Review of Linear Algebra for Statistics

Brian Vegetabile

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=ZK3O402wf1c`)
    - 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 Statisticians 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 $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$

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

  which is said to be "non-degenerate" when $\mathbf{\Sigma}$ is positive-definite.
- Additionally, $\mathbf{x}$ is a real-valued n-dimensional column vector and $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$
- 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 μ which we measure under error ,ϵ, 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 $\mathbf{x}$ are explanatory variables and each β_j is unknown and to be estimated

Motivation III

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

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

- Or more simply written

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

- We will eventually show that $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$.
- 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.

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

- 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 $\mathbf{A}$ and a scalar k, then

$$k\mathbf{A} = k\{a_{ij}\} = \{ka_{ij}\}$$

- Matrix Addition: Consider two matrices $\mathbf{A}$ and $\mathbf{B}$, if they are both of dimension $m \times n$ then we define addition between these two matrices. Specifically $\mathbf{A} + \mathbf{B}$ 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 $\mathbf{A} = \{a_{ij}\}$ represent an $m \times n$ matrix and $\mathbf{B} = \{b_{ij}\}$ a $p \times q$ matrix. When $n = p$ (when $\mathbf{A}$ has the same number of columns as $\mathbf{B}$ has rows), then the matrix product $\mathbf{AB}$ 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 $\mathbf{AB}$ is called the premultiplication of $\mathbf{B}$ by $\mathbf{A}$ or the postmultiplication of $\mathbf{A}$ by $\mathbf{B}$.
  - When $n \neq p$ then the matrix product $\mathbf{AB}$ is undefined.
  - Two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ are said to commute if $\mathbf{AB} = \mathbf{BA}$

Basic Matrix Operations III

- Matrix Transpose: The transpose of an $m \times n$ matrix $\mathbf{A}$, to be denoted $\mathbf{A}^T$ or $\mathbf{A}'$ is the $n \times m$ matrix whose ij^{th} element is the ji^{th} element of $\mathbf{A}$.
  - For any matrix $\mathbf{A}$, $(\mathbf{A}')' = \mathbf{A}$
  - For any two matrices $\mathbf{A}$ and $\mathbf{B}$ which are conformal for addition

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

  - Finally any two matrices $\mathbf{A}$ and $\mathbf{B}$ for which the product is defined,

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

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.

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

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

Symmetric Matrices

- A matrix $\mathbf{A}$ is said to be symmetric is $\mathbf{A}' = \mathbf{A}$
- Thus a symmetric matrix is a square matrix where the ij^{th} element equals the ji^{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

$$\begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_n \end{pmatrix}$$

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

Identity Matrix

- Often the most useful diagonal matrix is the identity matrix $\mathbf{I}_n$ where the subscript n denotes the dimension of the identity matrix $(n \times n)$. That is,

$$\mathbf{I}_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

  often the subscript n is dropped.

- An important property is

$$\mathbf{IA} = \mathbf{AI} = \mathbf{A}$$

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

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$$

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

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 $(\mathbf{I}_n)^{-1} = \mathbf{I}_n$.

Inverting a 2×2 Matrix. I

- Consider the following matrix

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

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

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

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

- By our previous definitions we should have that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$

Inverting a 2×2 Matrix. II

$$\begin{aligned}
\mathbf{A}\mathbf{A}^{-1} &= \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix} \\
&= \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{11}a_{22} - a_{12}a_{21} & -a_{11}a_{12} + a_{12} \\ a_{21}a_{22} - a_{22}a_{21} & -a_{21}a_{12} + a_{22} + \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}
\end{aligned}$$

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

Linear Dependence and Rank I

- 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

$$\sum_{i=1}^{p} a_i \mathbf{x}_i = 0$$

Linear Dependence and Rank II

- 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 $\mathbf{x}_i$ which form a linearly independent set.
- The matrix $\mathbf{X}'\mathbf{X}$ is a $p \times p$ matrix.
  - If $\mathbf{X}$ has rank p, so does $\mathbf{X}'\mathbf{X}$.
- Full Rank matrices always have an inverse
- Square matrices less than full rank never have an inverse

More Properties of Matrices I

Definition (Positive-Semidefinite Matrix)

A symmetric matrix $\mathbf{A}$ is said to be positive-semidefinite (p.s.d) if and only if

$$\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$$

for all $\mathbf{x}$

Definition (Positive-Definite Matrix)

A symmetric matrix $\mathbf{A}$ is said to be positive-definite (p.d.) if

$$\mathbf{x}'\mathbf{A}\mathbf{x} > 0$$

for all $\mathbf{x}, \mathbf{x} \neq 0$. Note that a matrix that is p.d. is also p.s.d.

More Properties of Matrices II

Definition (Idempotent Matrices)

A matrix $\mathbf{P}$ is idempotent if $\mathbf{PP} = \mathbf{P}^2 = \mathbf{P}$. 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

$$\text{tr}(\mathbf{A}) = a_{11} + a_{22} + \cdots + a_{nn}$$

Vector Differentiation

- 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 β.
- 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

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

Simple Linear Regression II

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

$$\begin{aligned}\sum_i^n \epsilon_i^2 &= \boldsymbol{\epsilon}'\boldsymbol{\epsilon} \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\end{aligned}$$

- That is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Simple Linear Regression III

Let's expand this function

$$\begin{aligned}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) &= \mathbf{Y}'\mathbf{Y}-\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y}-\mathbf{Y}'\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{Y}'\mathbf{Y}-2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y}+\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

where the above holds since $\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y}=\mathbf{Y}'\mathbf{X}\boldsymbol{\beta}$ which is a scalar.

Simple Linear Regression IV

Now

$$\frac{d}{d\boldsymbol{\beta}}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = \frac{d}{d\boldsymbol{\beta}}(\mathbf{Y}'\mathbf{Y}-2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y}+\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta})$$
$$= -2\mathbf{X}'\mathbf{Y}+2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

We can set this equal to zero and thus

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Now provided the inverse of $\mathbf{X}'\mathbf{X}$ exists we have.

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

Simple Linear Regression V

Let us consider $\mathbf{X}'\mathbf{X}$, its inverse will exist only if it is full rank and/or nonsingular.

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

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

The determinant is $det(\mathbf{X}'\mathbf{X}) = n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2$

Simple Linear Regression VI

Consider if $\mathbf{x} = \mathbf{1} = \begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix}^T$, Then

$$\begin{aligned} det(\mathbf{X}'\mathbf{X}) &= n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 \\ &= n^2 - n^2 = 0 \end{aligned}$$

We also see that

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & n \\ n & n \end{pmatrix}$$

which is not full rank. Thus one condition for inversion is that $\mathbf{x} \neq \mathbf{1}$

Simple Linear Regression VII

Continuing we can solve for $\hat{\boldsymbol{\beta}}$, by our formula for 2×2 inversions we have

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix}$$

and

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

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

Stephen Abbott. *Understanding analysis*. Springer, 2001.

George Casella and Roger L Berger. *Statistical inference*. Second edition, 2002.

EJ Dudewicz and SN Mishra. Modern mathematical statistics. john wilsey & sons. *Inc., West Sussex*, 1988.

Sheldon Ross. *A first course in probability*. Pearson, ninth edition, 2014.

Mark J Schervish. *Probability and Statistics*. 2014.