Review of Linear Algebra for Statistics

Brian Vegetable

2017 Statistics Bootcamp
Department of Statistics
University of California, Irvine

September 22nd, 2016
Overview I

• We wrap up the math topics by reviewing some linear algebra concepts
• Linear algebra will become an important tool for you as a statistician
• You’ll be using matrix operations most of the year, but the main necessity for linear algebra will come in STAT 200C.
Overview II

• Here are a few good references for reviewing undergraduate linear algebra in general
  • Introduction to Linear Algebra by Gilbert Strang
  • Gilbert Strang’s Lectures on YouTube (https://www.youtube.com/watch?v=ZK30402wf1c)
  • Linear Algebra and it’s Applications by David Lay
  • Linear Algebra by Friedberg, Insel, Spence (Upper division text)

• Graduate Level Linear Algebra References for Statistics
  • Matrix Algebra from a Statistician’s Perspective by David Harville
  • Appendix of Linear Regression Analysis by George Seber and Alan Lee
  • Appendix of Applied Linear Regression by Sanford Weisberg
Motivation I

- A familiarity with matrices will allow you to expand the types of statistics you can do.
- Consider the multivariate normal distribution
  \[ X = (X_1, X_2, \ldots, X_n)^T \]

  \[
  f(x) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)
  \]

  which is said to be “non-degenerate” when \( \Sigma \) is positive-definite.
- Additionally, \( x \) is a real-valued \( n \)-dimensional column vector and \( |\Sigma| \) is the determinant of \( \Sigma \)
- To investigate many of the properties of this distribution we’ll need matrix algebra
Motivation II

- We’ll specifically use this distribution to explore linear regression
- Let $Y$ be a random variable which has some mean $\mu$ which we measure under error, $\epsilon$, specifically
  \[ Y = \mu + \epsilon \]

- We will focus on linear models where
  \[ \mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} \]
  where $x$ are explanatory variables and each $\beta_j$ is unknown and to be estimated
Motivation III

- If we consider a random sample of $n$ observations we will have

$$
\begin{pmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{pmatrix} = \begin{pmatrix}
x_{10} & x_{11} & \cdots & x_{1,p-1} \\
x_{20} & x_{21} & \cdots & x_{2,p-1} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n0} & x_{n1} & \cdots & x_{n,p-1}
\end{pmatrix} \begin{pmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_{p-1}
\end{pmatrix} + \begin{pmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_n
\end{pmatrix}
$$

- Or more simply written

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

- We will eventually show that $\mathbf{Y} \sim N_n(\mathbf{X}\beta, \Sigma)$.
- Matrix algebra will play a very important role throughout understanding linear algebra
Defining a Matrix

- A rectangular array of real numbers is called a matrix.

\[
\begin{pmatrix}
a_{11} & a_{12} & \ldots & a_{1n} \\
a_{21} & a_{22} & \ldots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1} & a_{m2} & \ldots & a_{mn}
\end{pmatrix}
\]

- A matrix with \( m \) rows and \( n \) columns is referred to as an \( m \times n \) matrix.

- Matrices will often be denoted by boldface letters \( \mathbf{X} \).

- Additionally we can denote a matrix \( \mathbf{X} = \{a_{ij}\} \)
Basic Matrix Operations I

- **Scalar Multiplication:** Consider a matrix $A$ and a scalar $k$, then
  \[ kA = k\{a_{ij}\} = \{ka_{ij}\} \]

- **Matrix Addition:** Consider two matrices $A$ and $B$, if they are both of dimension $m \times n$ then we define addition between these two matrices. Specifically $A + B$ is the $m \times n$ matrix $\{a_{ij} + b_{ij}\}$ for all pairs $i, j$.
  - Matrix addition is commutative and associative
  - Additionally matrices having the same number of rows and columns are said to be conformal for addition (or subtraction).
Basic Matrix Operations II

- **Matrix Multiplication:** Let $A = \{a_{ij}\}$ represent an $m \times n$ matrix and $B = \{b_{ij}\}$ a $p \times q$ matrix. When $n = p$ (when $A$ has the same number of columns as $B$ has rows), then the matrix product $AB$ is defined to be the $m \times q$ matrix whose $ij^{th}$ element is

$$
\sum_{k=1}^{n} a_{ik} b_{kj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}
$$

- The formation $AB$ is called the premultiplication of $B$ by $A$ or the postmultiplication of $A$ by $B$.
- When $n \neq p$ then the matrix product $AB$ is undefined.
- Two $n \times n$ matrices $A$ and $B$ are said to commute if $AB = BA$.
Basic Matrix Operations III

- **Matrix Transpose:** The transpose of an $m \times n$ matrix $A$, to be denoted $A^T$ or $A'$, is the $n \times m$ matrix whose $i^{th}$ element is the $j^{th}$ element of $A$.
  - For any matrix $A$, $(A')' = A$
  - For any two matrices $A$ and $B$ which are conformal for addition,
    \[(A + B)' = A' + B'\]
  - Finally any two matrices $A$ and $B$ for which the product is defined, \[(AB)' = B'A'\]
Vectors

- A matrix with only one column

\[
\begin{pmatrix}
a_1 \\
a_2 \\
\vdots \\
a_m
\end{pmatrix}
\]

is called an \(m\)-dimensional column vector

- A matrix with only one row is called a row vector

- Vectors will often be denoted by lower case bold symbols \(\mathbf{x}\).

- Clearly the transpose of an \(m\)-dimensional column vector is an \(m\)-dimensional row vector
Square Matrices

- One of the most important types of matrices in all of statistics is the square matrix.
- A matrix having the same number of rows as it does columns is called a square matrix.
- An $n \times n$ square matrix is said to have order $n$.

$$
\begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{n1} & a_{n2} & \cdots & a_{nn}
\end{pmatrix}
$$

- The set of terms $\{a_{ii}\}$ are called the diagonal elements of the square matrix and the terms $\{a_{ij}\}, i \neq j$ are the off-diagonal terms.
Symmetric Matrices

• A matrix \( A \) is said to be symmetric is \( A' = A \)

• Thus a symmetric matrix is a square matrix where the \( i^j \)th element equals the \( j^i \)th element.

\[
\begin{pmatrix}
5 & 4 & 0 \\
4 & -10 & -2 \\
0 & -2 & 3
\end{pmatrix}
\]
Diagonal Matrix

- A diagonal matrix is a square matrix whose off-diagonal elements are zero, that is
  \[
  \begin{pmatrix}
    d_1 & 0 & \cdots & 0 \\
    0 & d_2 & \cdots & 0 \\
    \vdots & \vdots & \ddots & \vdots \\
    0 & 0 & \cdots & d_n
  \end{pmatrix}
  \]

- The effect of premultiplying an $m \times n$ matrix $A$ by a $m \times m$ diagonal matrix $D$, $DA$ is to multiply each element of the $i^{th}$ row of $A$ by the element $d_{ii}$.
Identity Matrix

- Often the most useful diagonal matrix is the identity matrix $I_n$ where the subscript $n$ denotes the dimension of the identity matrix ($n \times n$). That is,

$$
I_n = \begin{pmatrix}
1 & 0 & \ldots & 0 \\
0 & 1 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 1
\end{pmatrix}
$$

Often the subscript $n$ is dropped.

- An important property is

$$
IA = AI = A
$$
Matrix Inversion I

- For any scalar $c$ there is a number called the inverse of $c$, say $d$ such that the product of $cd = 1$.
  - For example, if $c = 3$, then $d = 1/c = 1/3$, and the inverse of 3 is 1/3.
- This can be extended to square matrices

**Definition (Matrix Inverse)**

An $n \times n$ square matrix $A$ is called invertible (also nonsingular and non-degenerate) if there exists an $n \times n$ square matrix $B$ such that

$$AB = BA = I_n$$

If this is the case, then the matrix $B$ is uniquely determined by $A$ and is called the inverse of $A$ denoted $A^{-1}$.
Matrix Inversion II

- The collection of matrices that have an inverse are called full rank, invertible, or nonsingular.
- A square matrix that is not invertible, is of less than full rank or singular.
- The identity matrix is its own inverse \((I_n)^{-1} = I_n\).
Inverting a $2 \times 2$ Matrix. 1

• Consider the following matrix

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

• the inverse of $A$ denoted $A^{-1}$ is

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

where the determinant of $A$, $|A| = a_{11}a_{22} - a_{12}a_{21}$

• By our previous definitions we should have that $AA^{-1} = I$
Inverting a $2 \times 2$ Matrix. II

\[
\mathbf{A} \mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}
\]

\[
= \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{11}a_{22} - a_{12}a_{21} & -a_{11}a_{12} + a_{12}a_{21} \\ a_{21}a_{22} - a_{22}a_{21} & -a_{21}a_{12} + a_{22}a_{21} \end{pmatrix}
\]

\[
= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}
\]

- This satisfies our requirement
Orthogonality

- Two vectors $\mathbf{a}$ and $\mathbf{b}$ (of the same length), are orthogonal if

$$\mathbf{a}'\mathbf{b} = 0$$

- An $r \times c$ matrix $\mathbf{Q}$ has orthonormal columns if its columns, viewed as a set $c \leq r$ different $r \times 1$ vectors, are orthogonal and in addition have length 1.

- This is equivalent to

$$\mathbf{Q}'\mathbf{Q} = \mathbf{I}$$

- Additionally a square matrix $\mathbf{A}$ is orthogonal if

$$\mathbf{A}'\mathbf{A} = \mathbf{A}\mathbf{A}' = \mathbf{I}$$

so $\mathbf{A}^{-1} = \mathbf{A}'$. 
• Consider an $n \times p$ matrix $\mathbf{X}$ with columns given by the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ (we only consider the case when $p \leq n$).

• We say that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ are linearly dependent if we can find multipliers $a_1, \ldots, a_p$ not all equal to 0, such that

$$
\sum_{i=1}^{p} a_i \mathbf{x}_i = 0
$$
• If no such multipliers exist, then we say the vectors are linearly independent, and the matrix is full-rank.
• In general the rank of a matrix is the maximum number of $x_i$ which form a linearly independent set.
• The matrix $X'X$ is a $p \times p$ matrix.
  • If $X$ has rank $p$, so does $X'X$.
• Full Rank matrices always have an inverse
• Square matrices less than full rank never have an inverse
### More Properties of Matrices I

<table>
<thead>
<tr>
<th>Definition (Positive-Semidefinite Matrix)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A symmetric matrix $A$ is said to be positive-semidefinite (p.s.d) if and only if $x'Ax \geq 0$ for all $x$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Definition (Positive-Definite Matrix)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A symmetric matrix $A$ is said to be positive-definite (p.d.) if $x'Ax &gt; 0$ for all $x, x \neq 0$. Note that a matrix that is p.d. is also p.s.d.</td>
</tr>
</tbody>
</table>
More Properties of Matrices II

Definition (Idempotent Matrices)

A matrix \( P \) is idempotent if \( PP = P^2 = P \). A symmetric idempotent matrix is called a projection matrix.
Trace of a Matrix

• An important operation on square matrices is called the trace.

• While not blatantly obvious at the moment, the trace of a square is encountered throughout statistics and therefore we’ll define it.

Definition (trace)

The trace of a square matrix \( \mathbf{A} = \{a_{ij}\} \) of order \( n \) is defined to be the sum of the \( n \) diagonal elements of \( \mathbf{A} \) and is said to be the symbol \( \text{tr}(\mathbf{A}) \). Thus

\[
\text{tr}(\mathbf{A}) = a_{11} + a_{22} + \cdots + a_{nn}
\]
Finally we introduce Differentiation for Vectors

• If \( \frac{d}{d\beta} = \left( \frac{d}{d\beta_i} \right) \), then

  1. Consider the vector \( \mathbf{a} \),

  \[
  \frac{d(\beta' \mathbf{a})}{d\beta} = \mathbf{a}
  \]

  2. If \( \mathbf{A} \) is a symmetric matrix, then

  \[
  \frac{d(\beta' \mathbf{A} \beta)}{d\beta} = 2\mathbf{A} \beta
  \]
Simple Linear Regression I

- Consider a random sample of $n$ observations such

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and independent observations.

- Here the $x_i$ are observed and known and we would like to estimate the parameter $\beta$.

- We can rewrite into matrix notation for the $n$ observations

$$
\begin{pmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{pmatrix} = 
\begin{pmatrix}
1 & x_{11} \\
1 & x_{21} \\
\vdots & \vdots \\
1 & x_{n1}
\end{pmatrix}
\begin{pmatrix}
\beta_0 \\
\beta_1
\end{pmatrix}
+ 
\begin{pmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_n
\end{pmatrix}
$$

or

$$Y = X\beta + \epsilon$$
Simple Linear Regression II

- One method that can be used to estimate $\beta$ is through the method of least squares
- The idea is to find the vector $\beta$ which minimizes the squared errors

$$\sum_{i=1}^{n} \epsilon_i^2 = \epsilon' \epsilon$$

$$= (Y - X\beta)'(Y - X\beta)$$

- That is

$$\hat{\beta} = \arg\min_{\beta} (Y - X\beta)'(Y - X\beta)$$
Let’s expand this function

\[(Y - X\beta)'(Y - X\beta) = Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta\]

\[= Y'Y - 2\beta'X'Y + \beta'X'X\beta\]

where the above holds since \(\beta'X'Y = Y'X\beta\) which is a scalar.
Simple Linear Regression IV

Now

\[
\frac{d}{d\beta} \left( (Y - X\beta)'(Y - X\beta) \right) = \frac{d}{d\beta} \left( Y'Y - 2\beta'X'Y + \beta'X'X\beta \right)
\]

\[
= -2X'Y + 2X'X\beta
\]

We can set this equal to zero and thus

\[
X'Y = X'X\beta
\]

Now provided the inverse of \(X'X\) exists we have.

\[
\hat{\beta} = (X'X)^{-1}X'Y
\]
Let us consider $X'X$, its inverse will exist only if it is full rank and/or nonsingular.

\[
X'X = \begin{pmatrix}
1 & 1 & \ldots & 1 \\
1 & x_1 & x_2 & \ldots & x_n \\
1 & x_1 & x_2 & \ldots & x_n \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
1 & x_1 & x_2 & \ldots & x_n \\
\end{pmatrix}
\begin{pmatrix}
1 & x_1 \\
1 & x_2 \\
1 & \vdots \\
1 & x_n \\
\end{pmatrix}
\]

\[
X'X = \begin{pmatrix}
\sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \\
\sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \\
\end{pmatrix}
\]

The determinant is $\det(X'X) = n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2$
Consider if \( x = 1 = (1 \ 1 \ \ldots \ 1)^T \), Then

\[
\det(X'X) = n \sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2 \\
= n^2 - n^2 = 0
\]

We also see that

\[
X'X = \begin{pmatrix} n & n \\ n & n \end{pmatrix}
\]

which is not full rank. Thus one condition for inversion is that \( x \neq 1 \)
Simple Linear Regression VII

Continuing we can solve for \( \hat{\beta} \), by our formula for \( 2 \times 2 \) inversions we have

\[
(X'X)^{-1} = \frac{1}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \left( \begin{array}{cc}
\sum_{i=1}^{n} x_i^2 & - \sum_{i=1}^{n} x_i \\
- \sum_{i=1}^{n} x_i & \frac{n}{n}
\end{array} \right)
\]

and

\[
X^T Y = \left( \begin{array}{c}
\sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} x_i y_i
\end{array} \right)
\]
Simple Linear Regression VIII

Without going into all fun of calculating this for you guys, it can be shown that

\[ \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \bar{y} - \hat{\beta}_1 \bar{x} \\ \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{pmatrix} \]
References


