variable X has a central chi-squared distribution with n *degrees of freedom*, written $X \sim \chi^2(n)$, if the density of X is given by

$$f(x) = \left(\frac{1}{\Gamma(\frac{n}{2})} \right) \left(\frac{1}{2} \right)^{\frac{n}{2}} x^{\frac{n}{2}-1} \exp\left\{ -\frac{x}{2} \right\},$$

where $\Gamma(a)$ is the complete gamma function, given by

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx.$$

The chi-squared distribution is asymmetric and restricted to positive numbers. Its degrees of freedom determine the mean and variance of the distribution.

The chi-squared distribution is related to the normal distribution. If the random variable $Z \sim N(0,1)$, then $Z^2 \sim \chi^2(1)$. In addition, if $Z_1, Z_2, \ldots, Z_n$ are independent, identically distributed $N(0,1)$ random variables, then $W = \sum_{i=1}^{n} Z_i^2$ has a chi-squared distribution with n degrees of freedom; that is, $W \sim \chi^2(n)$.

The mean of a $\chi^2(n)$ distribution is n, and its variance is $2n$.

If $Z \sim N(\mu, 1)$, then Z^2 follows a non-central chi-squared distribution with 1 degrees of freedom and non-centrality parameter μ^2.

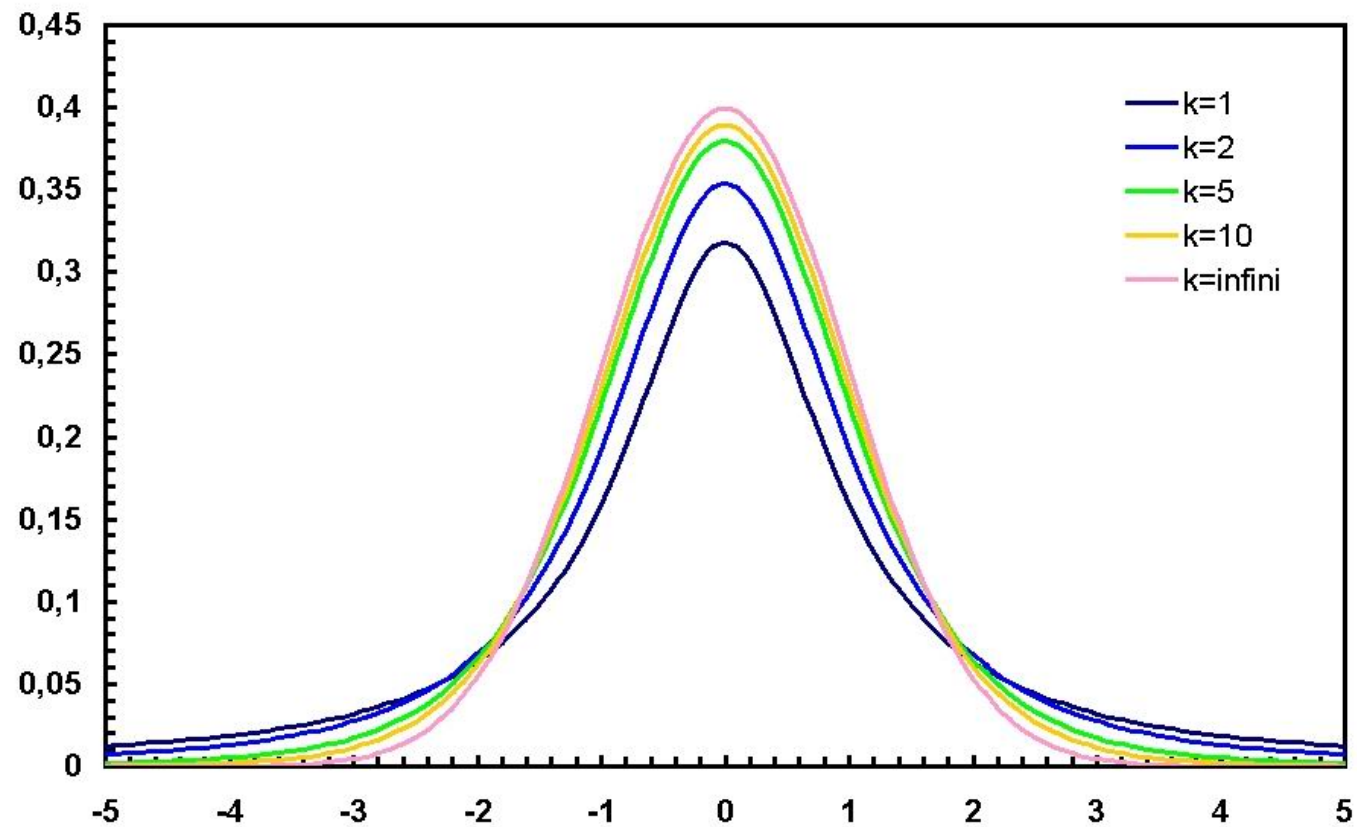The chi-squared distribution is used widely in the analysis of categorical data.

Student's t Distribution

Suppose $X \sim N(0, 1)$, $Y \sim \chi^2(n)$, with X and Y independent. The random variable

$$T = \frac{X}{\sqrt{\frac{Y}{n}}}$$

has a t distribution with n degrees of freedom. We write this as $T \sim t(n)$. Like the standard normal distribution, the t distribution is symmetric about 0.

If $X \sim N(\mu, 1)$, then T has a non-central t-distribution with n degrees of freedom and non-centrality parameter μ.

The *degrees of freedom* n determines the amount of variability in the distribution. As the number of degrees of freedom increases, the variability of the t distribution decreases. In fact, as the number of degrees of freedom gets large, the t distribution approximates the standard normal distribution. With smaller degrees of freedom, the t distribution resembles a normal distribution with fatter tails.

A $t(1)$ distribution, which has 1 degree of freedom, is not well-behaved and is called a *Cauchy distribution*.

Fisher's F Distribution

Suppose $X_1 \sim \chi^2(n_1)$ and $X_2 \sim \chi^2(n_2)$, and X_1 and X_2 are independent. The random variable

$$F = \frac{\left(\frac{X_1}{n_1}\right)}{\left(\frac{X_2}{n_2}\right)}$$

has a central F distribution with (n_1, n_2) degrees of freedom. We write this $F \sim F(n_1, n_2)$.

We call n_1 the *numerator* degrees of freedom and n_2 the *denominator* degrees of freedom. The central F distribution is used in hypothesis tests of nested linear models. If $T \sim t(\nu)$, then $T^2 \sim F(1, \nu)$. The F distribution is asymmetric and restricted to positive numbers.

If $X_1 \sim \chi^2(n_1; \mu^2)$, then F follows a non-central F distribution.
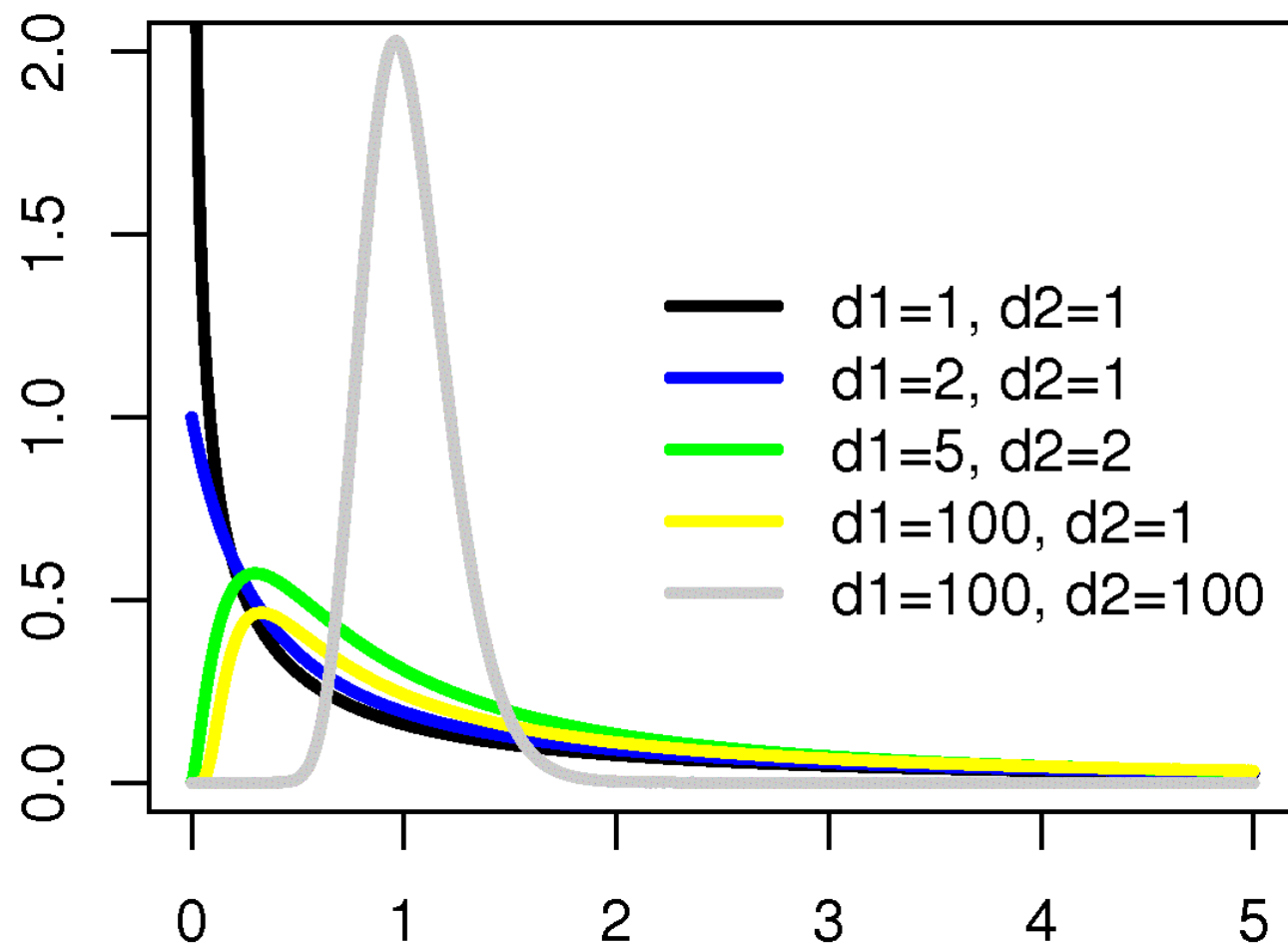
Maximum Likelihood Estimation

Maximum likelihood estimates (MLE's) have excellent large-sample properties and are applicable in a wide variety of situations. Examples of maximum likelihood estimates include the following:

- The sample average $\bar{X}$ of a group of independent and identically normally distributed observations $X_1, \ldots, X_n$ is a MLE.

- Parameter estimates in a linear regression model fit to normally distributed data are maximum likelihood estimates.

- Parameter estimates in a logistic regression model are maximum likelihood estimates.

- The estimate $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$ of the variance of a group of independent and identically normally distributed observations is *not* a MLE. (The MLE of the variance is $\left(\frac{n-1}{n}\right) s^2$.)

Finding Maximum Likelihood Estimates

Let $L(\mathbf{Y} \mid \boldsymbol{\theta})$ denote the *likelihood function* for $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ from some population described by the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$.

The maximum likelihood estimate of $\boldsymbol{\theta}$ is given by the estimator $\widehat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p)$ for which

$$L(\mathbf{Y} \mid \widehat{\boldsymbol{\theta}}) > L(\mathbf{Y} \mid \boldsymbol{\theta}^*),$$

where $\boldsymbol{\theta}^*$ is any other estimate of $\boldsymbol{\theta}$.

Thus the maximum likelihood estimate is the most "probable" or "likely" for the data.

Maximizing the likelihood $L(\mathbf{Y} \mid \boldsymbol{\theta})$ is equivalent to maximizing the natural logarithm $\ln(L(\mathbf{Y} \mid \boldsymbol{\theta})) = \ell(\mathbf{Y} \mid \boldsymbol{\theta})$, called the log-likelihood.

The maximum likelihood estimates are typically found as the solutions of the p equations obtained by setting the p partial derivatives of $\ell(\mathbf{Y} \mid \boldsymbol{\theta})$ with respect to each $\boldsymbol{\theta}_j$, $j = 1, \ldots, p$, equal to zero.

Why do we solve the derivatives for zero? The derivative gives us the slope of the likelihood (or log-likelihood), and when the slope is zero, we know that we are at either a local minimum or local maximum. (The second derivative is negative for a maximum and positive for a minimum.)
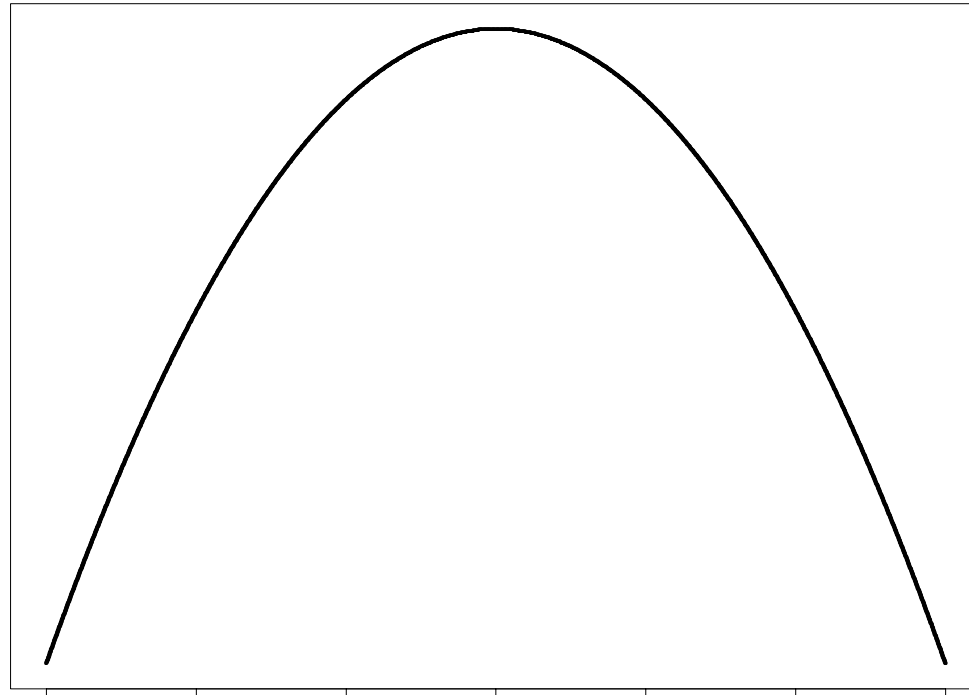
When closed form expressions for maximum likelihood estimates do not exist, computer algorithms may be used to solve for the estimates.

Example:

Let Y_i, $i = 1, \ldots, n$ be i.i.d.normal random variables with mean μ and variance σ^2, so $Y_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$.

The density of Y_i is given by

$$f(Y_i \mid \mu, \sigma^2) = (2\pi)^{-\frac{1}{2}} (\sigma^2)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2} (Y_i - \mu)^2 \right\}.$$

Find the maximum likelihood estimates of μ and σ^2.

Hypothesis Testing and Interval Estimation with MLE's

It can be shown (based on large-sample properties of MLE's) that

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{\mathsf{Var}}(\hat{\beta}_1)}}$$

is approximately $N(0, 1)$ when the sample size is large.

A test of $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$ can be based on the Z statistic $\frac{\hat{\beta}_1}{\sqrt{\widehat{\mathsf{Var}}(\hat{\beta}_1)}}$, which has approximately a $N(0, 1)$ distribution under H_0. This test is called a *Wald test*.

By a similar argument, an approximate $100(1 - \alpha)\%$ large-sample confidence interval for β_1 takes the form

$$\hat{\beta}_1 \pm Z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\mathsf{Var}}(\hat{\beta}_1)},$$

where $Pr(Z > Z_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2}$ when $Z \sim N(0, 1)$.

NOTE: Testing σ^2 is more complicated, and the Wald test for the hypothesis $\sigma^2 = 0$ is not recommended because the value $\sigma^2 = 0$ is on the boundary of the parameter space for σ^2.

Next: Simple Linear Regression

Reading Assignment:

- Weisberg Chapter 2: Simple Linear Regression