2 Matrix Algebra

2.1 Vectors & Matrices

An $n \times m$ **matrix** A is an array of numbers with n rows and m columns. A_i denotes the ith row and is itself a $1 \times m$ matrix. A^j denotes the jth column and is an $n \times 1$ matrix. Any $n \times 1$ matrix is also called a **vector** of size n. $\mathbb{R}^n$ denotes the set of all vectors of size n and $\mathbb{R}^{n \times m}$ denotes the set of all matrices that are $n \times m$.

Often we write $[a_{ij}]_{i=1,\ldots,n}^{j=1,\ldots,m}$ (or simply $[a_{ij}]$ when it is clear what n and m are) to denote the matrix A; and $[a_i]_{i=1,\ldots,n}$ (or simply $[a_i]$) to denote the $n \times 1$ matrix (i.e. vector) a. If $A = [a_{ij}]$ and $B = [b_{ij}]$ are both $n \times m$ matrices then $A + B$ is defined as the $n \times m$ matrix $[a_{ij} + b_{ij}]$. The **transpose** of the matrix $A = [a_{ij}]$ is the matrix $A' = [a_{ji}]$. The **dot product** of two vectors $a = [a_i]$ and $b = [b_i]$ is defined as the sum $\sum_{i=1}^n a_i b_i$ and is denoted $a'b$ or $b'a$ or $a \cdot b$. The **length** of a vector a of size n is $(a \cdot a)^{0.5}$ and is denoted $||a||$. If $a_i = 0$ for all $i = 1, \ldots, n$ then the vector a is called the **zero-vector** of size n and is denoted 0_n or just 0 when it is clear what n should be. If $a_i = 1$ for all $i = 1, \ldots, n$ then a is called the **one-vector** of size n and is denoted 1_n.

The product AB of an $n \times m$ matrix A and an $l \times k$ matrix B is not defined unless $l = m$, in which case it is the $n \times k$ matrix $[(A_i \cdot B^j)_{ij}]$. If c is a number then $c[a_{ij}] = [ca_{ij}]$. A **square matrix** is an $n \times n$ matrix, where n is called the **order** of the matrix. A **symmetric matrix** is one that is equal to its transpose.

A **lower triangular matrix** of order n is a square matrix of order n where $a_{ij} = 0$ for all $j > i$. An **upper diagonal matrix** of order n is a square matrix of order n whose transpose is a lower triangular matrix of order n. A **diagonal matrix** of order n is a lower triangular matrix of order n that is also an upper triangular matrix.

The **identity** matrix of order n is a diagonal matrix of order n where $a_{ij} = 1$ for all $i = j$. It is denoted I_n or just I when it is clear what n should be.

Exercise 7. Verify that (i) $A + B = B + A$, (ii) $(A + B) + C = A + (B + C)$, (iii) $(AB)C = A(BC)$, (iv) $A(B+C) = AB+AC$, (v) $(A+B)' = A'+B'$, (vi) $(AB)' = B'A'$, and (vii) $AI = A$ and $BI = B$ for any $n \times m$ matrices A and B (note that I does not denote the same matrix in the two equations: the two Is differ by their order so that the products are defined), and (viii) $I = I^2 = I^3 = \cdots$.

Exercise 8. Prove that if a and b are two vectors each of size n, then $|a \cdot b| \leq ||a||||b||$. This is known as the **Cauchy-Schwartz** inequality. *Hint:* If $b = 0_n$ then the result follows. If $b \neq 0_n$ then you can let $x = \frac{a \cdot b}{b \cdot b}$ and write $a = a - xb + xb$. Then you will have to show that $||a||^2 = ||a - xb||^2 + x^2||b||^2$ to get $x^2||b||^2 \leq ||a||^2$. You will be able (I hope) to derive the result from this inequality.

Exercise 9. Verify that the triangle inequality holds for vectors. That is, if x and y are each vectors of size n, $||x + y|| \leq ||x|| + ||y||$.

2.2 The Rank of a Matrix

Consider m vectors each of size n. Call the set of these vectors $V = \{a_1, \ldots, a_m\}$. A linear combination of V is an expression of the form $x_1a_1 + x_2a_2 + \cdots + x_ma_m$ where $x_1, \ldots, x_m$ are all numbers. V is said to be **linearly independent** if

$$x_1a_1 + x_2a_2 + \cdots + x_ma_m = 0_n \tag{1}$$

implies that $x_i = 0$ for all $i = 1, \ldots, m$. V is said to be **linearly dependent** if there are numbers $x_1, \ldots, x_m$, not all of which are equal to 0, such that

$$x_1a_1 + x_2a_2 + \cdots + x_ma_m = 0_n. \tag{2}$$

Any set like V is either linearly independent or linearly dependent.

Exercise 10. Let $a_1 = 3$, $a_2 = 7$, $b_1 = 2$, $b_2 = 4$, $c_1 = 0$, and $c_2 = 2$. Is the set $\{[a_i], [b_i], [c_i]\}$ linearly dependent or independent?

Let A be an $n \times m$ matrix. Take $\mathcal{A}^C = \{A^1, \ldots, A^m\}$, which is the set of columns of A, and let $\phi^C : \mathcal{P}(\mathcal{A}^C) \to \mathbb{R}$ be the function defined by

$$\phi^C(Z) = \begin{cases} 0 \text{ if } Z \text{ is linearly dependent} \\ |Z| \text{ if } Z \text{ is linearly independent} \end{cases} \tag{3}$$

Similarly take $\mathcal{A}_R = \{A_1, \ldots, A_n\}$, which is the set of rows of A, and let $\phi_R : \mathcal{P}(\mathcal{A}_R) \to \mathbb{R}$ be the function defined in exactly the same way as ϕ^C. Since n and m are both finite, ϕ^C and ϕ_R both achieve maximums on their domains. The **column rank** and **row rank** of A are then defined, respectively, as

$$c = \max_{Z \in \mathcal{P}(\mathcal{A}^C)} \phi^C(Z) \qquad \text{and} \qquad r = \max_{Z \in \mathcal{P}(\mathcal{A}_R)} \phi_R(Z).$$

The row rank (and column rank) of a matrix does not change when any of the following three operations are applied to the matrix:

1. interchanging any two rows (or columns)

2. multiplying each entry in a given row (or column) by a nonzero number

3. replacing any row (or column) by itself plus a number k times another row (or column)

Exercise 11. Convince yourself that the column and row ranks of a matrix are invariant to row and column operations above.

Theorem 5. *If A is an $n \times m$ matrix with c and r positive, then $r = c$; i.e., row rank equals column rank.*

Proof. Since $r > 0$ the matrix is not one where all of the entries are 0. Pick one nonzero component and through a series of successive row and column operations convert it to a matrix B where $b_{11} \neq 0$. This $b_{11} \neq 0$ is called the pivot entry. Now multiply the first row of this matrix by b_{21}/b_{11} and subtract it from the second row. Then multiply it by b_{31}/b_{11} and subtract it from the third row. Continue doing so down the rows. Then go across the columns doing the same thing until you get a matrix that has 0s in every row except the first, and in every column except the first. If there are any other entries that are nonzero, then you can pick any nonzero entry and after a series of column and row interchanges you can convert it to a matrix C where $c_{22} \neq 0$. Taking c_{22} to be the pivot entry, after a series of operations like those performed on B, you arrive at a matrix, D that has nothing but zeros in the second column and second row except in the the d_{22} position. Continue this process until you run out of candidates for pivot entries or you run out of spaces for pivot entries. Either way, you have a matrix of 0s except along a diagonal. Therefore, the column rank is equal to the row rank since the row and column ranks of this final matrix are equal to that of the matrix you started with. □

In light of this result, the column rank and row rank of a matrix are referred to simply as the **rank** of the matrix. An $n \times m$ matrix A is said to have **full rank** if the rank of the matrix is equal to $\min\{m, n\}$.

Exercise 12. Use row and column operations to calculate the rank of the matrix:

$$M = \begin{bmatrix} 1 & 2 & -3 \\ 2 & 1 & 0 \\ -2 & -1 & 3 \\ -1 & 4 & 2 \end{bmatrix} \tag{4}$$

2.3 The Determinant & Inverses

Square matrices are special because they are the only kinds of matrices for which we can calculate what is called the *determinant*. Consider the square matrix A of order n. Consider the $(n-1)\times(n-1)$ submatrix of A created by deleting row i and column j. Call that matrix $A(i,j)$. The (i,j)-**cofactor** of A is defined as

$$C_{ij}(A) = (-1)^{i+j} \det A(i,j),$$

where $\det A(i,j)$ is the determinant of the matrix $A(i,j)$. Now the **determinant** of a 1×1 matrix is the value of the single entry. For an $n\times n$ matrix A, the determinant is defined as

$$\det A = a_{11}C_{11}(A) + \cdots + a_{1n}C_{1n}(A). \tag{5}$$

You may object that this definition is circular since we use the notion determinant to define the cofactor. However, since we defined the determinant of a 1×1 matrix, the above equality helps us to recursively define determinants for any $n\times n$ matrix.

The determinant may seem like a mysterious concept to you. To make it more concrete, suppose that A is a 2×2 matrix and its columns are A^1 and A^2. The determinant of A is the area of the parallelogram spanned by the vectors A^1 and A^2. If A is a 3×3 matrix then the determinant of A is the volume of the parallelepiped spanned by the vectors A^1, A^2 and A^3. If A is a 4×4 matrix, then the determinant of A is ...

Exercise 13. Find a simple formula for the determinant of any 2×2 matrix. Use this formula and equation (5) to calculate the determinant of the matrix M given in (4) with the second row deleted.

Exercise 14. Show that the determinant of any lower- or upper- triangular matrix is simply the product of the diagonal entries.

After having done Exercise 14, and knowing that you can convert a matrix into a lower or upper triangular matrix using row and column operations, the following properties will be useful to you in calculating the determinant of any matrix.

Let A be any square matrix of order n.

1. Let $[A,j,w]$ denote the $n\times n$ matrix in which the vector jth column of A is replaced by the vector w of size n. If u and v are two column vectors each of size n, then

   $$\det[A,j,u+v] = \det[A,j,u] + \det[A,j,v].$$

   This decomposition holds for all $j=1,...,n$.

2. If the matrix B is obtained from A by interchanging any two rows (or columns) of A then $\det B = -\det A$.

3. If A and B are two matrices of order n then $\det AB = \det A \det B$.

4. If B is obtained from A by multiplying each entry of some given row (or column) of A by a nonzero constant k, then $\det B = k \det A$.

5. If B is obtained from A by replacing any row (or column) of A by itself plus k times some other row (or column), where k is any number, then the determinant remains unchanged.

6. $\det A = \det A'$.

Exercise 15. Convince yourself that the five properties of determinants listed above hold. Then use property 1 to show that if a matrix has a row (or column) of zeros then its determinant is 0.

Exercise 16. Show that a square matrix A has full rank if and only if $\det A \neq 0$.

Theorem 6. (**Cramer's Rule**) *Let $A = [A^1, \ldots, A^n]$ be a square matrix of order n where the columns are $A^1, \ldots, A^n$. Suppose that $\det A \neq 0$ and let v be a vector of size n. Then the system of equations $Ax = v$ where x is a vector of n variables has a unique solution where each x_i, $i = 1, ..., n$ is given by*

$$x_i = \frac{\det[A^1, \ldots, A^{i-1}, v, A^{i+1}, \ldots, A^n]}{\det A}.$$

Proof. By Exercise 16, the matrix A has full rank. By row operations of the kind described above the augmented $n \times n + 1$ matrix $[A^1, \ldots, A^n, v]$, where v is a vector of size n can be reduced to a matrix with zeros above and below the diagonal and 1s on the diagonal, as in

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & c_1 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & c_n \end{bmatrix}$$

Therefore, the system of equations $Ax = v$ where x is a vector of n variables has a unique solution. Call it x^*. Thus

$$\begin{aligned}\det[A^1, \ldots, A^{i-1}, v, A^{i+1}, \ldots, A^n] &= \det[A^1, \ldots, A^{i-1}, Ax^*, A^{i+1}, \ldots, A^n] \\ &= \sum_{j=1}^{n} x_j^* \det[A^1, \ldots, A^{i-1}, A^j, A^{i+1}, \ldots, A^n] \\ &= x_i^* \det A.\end{aligned}$$

which follows from the properties of determinants listed above. Divide both sides by $\det A$ to find the solution given in the theorem. □

Theorem 7. (**inverse of a matrix**) *If A is an $n \times n$ matrix with $\det A \neq 0$ there exists a unique matrix B (called the inverse of A) such that $AB = BA = I_n$. Moreover, this matrix is given by $B = [C_{ij}(A)/\det A]'$, where $C_{ij}(A)$ is the (i, j)-cofactor of A*

Proof. To see why B is unique suppose that there was another matrix C such that $CA = I_n$. Then $CAB = B$, but also $CAB = C(AB) = CI_n = C$. So $B = C$. The same holds if $AC = I_n$.

Now we prove that B exists. Let e_{jn} be the size n vector such that there is a 1 in the jth position and 0 everywhere else. Then for any $n \times n$ matrix $X = [x_{ij}]$ solving $AX = I_n$ we have $e_{jn} = AX^j$ where X^j is the jth column of X. Since $\det A \neq 0$, the matrix A has full rank (by Exercise 16), and thus the solution exists and is unique (by row reduction). We have left to show that $XA = I_n$. By the properties of matrix multiplication and determinants, we can find a matrix Y such that $A'Y = I_n$, which is equivalent to $Y'A = I_n$, and we have $I_n = Y'(AX)A = (Y'A)XA = XA$.

Finally, we derive the formula mentioned in the theorem. By Cramer's rule,

$$\begin{aligned}x_{ij} &= \det[A^1, \ldots, A^{i-1}, e_{jn}, A^{i+1}, \ldots, A^n]/\det A \\ &= \det[A^1, \ldots, A^{i-1}, e_{jn}, A^{i+1}, \ldots, A^n]'/\det A \\ &= C_{ji}(A)/\det A\end{aligned}$$

which gives us the formula. □

The unique matrix B that is the inverse of A is typically denoted A^{-1}, and this is how we will denote it from here on.

The following properties are useful. Whenever inverses exist,

1. $(A')^{-1} = (A^{-1})'$

2. $(AB)^{-1} = B^{-1}A^{-1}$

3. $\det A^{-1} = 1/\det A$

4. The inverse of a lower (or upper) triangular matrix is a lower (or upper) triangular matrix.

Exercise 17. Prove the four properties above and find the inverse of the matrix in (4) with the second row deleted (if it exists).

2.4 Eigenvectors and eigenvalues

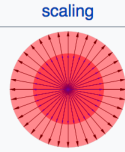
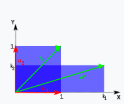
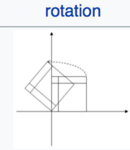
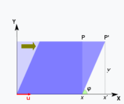
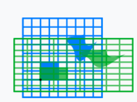
Let A be a square matrix of order n. A vector of size n is an **eigenvector** of A if there is a number λ such that $Av = \lambda v$. If $v \neq 0_n$ then λ is unique because $\lambda_1 v = \lambda_2 v$ implies $\lambda_1 = \lambda_2$. In that case, λ is said to be an **eigenvalue** of A belonging to v.

Theorem 8. *Let A be a square matrix of order n. Then λ is an eigenvalue of A belonging to some nonzero vector if and only if* $\det(A - \lambda I) = 0$.

Proof. Asume that λ is an eigenvalue of A. Then by definition, there is a vector $v \neq 0$ such that $Av = \lambda v$. In other words, $Av - \lambda v = 0_n$. This implies $(A - \lambda I_n)$ is a matrix with linearly dependent columns (since $v \neq 0$), so that the rank of the matrix is less than n. Therefore, by Exercise 16 it must be that $\det(A - \lambda I) = 0$.

Now interpret $(A - \lambda I_n)v = 0_n$ as a set of n equations. If $(A - \lambda I_n)$ does not have full rank then there is at least one equation that is a linear combination of the others. Eliminate one of the redundant equations. Now you are left with a system of equations with more variables than unknowns, which means that at least one variable can be set freely. That is equivalent to setting one entry of v freely. Make that one entry nonzero. Therefore, there is a vector $v \neq 0_n$ such that $Av = \lambda v$. □

Note that a matrix A represents the transformation of a vector space in the following sense: for each vector v, the vector Av is a new vector transformed by A. I reproduce examples of such transformations in the vector space $\mathbb{R}^2$ from Wikipedia, along with form of the matrix form of A that represents the transformation.

	scaling	unequal scaling	rotation	horizontal shear	hyperbolic rotation
illustration					
matrix	$\begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix}$	$\begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix}$	$\begin{bmatrix} c & -s \\ s & c \end{bmatrix}$ $c = \cos\theta$ $s = \sin\theta$	$\begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} c & s \\ s & c \end{bmatrix}$ $c = \cosh\varphi$ $s = \sinh\varphi$

With this in mind, an eigenvector of A is simply a vector whose direction does not change, or is completely reversed (which happens when the associated eigenvalue is negative), under the transformation represented by A. Thus under "scaling," all vectors are eigenvectors of the transformation matrix. Under "horizontal shear" only the vectors parallel to the x-axis are eigenvectors, and so on.

Suppose A is an $n \times n$ matrix and the solution to $\det(A - \lambda I) = 0$ yields n eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. When plugged back into $Av = \lambda v$, one can find corresponding eigenvectors associated with these eigenvalues. Since $Akv = \lambda kv$ is equivalent to $Av = \lambda v$ for any nonzero k, these eigenvectors are not unique. We may take ones that are normalized, i.e. ones whose lengths are set to 1. These are called the "unit eigenvectors." For each eigenvalue, we therefore have one eigenvector. Then create the diagonal matrix

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix} \tag{6}$$

and the matrix P whose first column is v_1, second is v_2, third is v_3 etc. all the way up to v_n—these are the eigenvectors. Now it turns out that

$$A = PDP^{-1}.$$

You may look at a proof in any advanced linear algebra textbook or try to come up with one yourself. If you try to come up with your own proof, think in terms of matrices shifting axes; e.g., in $\mathbb{R}^2$ the vectors $(0, 1)$ and $(1, 0)$ define unit movement in the x and y directions. Suppose we wanted to rotate our coordinate system and re-write the vector in the new system. How would we do that?

Exercise 18. Let

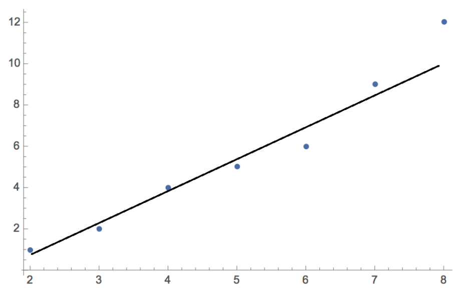
$$A = \begin{bmatrix} 2 & 3 \\ 1 & 0 \end{bmatrix} \tag{7}$$

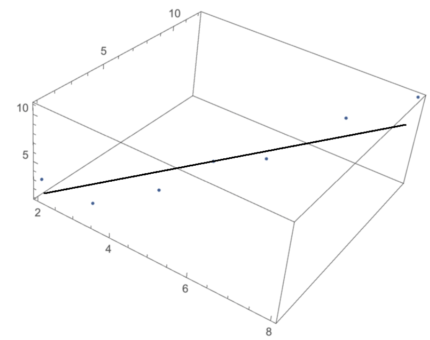
Find A^{19}.

2.5 Application: The Algebra of Least Squares

Let y be vector of size n and $\bar{X}$ an $n \times k$ matrix. We will refer to $(\bar{X}, y)$ together as the "data." To motivate what we plan to do, consider the case where $k = 1$. Then $\bar{X}$

too is a vector of size n. Suppose we plot values of (x_i, y_i), $i = 1, ..., n$, with the x's on the horizontal axis and y's on the vertical axis. We have a scatter of n points. We want to think of fitting a line through this scatter to "summarize" the relationship, linearly. When $k = 2$, we can imagine doing the same thing, but now we have a scatter of points in three-dimensional space. Again, we will want to think of fitting a line through this scatter. For $k > 2$, imagine doing the same thing even though it is hard to depict. It will be convenient to keep in mind the cases of $k = 1$ and $k = 2$ as we proceed.





A line is defined by its slope or **gradient**. In $(k+1)$-space, it can be written as $\tilde{y} = \tilde{x} \cdot m + b$ where m is a vector of size k, b is a number, $\tilde{x}$ is a vector of variables $(x_1, ..., x_k)$. Fitting the line means choosing values of m and b. How should these values be chosen? One way that we explore here is to "minimize the sum of squared residuals," also called the **ordinary least squares** (OLS). The point (x_i, y_i) from our data, plotted in $k+1$-space, is off from the fitted line by the amount $\varepsilon_i := y_i - (x_i \cdot m + b)$. Note that

$$\sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (y_i - (x_i \cdot m + b))^2 = (y - X\beta) \cdot (y - X\beta) =: S(\beta)$$

where X is the $n \times k + 1$ matrix of data with the 1_n vector appended to it (i.e., $\bar{X}$ with the 1_n vector added as a column at the end) and $\beta = (m, b)$ is the $(k + 1) \times 1$ matrix (i.e., vector) $(m_1, ..., m_k, b)$. We refer to each ε_i as the residual for observation i, and $S(\beta)$ as the sum of squared residuals.

We are interested in the value of $\beta = (m, b)$ that minimizes $S(m, b)$. You will solve this problem as an exercise later (after we introduce optimization). For now, note that

$$\begin{aligned} S(\beta) &= (y - X\beta)'(y - X\beta) \\ &= (y' - \beta' X')(y - X\beta) \\ &= y'y - \beta' X' y - y' X \beta + \beta' X' X \beta \\ &= y'y - 2y' X \beta + \beta' X' X \beta \end{aligned} \tag{8}$$

Exercise 19. Give reasons for why each of the equalities in (8) holds.