Topic 1 : Linear Algebra for Econometrics

Dr. Ani Dasgupta
Department of International Business, Mass. Maritime Academy
and
Department of Economics, Boston University

Draft lecture notes for a graduate econometrics course offered at Boston University, Spring 2017. Please do not circulate without permission.

1 Motivation

It is not that one cannot try and teach econometrics without matrices, but in my opinion, it is a misguided effort. A few lectures in this area will save us a lot of time and ink later in trying to communicate and understand certain ideas, and in fact there are certain complicated issues I doubt one can explain without using matrix notation. Please note that what follows is no substitute for a more solid, step-by-step account of linear algebra; and I do assume that you had one such basic course at some point.[1] However, for the most part the treatment here is self-contained, though I don't explain real basics such as how to multiply two matrices or how to take a transpose; also, some results are stated without proof. My purpose here is to provide a refresher for a few classical linear algebra concepts, but more importantly, I try to give you a whiff of why the results are useful in the study of econometrics. To motivate you in this direction, I begin with a few illustrative examples.

1.1 First Motivating Example: Describing The Classical Linear Model

The first area we will study in this course is the classical linear regression model. Topic 3 is devoted to it, but here is a brief idea of what the model is all about. Suppose we are interested

[1] If you haven't and wish to learn by studying a textbook on your own, I have a couple of recommendations. I really like "Applied Linear Algebra" by Ben Noble and James W. Daniel, Prentice Hall, 3rd. edition, but it is an old textbook and may be difficult to find. Another recommendation is "Matrix Analysis and Applied Linear Algebra", by Carl D. Meyer, SIAM.

in how a certain left-hand-side variable (the regressand) is related to a bunch of other right-hand-side variables (regressors)[2]. For a concrete example that we will shortly study, suppose we are interested in estimating the cost function in the electricity industry by collecting data from several electricity-generating firms (as Marc Nerlove was in his 1963 paper)[3]. In what is known as the parametric approach, we postulate that we know the functional form of the cost function except for certain parameters[4]. For instance, we assume that log of the cost is linearly related to log of output and log of input prices, which as we will see later, is exactly what the Cobb-Douglas production function predicts. So, consider a three-input scenario with labor, capital and fuel being the 3 inputs. Suppose for a typical firm in the sample, $y = ln$ (cost), $x_1 = 1$, $x_2 = ln$ (output) $x_3 = ln$ (wage rate), $x_4 = ln$ (price of capital) $x_5 = ln$ (price of fuel). We can then write our linear formulation as

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \tag{1}$$

or equivalently,

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \tag{2}$$

where the 'betas' are unknown coefficients to be estimated by the econometricians from data. However, we admit to ourselves that it is too much to expect that all firms have exactly this same cost function, and hence, we augment the above equation by adding a so-called 'error term' or a 'random disturbance term' to it. We call this term ε. So, the new functional form becomes

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon \tag{3}$$

The idea here is even if the econometrician knew perfectly the values of the 'betas', as well as the input prices and output of a particular firm, he will not be able to predict perfectly the cost incurred by that firm. Sometime, equation (2) will overpredict and sometimes it will underpredict because of 'random' factors that have either not been modeled, and/or beyond anyone's control. The error term thus is assumed to be random and take different values for different firms. If it is 0 in expectation, then equation (2) is to be interpreted as the 'average' firm's cost function.

Generalizing, a k-variable linear regression model, then, is given by

$$y = \beta_1 x_1 + \beta_1 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon \tag{4}$$

Notice that typically, in the k-variable model, there are $k - 1$ *true* right-hand-side variables. Sometimes, theory demands that there should be no intercept term in the linear equation above - in such cases, the k-variable model will indeed have k true right-hand-side variables.

[2] Notice that I have avoided saying, "...how a certain *dependent* variable is *affected* by a bunch of *independent* variables". There is a good reason for this. Regression analysis does not presuppose causation from the right hand side variables to the left hand side; only that the variables are statistically interrelated in a certain way. This cryptic remark will become more transparent later.

[3] Nerlove, Marc (1963): "Returns to Scale in Electricity Supply", in Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld, (Stanford University Press).

[4] The trick is to choose a functional form that can well-approximate actual production functions for different choice of parameters. It is economic theory that usually guides us in this choice.

Now, let us say the econometrician goes about and collects data on n different firms each of which obeys equation (4). Let, for the i-th firm, y_i be the observed value of the left hand side variable, ϵ_i the (unobserved) value of the random disturbance, and x_{ij} the observed value of the j-th right-hand-side variable. Then, in effect, we have an equation system:

$$\begin{array}{rcl} y_1 & = & \beta_1 x_{11} + \beta_1 x_{12} + \beta_3 x_{13} + \ldots + \beta_k x_{1k} + \varepsilon_1 \\ y_2 & = & \beta_1 x_{21} + \beta_1 x_{22} + \beta_3 x_{23} + \ldots + \beta_k x_{2k} + \varepsilon_2 \\ \ldots & \ldots & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n & = & \beta_1 x_{n1} + \beta_1 x_{n2} + \beta_3 x_{n3} + \ldots + \beta_k x_{nk} + \varepsilon_n \end{array} \tag{5}$$

Now, to see the usefulness of matrices in this context, define the following vectors and matrices:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \tag{6}$$

Note that $\mathbf{y}, \boldsymbol{\varepsilon}$ are $n \times 1$ matrices (i.e. n-vectors), $\boldsymbol{\beta}$ is a $k \times 1$ matrix (i.e. a k-vector) and $\mathbf{X}$ is $n \times k$ (this matrix is sometime called the *design* matrix). With these symbols in place, using the definitions of matrix multiplication and addition, it is easy to verify that equation system (5) can simply be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{7}$$

It is much simpler and neater to state that you have a model given by equation (7) than by equation (5). Don't you agree?

1.2 Second Motivating Example: Calculating Variances and TSS

Now, suppose we want to obtain an expression for the (sample) variance in the y-data. As you might know from a prior statistics course, this requires calculating the expression: $\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ where $\bar{y}$ is the average of the y_i's. Here is a way to create a neat matrix expression for the sum. Define $\boldsymbol{\iota}$ (read 'iota') to be a $n \times 1$ (column) vector of 1's. Then $\sum_{i=1}^{n} y_i = \boldsymbol{\iota}'\mathbf{y}$. Now, let us create a vector $\mathbf{y_d}$ of 'deviations from the mean'. We have

$$\mathbf{y_d} \equiv \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix} = \mathbf{y} - \begin{bmatrix} 1/n \sum y_i \\ \vdots \\ 1/n \sum y_i \end{bmatrix} = \mathbf{y} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}'\mathbf{y} = \left[\mathbf{I} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}'\right]\mathbf{y} = \mathbf{D}\mathbf{y} \tag{8}$$

where $\mathbf{D} = \left[\mathbf{I} - \frac{\mathbf{1}}{\mathbf{n}}\boldsymbol{\iota}\boldsymbol{\iota}'\right]$. Let me make a couple of remarks about the $\mathbf{D}$ matrix. It is a symmetric matrix (i.e. $\mathbf{D}' = \mathbf{D}$). It is also an idempotent matrix which means that multiplied to itself it gives itself ($\mathbf{DD} = \mathbf{D}$). You should have little trouble verifying these claims (do it!). The matrix

D acting on a vector of any set of observations creates deviations from the mean, and is a very useful matrix. More generally, symmetric, idempotent matrices are ubiquitous in econometrics - we will have a lot to say about them later.

Now, going back to the issue of finding a (matrix) expression for variance, note that

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathbf{y}_\mathbf{d}^{'}\mathbf{y}_\mathbf{d} = (\mathbf{Dy})'(\mathbf{Dy}) = \mathbf{y}'\mathbf{D}'\mathbf{Dy} = \mathbf{y}'\mathbf{DDy} = \mathbf{y}'\mathbf{Dy} \tag{9}$$

Hence, the sample variance we were seeking can be written as $\frac{1}{n-1}\mathbf{y}'\mathbf{Dy}$. The term on the left-hand-side of equation (9) is called in the regression context, total sum of squares (TSS), while expressions such as $\mathbf{y}'\mathbf{Dy}$ which is a n-dimensional row vector multiplied by an $n \times n$ matrix multiplied by an n-dimensional column vector is called a 'quadratic form' (it is a scalar, quadratic function of the individual y_i's). We have just established that TSS is a quadratic form. In the next topic, we will explore statistical properties of quadratic forms when the elements of $\mathbf{y}$ (but not of $\mathbf{D}$) are random.

1.3 Third Motivating Example: Calculating Covariances

Another important set of numerical data items are sums of cross-products such as $\sum_{i=1}^{n} y_i x_{ij}$ or $\sum_{i=1}^{n} x_{ij}x_{ik}$ (Here j and k refer to indices of two different right-hand-side variables. i as usual refers to observation index). These are easily obtainable from the matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$. For instance,

$$\begin{aligned}
\mathbf{X}'\mathbf{X} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{bmatrix} \\
&= \begin{bmatrix} n & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i2}^2 & \cdots & \sum_{i=1}^{n} x_{i2}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{i2}x_{ik} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \qquad (10)
\end{aligned}$$

As you can see, the j-th row and k-th column element of this matrix is precisely $\sum_{i=1}^{n} x_{ij}x_{ik}$.

The $\mathbf{X}'\mathbf{X}$ matrix is a fundamental object of interest to the classical model, and whether it has an inverse or not determines how the estimates of the 'beta' coefficients are expressed and calculated. We will shortly examine this question.

Now, suppose we are interested in calculating sample covariance between the jth and the kth variable. Recall that the expression for this object is $\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ where $\bar{x}_j$

and $\bar{x}_k$ are the sample means of the two variables. The question is can we we obtain the information via matrix manipulations? The answer is, yes we can. The required number can be found in the (jk)-th entry of the matrix expression $\frac{1}{n-1}\mathbf{X}'\mathbf{D}\mathbf{X}$. To see this, note that $\mathbf{X}'\mathbf{D}\mathbf{X}$ is simply $(\mathbf{D}\mathbf{X})'(\mathbf{D}\mathbf{X})$ and now note that the (jk)-th entry of this object is the jth row of $(\mathbf{D}\mathbf{X})'$ multiplying the kth column of $(\mathbf{D}\mathbf{X})$. A moment's thought about what these rows and columns are should convince you that as a result of this multiplication, we obtain $\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)$, and the claim follows.

1.4 Fourth Motivating Example: Models Using Partitioned Matrices

Matrix notation is also useful in describing situations where we are concerned with two sets of regressors or two sets of data. Suppose, we have a k-variable model

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{k_1} x_{ik_1} + \beta_{k_1+1} x_{i\,k_1+1} + \ldots + \beta_k x_{ik} + \varepsilon_i \tag{11}$$

where $x_{i1} \ldots x_{ik_1}$ denote one set of regressors and $x_{i\,k_1+1} \ldots x_{ik}$ denote another set of regressors. You should convince yourself that using matrix notation, we can rewrite the model in equation (7) as

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \\ &= \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \boldsymbol{\varepsilon} \end{aligned} \tag{12}$$

where $\mathbf{X}_1$ is $n \times k_1$, $\mathbf{X}_2$ is $n \times (k-k_1)$, $\boldsymbol{\beta}_1$ is $k_1 \times 1$ and $\boldsymbol{\beta}_2$ is $(k-k_1) \times 1$. The second line of the equation above uses the notion of partitioned matrices. Partitioned matrices are convenient to manipulate; you can simply pretend that the (sub)matrices sitting inside partitions are just like simple matrix elements and then take transposes and multiply big matrices expressed in terms of partitions involving smaller matrices following standard rules of transposing and multiplication (make sure however that whenever you multiply two submatrices they are conformable for multiplication). Thus, for instance, in the above example, one can express $\mathbf{X}'\mathbf{X}$ as

$$\begin{aligned} \mathbf{X}'\mathbf{X} &= \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \end{aligned} \tag{13}$$

Later, we will encounter formulas for finding inverses of partitioned matrices.

Now suppose, I have a model where

$$y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \eta_k$$

and

$$z_i = \gamma_1 + \gamma_2 x_{i2} + \ldots + \gamma_k x_{ik} + \xi_k$$

I have n observations on y, z and the x's. The η's, ξ's are random disturbances. This kind of model may easily occur in a situation, where say, for each observational unit, there are two left hand side variables of interest.

This model too can be written in matrix form so that it looks like (7) which I leave to you as an exercise. But why would we want to do that? The answer is in general, since the classical linear model is written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, and we know almost everything there is to know about this model, any other model that can be cast in this format, also becomes analyzable.

Section End Questions

1. I am estimating a model with an intercept term and two other right hand side variables called w and z. I have 4 observations. the value for the 4 observations on w are 4, 9, 10 and 7. The values of z for the 4 observations for z are 1, -6, -2 and 4. What is my design matrix?
2. In the previous example, without actually performing any matrix multiplication, tell me what is the first row, first column element of $\mathbf{X'X}$? What about the second row, third column element?
3. I defined a quadratic form to be an object of the type $\mathbf{y'Ay}$ where $\mathbf{y}$ is an n-vector and $\mathbf{A}$ is an $n \times n$ matrix. Without loss of generality, $\mathbf{A}$ is usually assumed to be a symmetric matrix. Why do you think this is the case?
4. In the context of the last model described in sebsection 1.4, give an example where such a model could arise in a practical example. In your example, do you expect η and ξ to be independent random variables? Why / why not? Now, show how the model can be expressed in matrix form.

2 Vector Spaces, Subspaces, Linear Independence

DEFINITION: A set of n-vectors[5] belonging to n-dimensional Euclidean space $\mathcal{R}^n$ will be called a vector space if given any vector v and a scalar α, αv also belongs to the set and given any v_1 , v_2 belonging to the set, $v_1 + v_2$ also belongs to the set.

Thus these sets are '*closed* under (vector) addition and scalar multiplication'. It is very simple to figure out what these sets geometrically look like in familiar contexts such as $\mathcal{R}^3$. First note

[5] An n-vector of course, is just an ordered array of n numbers, such as $\begin{bmatrix} 4 \\ 6 \\ 27 \\ -3 \end{bmatrix}$ which is a 4-vector. Unless specifically mentioned otherwise, whenever we talk about a vector, we will mean a column vector.

that because of closure under scalar multiplication, the zero vector must be part of any vector space; also, if a particular (nonzero) point belongs to a vector space, the whole line connecting that point and the zero, must belong to the vector space. Lastly, using the parallelogram addition law of vectors,[6] you can see that given two vectors, the whole plane in which these two vectors lie belong to the set. Given these observations, it is clear in $\mathcal{R}^3$, there can be only 4 types of vector spaces: the origin, 1-dimensional lines passing through the origin, 2-dimensional planes containing the origin and the whole of $\mathcal{R}^3$ itself.

DEFINITION: When a vector space lies fully inside another larger vector space, we call the former a subspace of the latter. Thus, a 2-dimensional plane (containing the origin) is a subspace of $\mathcal{R}^3$.

DEFINITION: Suppose we have p n-vectors: $\mathbf{v_1}, \mathbf{v_2}, \ldots \mathbf{v_p}$. Then the n-vector $\alpha_1 \mathbf{v_1} + \ldots \alpha_p \mathbf{v_p}$, where $\alpha_1, \ldots \alpha_p$ are scalars, is called a linear combination of the vectors $\mathbf{v_1}, \mathbf{v_2}, \ldots \mathbf{v_p}$.

DEFINITION: A set of vectors $\mathbf{v_1}, \mathbf{v_2}, \ldots \mathbf{v_p}$ are called linearly independent if $\alpha_1 \mathbf{v_1} + \ldots \alpha_p \mathbf{v_p} = \mathbf{0}$ implies $\alpha_1, \ldots \alpha_p$ are all equal to 0. Thus, no non-trivial linear combinations of such vectors can be made to yield the zero vector.

As an example, consider vectors $\begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}$, and $\begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$. These 3 vectors are linearly independent because if we set

$$\alpha_1 \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} + \alpha_2 \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} + \alpha_3 \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{14}$$

we get the equation system

$$\begin{aligned} \alpha_1 + 2\alpha_2 &= 0 \\ 2\alpha_1 + \alpha_3 &= 0 \\ \alpha_2 + 2\alpha_3 &= 0 \end{aligned} \tag{15}$$

which upon solving (by method of substitution for example) gives $\alpha_1 = \alpha_2 = \alpha_3 = 0$ as one can easily verify.

[6] This is how the law works. Suppose you have two vectors: $\begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} 4 \\ -1 \\ 5 \end{bmatrix}$. Draw two arrows from the origin to these two points. Now complete the parallelogram with these two arrows as adjacent sides. Draw the arrow (along a diagonal) from the origin to the opposite tip of the parallelogram. This arrow represents the vector which is the sum of the two given vectors. In other words, the tip of the arrow has coordinates $\begin{bmatrix} 5 \\ 2 \\ 7 \end{bmatrix}$.

It should be clear that if one has a set of linearly independent set of vectors then a subset of these vectors must be linearly independent as well. Also, obvious should be the fact that the zero vector cannot be part of a linearly independent set.

DEFINITION: A set of vectors are said to be linearly dependent if they are not linearly independent.[7]

DEFINITION: A set of vectors $\mathcal{S}$ in a vector space $\mathcal{V}$ is said to *span* $\mathcal{V}$ if any vector belonging to the vector space can be expressed as a linear combination of the vectors in $\mathcal{S}$. If moreover, the vectors in $\mathcal{S}$ are linearly independent, then they are said to be form a *basis* for $\mathcal{V}$.

I now state a fundamental result that leads to the concept of the dimension of a vector space.

A Fundamental Proposition for Vector Spaces: Suppose p vectors $(\mathbf{u_1}, \mathbf{u_2}, \ldots \mathbf{u_p})$ constitute a basis for some vector space $\mathcal{V}$. Suppose the set of vectors $(\mathbf{v_1}, \mathbf{v_2}, \ldots \mathbf{v_p})$ belong to $\mathcal{V}$ and are linearly independent. Then $(\mathbf{v_1}, \mathbf{v_2}, \ldots \mathbf{v_p})$ is a basis for $\mathcal{V}$ as well.

Sketch of proof (fill out the details): Since $\mathbf{v_1}$ can be expressed as a linear combination of the $\mathbf{u}$'s, there is a $\mathbf{u}_i$ which can be written as a linear combination of $\mathbf{v_1}$ and the other $\mathbf{u}'s$. Wlog, let this be $\mathbf{u_1}$. Now, $\mathbf{v_1}, \mathbf{u_2}, \ldots \mathbf{u_p}$ span $\mathcal{V}$. Again $\mathbf{v_2}$ can be written as a linear combination of these vectors. Hence, there is another $\mathbf{u}$, say, wlog $\mathbf{u_2}$ which can be written in terms of $\mathbf{v_1}, \mathbf{v_2}, \ldots \mathbf{u_p}$ and hence, $\mathbf{v_1}, \mathbf{v_2}, \mathbf{u_3}, \ldots \mathbf{u_p}$ span $\mathcal{V}$. Proceeding inductively, $\mathbf{v_1}, \mathbf{v_2}, \ldots \mathbf{v_p}$ span $\mathcal{V}$.

The proposition and its proof technique immediately offers a few corollaries. See if you can verify them.

COROLLARY 1: Any two basis have the same number of vectors, and hence, we can talk about the *dimension* of a vector space which is the number of vectors in any basis.

COROLLARY 2: Every basis of $\mathcal{R}^n$ must contain exactly n vectors since the unit vectors $\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$ are clearly linearly independent and span $\mathcal{R}^n$.

COROLLARY 3: Any set of q vectors in a p-dimensional space are linearly dependent if $q > p$.

COROLLARY 4: Any set of q vectors in a p-dimensional space can not span the space if $q < p$.

[7] It is not a completely trivial job to determine whether a given set of n-vectors are linearly dependent or not, though sometimes this can be detected by inspection. To examine the issue formally, a technique known as Gaussian reduction (or Gaussian elimination) is to be employed. If you are curious about it consult a linear algebra text.

We end this section introducing a very important subspace of $\mathcal{R}^n$. Suppose $\mathbf{A}$ is an $n \times k$ matrix. Thus, it has k columns, each column being an n-vector. Now consider the object $\mathbf{Ax}$ where $\mathbf{x}$ is a k-vector. Realize that this object is an n-vector which happens to be a linear combination of the columns of $\mathbf{A}$. This becomes clear since we could write $\mathbf{A}$ as a partitioned matrix $(\mathbf{A_{.1}}, \ldots, \mathbf{A_{.k}})$ (where $\mathbf{A_{.j}}$ refers to the j'th column of $\mathbf{A}$), the vector $\mathbf{x}$ partitioned as $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ and hence,

$$\mathbf{Ax} = x_1\mathbf{A_{.1}} + x_2\mathbf{A_{.2}} + \ldots + x_k\mathbf{A_{.k}} \tag{16}$$

Similarly $\mathbf{y'A}$ where $\mathbf{y'}$ is a row vector represents a linear combination of the rows of $\mathbf{A}$.

DEFINITION: Consider the set $\mathcal{C}(\mathbf{A}) = \{\mathbf{z} : \mathbf{z} = \mathbf{Ax} \quad \text{for some } \mathbf{x}\}$; i.e. $\mathcal{C}(\mathbf{A})$ is simply the set of all linear combinations of the columns of $\mathbf{A}$. We call such a set the *column span* or *column space* of $\mathbf{A}$. It is easy to verify that $\mathcal{C}(\mathbf{A})$ is a vector space (it is closed under vector addition and scalar multiplication). As we work through the classical model, the column span of the design matrix $\mathbf{X}$ will be of considerable interest to us.

Section End Questions

1. Write down the details of the proof of the main proposition in this section. Where exactly are we using the linear independence of the $\mathbf{v}$'s?
2. Justify the corollaries stated in this section.
3. Is it possible that I have 4 vectors in $\mathcal{R}^4$ which are linearly dependent but any three of them are linearly independent?
4. If the columns of matrix $\mathbf{A}$ are linearly independent and $\mathbf{y}$ is some vector, then must there exist a vector $\mathbf{z}$ such that $\mathbf{Az} = \mathbf{y}$? (Your answer should depend on the size and number of vectors in $\mathbf{A}$).
5. Is $\mathbf{ABx}$ a linear combinations of the columns of $\mathbf{A}$ or $\mathbf{B}$ or both?

3 Rank and Inverse

DEFINITION: The maximal number of linearly independent columns of a matrix is known as the column rank of a matrix.

It should be clear that the corresponding columns will span the column space of the matrix and will form a basis for this vector space.

DEFINITION: Similarly, the row rank of a matrix refers to the maximal number of linearly independent rows in the matrix.

I now wish to provide a proof of a very useful theorem; hopefully you will see how many of the concepts developed earlier (partitioned matrices, basis, column space) are used in this proof.

Theorem 1 (Rank Equivalence Theorem) *For any $m \times n$ matrix $\mathbf{A}$, row rank r = column rank c.*

Proof: Since rearrangement of rows (columns) does not change the fact that a set of columns (rows) are linearly independent, I will for notational convenience, and without loss of generality assume that the first r rows and the first c columns of $\mathbf{A}$ are linearly independent. Consider the $r \times n$ matrix $\tilde{\mathbf{A}}$ obtained by discarding the last $m-r$ rows. I claim that column rank of $\tilde{\mathbf{A}} = c$.[8] To see this, think of $\mathbf{A}$ in terms of partitions:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \tag{17}$$

where $\mathbf{A}_{11}$ is $r \times c$, $\mathbf{A}_{12}$ is $r \times (n-c)$, $\mathbf{A}_{21}$ is $(m-r) \times c$ and $\mathbf{A}_{22}$ is $(m-r) \times (n-c)$. Note that $\tilde{\mathbf{A}} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \end{bmatrix}$.

We observe that there exists a $(m-r) \times r$ matrix $\mathbf{B}$ such that

$$\mathbf{B} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \tag{18}$$

since that is exactly what is meant by saying that the last $m-r$ rows of $\mathbf{A}$ can be expressed as linear combinations of the first r rows.[9] This, in turn, implies that $\mathbf{A}_{21} = \mathbf{B}\mathbf{A}_{11}$. Now, if the columns of $\mathbf{A}_{11}$ were <u>not</u> linearly independent, then there would exist an $\mathbf{x} \in \mathcal{R}^{\mathbf{c}}$ such that $\mathbf{A}_{11}\mathbf{x} = \mathbf{0}$. But that would imply

$$\begin{bmatrix} \mathbf{A}_{11} \\ \mathbf{A}_{21} \end{bmatrix} \mathbf{x} = \mathbf{0} \tag{19}$$

since $\mathbf{A}_{21} = \mathbf{B}\mathbf{A}_{11}$. But this in turn will imply that the first c columns of $\mathbf{A}$ were not linearly independent, a contradiction. Hence, our claim that the first c columns of $\tilde{\mathbf{A}}$ are linearly independent must be true. Now each column of $\tilde{\mathbf{A}}$ is an r-vector. Hence, $c \leq r$ since, there can at most be r linearly independent vectors in $\mathcal{R}^r$. By an exactly analogous argument, one can show that $r \leq c$. Hence, $r = c$. ♣

Here is another useful result:

Theorem 2 (Product Rank Theorem) *Let $\mathbf{A}$, $\mathbf{B}$ be two matrices such that the product $\mathbf{AB}$ is defined. Then rank$(\mathbf{AB}) \leq$ min(rank$(\mathbf{A})$, rank$(\mathbf{B})$).*

[8] Generally speaking, if you have a set of linearly independent vectors and you chop off some coordinates, you may lose linear independence. However, I claim that the first c columns are still linearly independent.

[9] Thus if $\mathbf{B}_{1.}$ represents the first row of $\mathbf{B}$, then the $(r+1)$-th row is given by $\mathbf{B}_{1.} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \end{bmatrix}$.

Proof: Since the columns of $\mathbf{AB}$ are linear combinations of the columns of $\mathbf{A}$, it follows that $\mathcal{C}(\mathbf{AB}) \subset \mathcal{C}(\mathbf{A})$ and column rank of $\mathbf{AB}$ is less than or equal to the column rank of $\mathbf{A}$. Similarly, the row rank of $\mathbf{AB}$ is less than equal to the row rank of $\mathbf{B}$. Since row rank equals column rank of any matrix by the previous theorem, the result follows. ♣

DEFINITION: An $n \times n$ matrix $\mathbf{A}$ is said to have an *inverse* in a matrix $\mathbf{B}$ if $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$ where $\mathbf{I}$ is the $n \times n$ identity matrix.

Inverses are intimately related to finding solutions to equation systems. An n-equation system in n variables, as you know, can be represented in matrix notation as $\mathbf{Ax} = \mathbf{b}$ where $\mathbf{x}$ is the vector of unknowns, $\mathbf{A}$ a matrix of coefficients and $\mathbf{b}$ a vector of right hand side. If we have an inverse of $\mathbf{A}$ available, we can quickly express our solution as $\mathbf{x} = \mathbf{A^{-1}b}$. In this regard, the following is a well-known fact:

Fact 1 *If an $n \times n$ matrix $\mathbf{A}$ has rank n, then it has an inverse. Conversely, if the matrix is invertible (also called nonsingular), it has rank n.*

Proof: For $i = 1, \ldots n$, consider the matrix equation $\mathbf{Ab_i} = \mathbf{e_i}$ where $\mathbf{e_i}$ refers to a unit n-vector with a 1 in the i-th coordinate and 0 everywhere else. Since the n columns of $\mathbf{A}$ are linearly independent, they form a basis of $\mathcal{R}^n$ which includes $\mathbf{e_i}$; hence a $\mathbf{b_i}$ satisfying the above matrix equation must exist. Now stack those $\mathbf{b_i}$ vectors side by side and call the resulting matrix $\mathbf{B}$. It is clear that $\mathbf{AB} = \mathbf{I}$. By a similar argument, but this time making use of the fact that rows of $\mathbf{A}$ are linearly independent and hence form a basis of $\mathcal{R}^n$, one can see that there exists a matrix $\mathbf{C}$ such that $\mathbf{CA} = \mathbf{I}$. Postmultiplying this equation by $\mathbf{B}$ we get $\mathbf{CAB} = \mathbf{B}$, which implies $\mathbf{C} = \mathbf{B}$. This proves the first half of the fact. To see the other half, let us assume that $\mathbf{A}$'s rank is less than n, i.e. its columns are linearly dependent. This implies there exists a non-zero vector $\mathbf{x}$ with the property $\mathbf{Ax} = \mathbf{0}$. Multiplying both sides of this equation by $\mathbf{A^{-1}}$ shows $\mathbf{x} = \mathbf{0}$, a contradiction. Hence, etc... ♣

Now, let me state another useful fact.

Fact 2 *For any matrix, $\mathbf{A}$, rank $(\mathbf{AB})$ = rank $(\mathbf{A})$ where $\mathbf{B}$ is a nonsingular matrix and the product $\mathbf{AB}$ is defined.*

Proof: First observe that rank $(\mathbf{AB}) \leq$ rank $(\mathbf{A})$ on one hand (by the product rank theorem), and also, rank $(\mathbf{A})$ = rank $(\mathbf{ABB^{-1}}) \leq$ rank $(\mathbf{AB})$, also by the product rank theorem. Hence, the result follows. ♣

Lastly, comes the main theorem of this section, of some interest to econometricians. As we will see later in this topic, the classical least squares estimates of a linear model can be expressed in terms of a system of equations, which expressed in matrix notation, features $\mathbf{X}'\mathbf{X}$ as the coefficient matrix. Naturally, we are interested in knowing if this matrix has an inverse, and *what condition(s) on the* $\mathbf{X}$ *matrix will guarantee that.* The following result, which you should try and prove, is useful in this regard.

Theorem 3 *For any matrix* $\mathbf{X}$*, rank* $(\mathbf{X}'\mathbf{X})$ = *rank* $(\mathbf{X})$*.*

To see the usefulness of the theorem, notice that the matrix $(\mathbf{X}'\mathbf{X})$ is non-invertible precisely when the matrix $\mathbf{X}$ has less than full column rank. Next, ask yourself, how an applied researcher may inadvertently use a model where the design matrix may indeed have less than full column rank.

We close this section with a few properties of matrix inverses. You should be able to verify the first two these without much trouble. The proof of the third, which asks you to provide formula(s) for the inverse of a partitioned matrix, is not hard, but tedious. Write down the inverse of a partitioned matrix also in partitioned form and obtain a bunch of equations following the definition of a matrix inverse. Using these equations, you can solve for the submatrices in the partitioned form of the inverse matrix.

Property 1. (Inverse of Transpose is Transpose of Inverse) $(\mathbf{A}')^{-1} = (\mathbf{A^{-1}})'$.

Property 2. (Inverse of Product is Product of Inverse with Order Reversed) $(\mathbf{AB})^{-1} = \mathbf{B^{-1}A^{-1}}$.

Property 3. (Inverse of a Partitioned matrix) Let $\mathbf{A} = \begin{bmatrix} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{22}} \end{bmatrix}$. Then $\mathbf{A^{-1}}$ can be written as

$$\mathbf{A^{-1}} = \begin{bmatrix} \mathbf{B_{11}} & -\mathbf{B_{11}A_{12}A_{22}^{-1}} \\ -\mathbf{A_{22}^{-1}A_{21}B_{11}} & \mathbf{A_{22}^{-1}} + \mathbf{A_{22}^{-1}A_{21}B_{11}A_{12}A_{22}^{-1}} \end{bmatrix} \tag{20}$$

where $\mathbf{B_{11}} = (\mathbf{A_{11}} - \mathbf{A_{12}A_{22}^{-1}A_{21}})^{-1}$, or as

$$\mathbf{A^{-1}} = \begin{bmatrix} \mathbf{A_{11}^{-1}} + \mathbf{A_{11}^{-1}A_{12}B_{22}A_{21}A_{11}^{-1}} & -\mathbf{A_{11}^{-1}A_{12}B_{22}} \\ -\mathbf{B_{22}A_{21}A_{11}}^{-1} & \mathbf{B_{22}} \end{bmatrix} \tag{21}$$

where $\mathbf{B_{22}} = (\mathbf{A_{22}} - \mathbf{A_{21}A_{11}^{-1}A_{12}})^{-1}$.

Section End Questions

1. Show that chopping off some coordinates of a set of linearly indepenedent vectors can cause them to become linearly dependent.

2. Each column of the matrix
$$\mathbf{A} = \begin{bmatrix} 3 & 2 \\ 4 & 3 \end{bmatrix} \tag{22}$$
can be written as linear combinations of the columns of the matrix
$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \tag{23}$$
Verify that. Hence, write **A** in terms of **B** postmultiplied by some matrix.

3. Each row of the matrix
$$\mathbf{A} = \begin{bmatrix} 3 & 2 \\ 4 & 3 \end{bmatrix} \tag{24}$$
can be written as linear combinations of the rows of the matrix
$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \tag{25}$$
Verify that. Hence, write **A** in terms of **B** premultiplied by some matrix.

4 Determinant and Trace

Two important scalar functions of matrices are determinants and trace, both of which come useful when we study the distribution of random *vectors.* In Topic 2, we will study the multivariate normal density function which will feature what is known as the variance-covariance matrix and its determinant. We will also study, in the context of hypothesis testing, distribution of random quadratic forms, which look like $\mathbf{x}'\mathbf{A}\mathbf{x}$, with $\mathbf{x}$ being a normally distributed random vector and **A** being an idempotent matrix. A result which says that the rank of an idempotent matrix is equal to its trace, to be discussed in connection with eigenvalues, will help immensely in the study of such quadratic forms.

4.1 Determinant

The determinant of a matrix **A**, written sometimes as $det(\mathbf{A})$ and sometimes as $|\mathbf{A}|$ is one of the fundamental concepts in linear algebra. Most people learn about determinants by looking at 2 x 2 examples first. For instance they are told that
$$det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11} \times a_{22} - a_{12} \times a_{21}. \tag{26}$$

They then are told that determinants for 3 x 3 matrices can be figured out by calculating 2 x 2 determinants. For instance,

$$det\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = a_{11} \times det\begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} - a_{12} \times det\begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{32} \end{bmatrix} + a_{31} \times det\begin{bmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \tag{27}$$

And generally, one can find the determinant of an $n \times n$ matrix, by expanding along the first row or first column. For example, if you are expanding along the first row, you can write

$$det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{1+j}\, a_{1j}\, det(\bar{\mathbf{A}}_{1j}) \tag{28}$$

where $\bar{\mathbf{A}}_{1j}$ is the $n-1 \times n-1$ matrix you obtain from $\mathbf{A}$ by removing its first row and jth column.

Even more generally, if you are expanding along the i th row, you can write

$$det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{i+j}\, a_{ij}\, det(\bar{\mathbf{A}}_{ij}) \tag{29}$$

where $\bar{\mathbf{A}}_{ij}$ is the $n-1 \times n-1$ matrix you obtain from $\mathbf{A}$ by removing its ith row and jth column.

There is nothing special about expanding along rows. You can get the determinant by expanding along the jth column as well.

$$det(\mathbf{A}) = \sum_{i=1}^{n} (-1)^{i+j}\, a_{1j}\, det(\mathbf{A}_{ij}). \tag{30}$$

Here is a different (but equivalent) definition that is often better suited to proving properties of determinants.

DEFINITION: The *determinant* of a $n \times n$ square matrix $\mathbf{A}$, written as $|\mathbf{A}|$ or det $\mathbf{A}$ is defined as

$$|\mathbf{A}| = \sum_{\text{All permutations}\,(j_1,j_2,\ldots j_n)\,\text{of}\,(1,2,\ldots n)} (-1)^{\phi(j_1,j_2,\ldots j_n)} (a_{1j_1}\, a_{2j_2}\, \cdots\, a_{nj_n}) \tag{31}$$

where a_{pq} refers to the p-th row and q-th column element of $\mathbf{A}$, and $\phi(j_1, j_2, \ldots j_n)$ refers to the number of *inversions* in the permutation $(j_1, j_2, \ldots j_n)$ which is the total number of times a larger number appears before a smaller number.

The above appears to be a formidable definition; so let us take it apart slowly. First of all, a permutation of $(1, 2, \ldots, n)$ is simply a rearrangement of $(1, 2, \ldots, n)$; so $(3, 1, 2)$, for instance, is a permutation of $(1, 2, 3)$. In this permutation, the number of inversions is 2 (3 comes before 1 and 3 comes before 2; 1 coming before 2 doesn't create an inversion). Hence, $\phi(3, 1, 2) = 2$. Thus a typical member of the above summation will be $(-1)^2$ times the product of a_{13}, a_{21} and a_{32} ($j_1 = 3$, $j_2 = 1$, $j_3 = 2$). Hence, for a 2×2 matrix $\mathbf{A}$, it follows that the formula for determinant is

$$\begin{aligned} |\mathbf{A}| &= (-1)^0 a_{11}\, a_{22} + (-1)^1 a_{12}\, a_{21} \\ &= a_{11}\, a_{22} - a_{12}\, a_{21} \end{aligned} \tag{32}$$

and the formula for the determinant of a 3×3 matrix is

$$\begin{aligned} |\mathbf{A}| &= (-1)^0 a_{11}\, a_{22}\, a_{33} + (-1)^1 a_{11}\, a_{23}\, a_{32} + (-1)^1 a_{12}\, a_{21}\, a_{33} + (-1)^2 a_{12}\, a_{23}\, a_{31} \\ &\quad +(-1)^2 a_{13}\, a_{21}\, a_{32} + (-1)^3 a_{13}\, a_{22}\, a_{31} \\ &= a_{11}\, a_{22}\, a_{33} - a_{11}\, a_{23}\, a_{32} + -a_{12}\, a_{21}\, a_{33} + a_{12}\, a_{23}\, a_{31} + a_{13}\, a_{21}\, a_{32} - a_{13}\, a_{22}\, a_{31} \end{aligned} \tag{33}$$

Comparing the last two equations with equations (26) and (27) you can see the equivalence between the two definitions; you can also see how to build higher-order determinants from lower order ones.

Here are some important properties of determinants, all of which are verifiable from the given definition.

Property 1. $|\mathbf{A}'| = |\mathbf{A}|$.
Property 2. If every element of a row or column of a matrix is multiplied by k, then the determinant gets multiplied by k.
Property 3. If matrix $\mathbf{B}$ is obtained by interchanging two rows (or columns) of matrix $\mathbf{A}$, then, $|\mathbf{B}| = -|\mathbf{A}|$.
Property 4. If matrix $\mathbf{B}$ is obtained from matrix $\mathbf{A}$ by adding to a row (column) a scalar multiple of another row (column), then, $|\mathbf{B}| = |\mathbf{A}|$.
Property 5. The rows (columns) of $\mathbf{A}$ are linearly independent if and only if $|\mathbf{A}| \neq 0$.
Property 6. $|\mathbf{AB}| = |\mathbf{A}|\,|\mathbf{B}|$.
Property 7. If a square matrix $\mathbf{A}$ is expressible as $\begin{bmatrix} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{22}} \end{bmatrix}$, where $\mathbf{A_{11}}, \mathbf{A_{22}}$ are themselves square matrices, then $|\mathbf{A}| = |\mathbf{A_{11}}|\,|\mathbf{A_{22}}|$ if either $\mathbf{A}_{12} = \mathbf{0}$ or $\mathbf{A}_{21} = \mathbf{0}$.

4.2 Trace

DEFINITION: The *trace* of a (square) matrix, is simply the sum of its diagonal elements.

The following, easily verifiable property of trace will prove quite useful later.

Property 1. Assuming $\mathbf{A}$ and $\mathbf{B}$ are matrices such that both $\mathbf{AB}$ and $\mathbf{BA}$ are defined, trace $(\mathbf{AB})$ = trace $(\mathbf{BA})$.
It follows that
Property 2. For matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, where the following products are well-defined,
trace($\mathbf{ABC}$) = trace($\mathbf{BCA}$) = trace($\mathbf{CAB}$).

Section End Questions

1. How many terms are there in the expression for a determinant of a 4 x 4 matrix? How many of them are positive and how many negative?

2. Calculate the determinant of the following matrix without using a calculator

$$\mathbf{A} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 5 \\ 0 & 2 & 6 \end{bmatrix}$$

3. If all the diagonal elements of two matrices are 0, does that mean that the trace of their product is 0?

5 Orthogonality, Projection and the Least Squares Problem

DEFINITION: The standard or Euclidean *norm* or the length of a vector $\mathbf{v} = (v_1, \ldots v_n)'$ sitting inside the n-dimensional Euclidean space is written as $\| \mathbf{v} \|$ and defined as $(v_1^2 + \ldots v_n^2)^{\frac{1}{2}}$, which of course, simply measures the distance from the origin to the tip of the vector.

DEFINITION: The *inner product* or the *dot product* of two vectors given by $\mathbf{u} = (u_1, \ldots v_n)'$ and $\mathbf{v} = (v_1, \ldots v_n)'$, written as $\mathbf{u}.\mathbf{v}$ or $\mathbf{u}'\mathbf{v}$ is defined as $u_1 v_1 + \ldots u_n v_n$. We say $\mathbf{u}$ and $\mathbf{v}$ are *orthogonal* if $\mathbf{u}.\mathbf{v} = 0$ (sometimes written as $\mathbf{u} \perp \mathbf{v}$). The following fact can be easily verified using basic algebraic manipulation:

Fact 3 $\| \mathbf{u} + \mathbf{v} \|^2 = \| \mathbf{u} \|^2 + \| \mathbf{v} \|^2 + 2\,\mathbf{u}.\mathbf{v}$

Hence, in the spirit of Pythagorus's theorem, it does make sense to call $\mathbf{u}$ and $\mathbf{v}$ orthogonal as in this case, the square of the length of $\mathbf{u} + \mathbf{v}$ (which is the third side of the triangle formed with $\mathbf{u}$ and $\mathbf{v}$ as adjacent sides) becomes the sum of the squared lengths of $\mathbf{u}$ and $\mathbf{v}$.

The least squares method of calculating $\mathbf{b}$, estimate of the β coefficients in our model given by equation (7) is to choose $\mathbf{b}$ so as to minimize the length of the vector $\| \mathbf{y} - \mathbf{Xb} \|$, or equivalently, "the sum of the squared errors". We will later talk about the virtues of this criterion, but first, let us take a minute to understand the criterion. Suppose, after examining the data, we arrive at the following estimates of the beta coefficients: $b_1, b_2, \ldots b_k$. Then, ex-post for the i-th observation, the 'unexplained' error (of the model from the observation) is $e_i = y_i - b_1 - b_2 x_{i2} - \ldots b_k x_{ik}$[10]. We square this error and sum it across observations, and the resulting object, which is called RSS (residual sum of squares) is what we seek to minimize (by choosing proper values of $(b_1, \ldots b_k)$.

For instance, suppose we have a very simple bivariate model of regression, i.e. there are two right hand side variables: 1 (representing the intercept) and some other variable called x. Let us say we have three observations which in terms of the (x, y) values are: (1,3), (3,8) and (4,7). In the diagram below these three points are plotted (the diamonds). What we are trying to do in regression analysis is to postulate that $y = \beta_1 + \beta_2 x$ is a good enough chracterization of the relationship between y and x and come up with estimates of β_1 and β_2 (which we will call b_1 and b_2). Choosing particular values of b_1 and b_2 is essentially the same as fixing a line in the $x - y$ graph. For any such line, we define the 'error' of a particular observation e_i as $y_i - (b_1 - b_2 x_i)$. Thus, if the estimated intercept (b_1) was 2 and the estimated slope (b_2) was 1.5, then $e_1 = -.5, e_2 = 1.5$ and $e_3 = -1$. These are the distances from the diamonds to the squares - the predicted values (see figure below). Now, the RSS for this particular choice of the parameter estimates $(b_1 = 2, b_2 = 1.5)$ is 3.5. As it turns out, this is indeed the optimal choice; if we chose another line (say given by $y = 1 + 2x$, the RSS we will get will be higher(5)). Hence, we are indeed looking at the least squares line in the picture below.

[10] Note that e_i and ε_i are very different animals. We will never know the actual realization of ε_i, but e_i is determined once we choose certain estimated parameters. Also note that $\mathbf{y}_i$ can be written in two different ways: $\mathbf{y}_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ (which represents the true model) as well as $\mathbf{y}_i = \mathbf{x}_i'\mathbf{b} + e_i$ (which represents the fitted model).

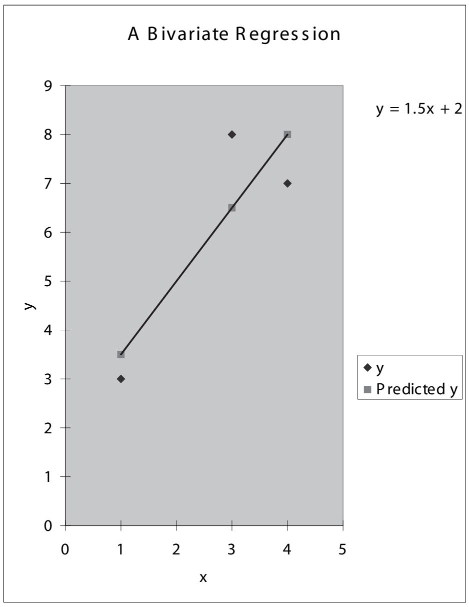

Figure 1. Regression Line and Errors

While the diagram above gives us a sense of what is happening with bivariate regression, it gives us little intuition about how we should go about choosing the parameters in situations where there are many right hand side variables. As it so happens, to get some intuition about the multivariate situation, we have to look at a different sort of a diagram; one which can house the $\mathbf{y}$ vector and each column (vector) of the $\mathbf{X}$ matrix. Such a diagram is Figure 2, where $\mathbf{X_{.j}}$ represents the jth column of the $\mathbf{X}$ matrix. Hence, we are talking about n-dimensional space here.

Now, since the vector of errors is just $\mathbf{e} = \mathbf{y} - \mathbf{Xb}$, minimizing RSS $= \sum_{i=1}^{n} e_i^2$ is the same as minimizing $\| (\mathbf{y} - \mathbf{Xb}) \|^{\mathbf{2}}$.[11] This is the famous Least Squares Problem. Noting that any vector of the form $\mathbf{Xb}$ is just a member of the $\mathcal{C}(X)$, let us now break the problem in two stages:

Stage 1. First find $\hat{\boldsymbol{\mu}}$ such that $\hat{\boldsymbol{\mu}}$ solves: Min $\| (\mathbf{y} - \hat{\boldsymbol{\mu}}) \|^2$ subject to $\hat{\boldsymbol{\mu}} \in \mathcal{C}(X)$.

Stage 2. Next find $\mathbf{b}$ such that $\hat{\boldsymbol{\mu}} = \mathbf{Xb}$.

[11] Of course, minimizing $\| (\mathbf{y} - \mathbf{Xb}) \|^{\mathbf{2}}$ and minimizing $\| (\mathbf{y} - \mathbf{Xb}) \|$ are completely equivalent exercises.

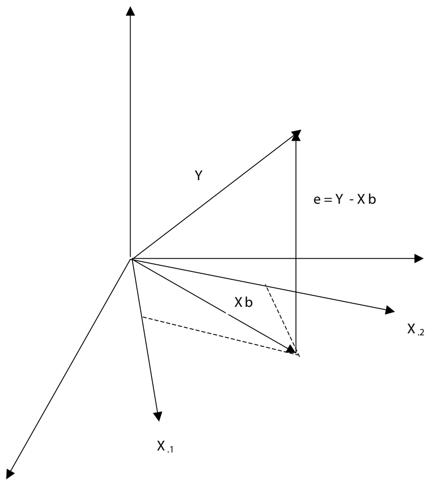

Figure 2. The Geometry of Projection

The first task is facilitated by the following result.

Theorem 4 (Projection Theorem) *Let* $\mathbf{y} \in \mathcal{R}^{\mathbf{n}}$ *and* $S \subset \mathcal{R}^n$ *be a linear subspace. Then* $\hat{\boldsymbol{\mu}}$ *solves Min* $\| (\mathbf{y} - \hat{\boldsymbol{\mu}}) \|^2$ *subject to* $\hat{\boldsymbol{\mu}} \in S$ *if and only if* $\mathbf{y} - \hat{\boldsymbol{\mu}}$ *is orthogonal to* S*, i.e. for any* $\mathbf{z} \in S$*,* $(\mathbf{Y} - \hat{\boldsymbol{\mu}})'\mathbf{z} = \mathbf{0}$*. Moreover, such a* $\hat{\boldsymbol{\mu}}$ *is unique.*

Proof: (If part) We show that if $(\mathbf{y} - \hat{\boldsymbol{\mu}}) \perp S$, then $\| \mathbf{y} - \hat{\boldsymbol{\mu}} \|^2 \leq \| \mathbf{y} - \tilde{\boldsymbol{\mu}} \|^2$ for all $\tilde{\boldsymbol{\mu}} \in S$. Now,

$$\begin{aligned}
\| \mathbf{y} - \tilde{\boldsymbol{\mu}} \|^2 &= \| \mathbf{y} - \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}} \|^2 \\
&= \| \mathbf{y} - \hat{\boldsymbol{\mu}} \|^2 + \| \hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}} \|^2 \text{ since } \hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}} \in S \text{ and } \mathbf{y} - \hat{\boldsymbol{\mu}} \text{ is orthogonal to S} \\
&\geq \| \mathbf{y} - \hat{\boldsymbol{\mu}} \|^2
\end{aligned}$$

(Only if part) Suppose $\hat{\boldsymbol{\mu}}$ is indeed the minimizer but $\mathbf{y} - \hat{\boldsymbol{\mu}}$ is not orthogonal to S. In particular suppose there exists a $\boldsymbol{\delta}$ in S such that $\boldsymbol{\delta}'(\mathbf{y} - \hat{\boldsymbol{\mu}}) \neq 0$. Define $\tilde{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}} + \frac{\boldsymbol{\delta}'(\mathbf{y} - \hat{\boldsymbol{\mu}})}{\boldsymbol{\delta}'\boldsymbol{\delta}} \boldsymbol{\delta}$ (notice that the fraction there is one scalar divided by another scalar). You should be now able to show (do it!) $\tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}$ is orthogonal to $(\mathbf{y} - \tilde{\boldsymbol{\mu}})$ and hence the vector $\mathbf{y} - \hat{\boldsymbol{\mu}}$ has a strictly larger length that $\mathbf{y} - \tilde{\boldsymbol{\mu}}$ which creates a contradiction.

(Uniqueness) The idea of this proof is almost identical to that of the if part. Suppose there exists two minimizers to the problem mentioned in the statement of the theorem: $\hat{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\mu}}$.

Since,

$$\| \mathbf{y} - \tilde{\boldsymbol{\mu}} \|^2 = \| \mathbf{y} - \hat{\boldsymbol{\mu}} \|^2 + \| \hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}} \|^2$$

and $\| \mathbf{y} - \tilde{\boldsymbol{\mu}} \|^2 = \| \mathbf{y} - \hat{\boldsymbol{\mu}} \|^2$, it must be the case that $\| \hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}} \|^2 = 0$ which can only happen if $\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}$. ♣

Putting together all the above insights, we realize that any solution $\mathbf{b}$ to the least squares problem has the property that the vector of error term $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$ must be orthogonal to column span of $\mathbf{X}$, which it will be if it is orthogonal to each of the columns of $\mathbf{X}$. This results in the following characterization of $\mathbf{b}$:

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0} \tag{34}$$

or,

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b} \tag{35}$$

The above equation is called a normal equation.[12] Now, a solution of the above equation exists uniquely, if $\mathbf{X}'\mathbf{X}$ has an inverse. The results discussed in Section 3 tell you exactly when that happens: it happens when the $\mathbf{X}$ matrix has the full column rank k. Under such circumstance, we can write the solution to the least squares problem as[13]

$$\mathbf{b_{OLS}} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y} \tag{36}$$

Memorize this formula like its your mother's birthday or that of your significant other. It will be really embarrassing later if you don't remember it.

Section End Questions

1. Can n pairwise orthogonal vectors in $\mathcal{R}^n$, none of which is a zero vector be linearly dependent?
2. Find a vector that is orthogonal to both $(1, 2, 3, 4)'$ and $(5, 6, 7, 8)'$.
3. Is the space of all vectors that are orthogonal to the above two vectors a subspace? If so, what is its dimension?
4. Suppose the $\mathbf{X}$ matrix does not have full column rank. Does it mean that the least squares problem has no solution?

[12] It is so called since the equation is characterized in terms of a normal or perpendicular to a plane.

[13] Below, the acronym OLS stands for Ordinary Least Squares. Why use the qualifier Ordinary and not just Least Squares? Because there is a GLS or a Generalized Least Squares estimator also!

6 Matrix Differentiation

An alternative technique of deriving the OLS formula is to use the matrix differentiation approach which you will find useful in many multivariable optimization context (we will use it to derive an expression of restricted least squares formula in a few weeks).

Let $\mathbf{x}$ be a vector in $\mathcal{R}^n$ and let $f : \mathcal{R}^n \mapsto \mathcal{R}$ be a scalar-valued function which takes n-vectors as arguments. Then by $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$ we mean the vector

$$\begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

and call it the gradient vector of f (Note that some authors like to think of it as a row vector). On the other hand, the notation $\frac{\partial \mathbf{f}}{\partial \mathbf{x}'}$ will simply denote the same vector as above but laid out as a row. Here are two useful facts.

Fact 4 $\frac{\partial}{\partial \mathbf{x}}(\mathbf{a}'\mathbf{x}) = \mathbf{a}$ and $\frac{\partial}{\partial \mathbf{x}'}(\mathbf{a}'\mathbf{x}) = \mathbf{a}'$.

Fact 5 $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}')\mathbf{x}$.

Fact 4 can be easily verified. To see Fact 5, write $\mathbf{x}'\mathbf{A}\mathbf{x}$ as $\sum_{i=1}^n \sum_{j=1}^n x_i a_{ij} x_j$. Hence,

$$\frac{\partial}{\partial x_i}(\mathbf{x}'\mathbf{A}\mathbf{x}) = 2x_i a_{ii} + \sum_{j \neq i} a_{ij} x_j + \sum_{j \neq i} a_{ji} x_j \tag{37}$$

On the other hand, consider the i-th row of $(\mathbf{A} + \mathbf{A}')\mathbf{x}$, which is

$$\sum_{j=1}^n a_{ij} x_j + \sum_{j=1}^n a_{ji} x_j \tag{38}$$

If you compare the two expressions above, you will find that they are exactly the same, which is what the Fact is asserting. From Fact 5, the following is immediate.

Fact 6 *If* $\mathbf{A}$ *is symmetric,* $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{A}\mathbf{x}) = (\mathbf{2}\mathbf{A}\mathbf{x})$.

Now let us attack the least squares problem of minimizing $f(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ by choosing $\mathbf{b}$ optimally. You know what to do - differentiate the objective with respect to the optimizing variable, in this case a vector and set it equal to zero. But before we do that, let us rewrite the objective:

$$\begin{aligned} f(\mathbf{b}) &= (\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb}) \\ &= (\mathbf{y}'-\mathbf{b}'\mathbf{X}')(\mathbf{y}-\mathbf{Xb}) \\ &= \mathbf{y}'\mathbf{y}-\mathbf{y}'\mathbf{Xb}-\mathbf{b}'\mathbf{X}'\mathbf{y}+\mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b} \\ &= \mathbf{y}'\mathbf{y}-\mathbf{2y}'\mathbf{Xb}+\mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b} \quad \text{since } \mathbf{y'Xb},\ \mathbf{b'X'y} \text{ are both the same scalar} \\ &= \mathbf{y}'\mathbf{y}-\mathbf{2}(\mathbf{X}'\mathbf{y})'\mathbf{b}+\mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b} \qquad (39) \end{aligned}$$

Now it is easy to see that the derivative of this with respect to **b** is

$$-\mathbf{2}(\mathbf{X}'\mathbf{y})+\mathbf{2}(\mathbf{X}'\mathbf{X})\mathbf{b} \qquad (40)$$

which when set equal to the **0** vector gives us the same formula for the least squares estimates (equation (36)) as in the last section (verify!).

One needs to check second order conditions too in order to insure that we have indeed minimized the objective. In order to discuss that I now need to introduce the notion of derivative of one vector with respect to another vector. If **x** and **y** are both (column) vectors with each element of **y** being a function of all the elements of **x**, we will write $\frac{\partial \mathbf{y}}{\partial \mathbf{x}'}$ to mean the matrix the ij-th element of which is $\frac{\partial y_i}{\partial x_j}$. Similarly the notation $\frac{\partial \mathbf{y}'}{\partial \mathbf{x}}$ simply refers to the transpose of the previous matrix. Now, it should be straightforward to verify the following:

Fact 7 $\frac{\partial}{\partial \mathbf{x}'}(\mathbf{Ax}) = \mathbf{A}$ and $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{A}) = \mathbf{A}$.

Let us go back to the problem of minimizing $f(\mathbf{b})$ with respect to **b**. In order to do this the second order condition involves creating the Hessian matrix $\frac{\partial^2 f}{\partial \mathbf{b}\partial \mathbf{b}'}$ and checking that this matrix is *positive definite.*

DEFINITION: A matrix **A** is positive (negative) definite if for any nonzero vector **x**, the quadratic form $\mathbf{x}'\mathbf{Ax}$ is positive (negative). Also, a matrix **A** is positive (negative) semidefinite if for any vector **x**, the quadratic form $\mathbf{x}'\mathbf{Ax}$ is nonnegative (nonpositive).

Now, it is easy to check that the Hessian matrix for our problem is $2(\mathbf{X}'\mathbf{X})$, and this matrix is indeed positive definite if **X** has full column rank (why?). Thus $\mathbf{b_{OLS}}$ indeed minimizes the sum of squares.

Section End Questions

1. Check that the Hessian matrix of f with respect to **b**, the ijth entry is $\frac{\partial^2 f}{\partial b_i \partial b_j}$ is indeed $\frac{\partial^2 f}{\partial \mathbf{b}\partial \mathbf{b}'}$.

2. Verify Fact 7.
3. Why is the matrix $\mathbf{X}'\mathbf{X}$ positive definite if $\mathbf{X}$ has full column rank?
4. In this section, we have discussed unconstrained optimization using matrix differentiation techniques. But the method of Lagrange multiplier can be used for unconstrained optimization as well (if there are multiple constraints, the multiplier will be a vector). Use what you have learnt in this section to solve the following optimization problem: Maximize $5x + 3y + 4z$ subject to $x^2 + y^2 + z^2 = 2$ and verify your answer using EXCEL's SOLVER.

7 Eigenvalues, Eigenvectors and More on Quadratic Forms

DEFINITION: Let $\mathbf{A}$ be a (real-valued) $n \times n$ square matrix. Then a (possibly compex-valued[14]) scalar λ is said to be an eigenvalue of $\mathbf{A}$, if there is a non-zero n-vector (possibly complex-valued) such that

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \tag{41}$$

The study of eigenvalues and eigenvectors will lead us to a very important result called the Spectral Decomposition Theorem. Among other things, it will tell us how to take the 'square root' of certain matrices, and will be of immense help when we study multivariate normal distributions and quadratic forms derived from such distributions. These results in turn will form the backbone of hypothesis testing in the context of linear models.

The definition above leads us naturally to the computation of eigenvalues and eigenvectors. If λ is an eigenvalue, from the definition above it follows, that the matrix $\mathbf{A} - \lambda\mathbf{I}$ must be singular (since there exists a nonzero vector that postmultiplies it to give the zero vector). Hence, any eigenvalue of $\mathbf{A}$ must be a root of the determinantal equation

$$|\mathbf{A} - \lambda\mathbf{I}| = 0. \tag{42}$$

This is a polynomial equation in λ that is usually called the characteristic equation. As an example, consider the matrix $\mathbf{A} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$. To find its eigenvalues we set

$$|\mathbf{A} - \lambda\mathbf{I}| = \begin{vmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{vmatrix} = (4-\lambda)(3-\lambda) - 2 = 12 - 7\lambda + \lambda^2 - 2 = 10 - 7\lambda + \lambda^2 = 0$$

Since the roots of the quadratic equation above are 5 and 2, these are the precisely the eigenvalues of $\mathbf{A}$.

[14] Recall that a typical complex number is of the form $a + bi$ where a, b are real numbers and $i = \sqrt{-1}$.

Next, to determine the eigenvectors, make use of the equation $(\mathbf{A}-\lambda\mathbf{I})\mathbf{x}=\mathbf{0}$. For instance, corresponding the eigenvalue 5, this gives us two equations

$$\begin{aligned}-x_1+x_2&=0\\2x_1-2x_2&=0\end{aligned}$$

and we need to solve for x_1, x_2 from these equations. Actually, there is only 1 (independent) equation here and we can obtain multiple solutions. This is typical. Equation (41) for a given eigenvalue λ does not pin down an $\mathbf{x}$ (the solution set is a vector space of at least dimension 1 and can have dimension larger than 1). One solution is $x_1 = x_2 = \frac{1}{\sqrt{2}}$, where we have chosen to 'normalize' the eigenvector (i.e. make its norm or length equal 1). Proceeding similarly, you should have no trouble showing that an eigenvector corresponding to the eigenvalue 2 obeys the equation $2x_1 + x_2 = 0$ and hence, a normalized eigenvector is $\left(\frac{1}{\sqrt{5}}, -\frac{2}{\sqrt{5}}\right)'$.

Two simple facts about eigenvalues are stated next:

Fact 8 *The product of the eigenvalues of a matrix is equal to its determinant.*

Proof: To see the validity of this, consider the characteristic equation (42) and realize that this is the same equation as

$$(\lambda_1-\lambda)(\lambda_2-\lambda)\cdots(\lambda_n-\lambda)=0 \tag{43}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\mathbf{A}$ (note that some of the roots may be repeated). Now, if you consider the coefficient of the constant term, in equation (42) that is just the determinant of the matrix, whereas in (43) it is the product $\lambda_1\lambda_2\cdots\lambda_n$. ♣

Fact 9 *The sum of eigenvalues of a matrix is equal to its trace.*

Proof: Now if you equate the coefficient of λ^{n-1} in equations (42) and (43), you will get Fact (9). ♣

To really understand the logic behind these two (cryptic) proofs, write out the full-blown forms for equations (41) and (42) for a specific 2 x 2 or a 3 x 3 matrix and check for yourself. The next two facts are easy to verify and are left as exercises.

Fact 10 *The eigenvalues of* $\mathbf{A}^2$ *are squares of the eigenvalues of* $\mathbf{A}$ (*the eigenvectors of both are the same*).

Fact 11 *The eigenvalues of* $\mathbf{A}^{-1}$ (*assuming* $\mathbf{A}$ *is nonsingular*) *are reciprocals of the eigenvalues of* $\mathbf{A}$ (*the eigenvectors of both are the same*).

There is no guarantee that eigenvalues are always real-valued, for a polynomial equation with real coefficients can yield complex-valued roots (that occur in conjugate pairs[15]). A very important situation where we are assured of real eigenvalues is provided by the next fact.

Fact 12 *The eigenvalues of a real symmetric matrix are all real (which then guarantees real eigenvectors as well).*

Proof: Let $\mathbf{A}$ be a real, symmetric square matrix, let $\lambda + i\mu$ be one of its eigenvalues, and let $\mathbf{x} + i\mathbf{y}$ be a corresponding eigenvector, where $\lambda, \mu, \mathbf{x}, \mathbf{y}$ are real. Note that both $\mathbf{x}$ and $\mathbf{y}$ can not be zero vectors. We have

$$\mathbf{A}(\mathbf{x} + i\mathbf{y}) = (\lambda + i\mu)(\mathbf{x} + i\mathbf{y}).$$

Equating the real and imaginary parts of both sides, we get two equations:

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} - \mu\mathbf{y}$$

$$\mathbf{A}\mathbf{y} = \lambda\mathbf{y} + \mu\mathbf{x}$$

Premultiply the first of this by $\mathbf{y}'$ and the second by $\mathbf{x}'$ to get

$$\mathbf{y}'\mathbf{A}\mathbf{x} = \lambda\mathbf{y}'\mathbf{x} - \mu\mathbf{y}'\mathbf{y}$$

$$\mathbf{x}'\mathbf{A}\mathbf{y} = \lambda\mathbf{x}'\mathbf{y} + \mu\mathbf{x}'\mathbf{x}$$

But $\mathbf{y}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{A}'\mathbf{y}$ (because the transpose of a scalar is itself) which in turn is $\mathbf{x}'\mathbf{A}\mathbf{y}$ (because $\mathbf{A}$ is symmetric). Hence subtracting the first equation from the second one gives us $\mu(\mathbf{y}'\mathbf{y} + \mathbf{x}'\mathbf{x}) = \mathbf{0}$ (since $\mathbf{y}'\mathbf{x} = \mathbf{x}'\mathbf{y}$). But $(\mathbf{y}'\mathbf{y} + \mathbf{x}'\mathbf{x})$ is nonzero (the only way it can be 0 is for $\mathbf{x}$ and $\mathbf{y}$ to be zero vectors, and eigenvectors can not be zero vectors). Hence, $\mu = 0$. Now, that we know the eigenvalue is real, $(\mathbf{A} - \lambda\mathbf{I})$ being a real, singular matrix, there exists a real vector $\mathbf{x}$, such that $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$. $\clubsuit$

For our next result, we need a definition.

DEFINITION: A real matrix $\mathbf{A}$ is called orthogonal, if its transpose is its inverse, i.e. if $\mathbf{A}\mathbf{A}' = \mathbf{A}'\mathbf{A} = \mathbf{I}$.[16] (Thus, for an orthogonal matrix both its rows and columns are pairwise orthogonal and each column has length 1. We call such a set of vectors *orthonormal*. The same statement is true about the rows of an orthogonal matrix as well.)

Theorem 5 (The Spectral Decomposition Theorem) *Let $\mathbf{A}$ be a real symmetric matrix. Then there exists an orthogonal matrix $\mathbf{X}$, and a diagoinal matrix $\mathbf{\Lambda}$, such that $\mathbf{X}'\mathbf{A}\mathbf{X} = \mathbf{\Lambda}$. Furthermore, the diagonal entries of $\mathbf{\Lambda}$ are the n eigenvalues of $\mathbf{A}$ (possibly repeated) and the matrix $\mathbf{X}$ consists of its (normalized) eigenvectors.*

[15] Recall that this means if $a + ib$ is a root, so is $a - ib$.

[16] Note that for any square matrix $\mathbf{A}$, as soon we establish $\mathbf{A}'\mathbf{A} = \mathbf{I}$, it automatically follows that $\mathbf{A}\mathbf{A}' = \mathbf{I}$.

Proof: The proof is by induction on the size of the matrix. Clearly the result is true for $n = 1$, when $\mathbf{A}$ is actually a scalar, which is the same as its eigenvalue, and $\mathbf{X}$ is just 1. We now show that if the result is true for $n-1$, then it must be true for n as well.

Let λ_1 be an eigenvalue of $\mathbf{A}$ and let $\mathbf{x_1}$ be a corresponding (normalized) eigenvector. Thus, we have $\mathbf{x_1}'\mathbf{x_1} = 1$. Find $n-1$ orthonormal vectors which forms a basis for the space that is orthogonal to $\mathbf{x_1}$ and stack them in a $n \times n-1$ matrix called $\mathbf{B}$.

Consider the matrix $\mathbf{B}'\mathbf{A}\mathbf{B}$. It is a symmetric $n-1 \times n-1$ matrix as can be easily verified. Because we assume that the result is true for $n-1$, there exists a matrix $\mathbf{Y}$, such that $\mathbf{Y}'\mathbf{B}'\mathbf{A}\mathbf{B}\mathbf{Y} = \mathbf{D}$ where $\mathbf{Y}$ is an $n-1 \times n-1$ orthogonal matrix of eigenvectors of $\mathbf{B}'\mathbf{A}\mathbf{B}$, and $\mathbf{D}$ is a diagonal matrix, with the eigenvalues of $\mathbf{B}'\mathbf{A}\mathbf{B}$ along its diagonal. We now claim that the desired $\mathbf{X} = [\mathbf{x_1} \ \ \mathbf{BY}]$, while the $\mathbf{\Lambda}$ is $\begin{bmatrix} \lambda_1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{D} \end{bmatrix}$.

To show this, first we need to show that $\mathbf{X}$ is orthogonal. This involves demonstrating that 1) $\mathbf{x_1}'\mathbf{x_1} = 1$, 2) $\mathbf{x_1}'\mathbf{B}\mathbf{Y} = \mathbf{0}'$ and 3) $(\mathbf{BY})'(\mathbf{BY}) = \mathbf{I}$. 1) follows from the construction of $\mathbf{x}_1$. 2) follows from the fact that $\mathbf{x}_1$ is orthogonal to each vector in $\mathbf{B}$. 3) follows from the fact that $(\mathbf{BY})'(\mathbf{BY}) = \mathbf{Y}'\mathbf{B}'\mathbf{B}\mathbf{Y} = \mathbf{Y}'\mathbf{Y} = \mathbf{I}$, where we have utilized the fact that both the columns of $\mathbf{B}$ and those of $\mathbf{Y}$ are orthonormal.

Next we need to show $\mathbf{X}'\mathbf{A}\mathbf{X} = \mathbf{\Lambda}$, i.e. $\begin{bmatrix} \mathbf{x}_1' \\ \mathbf{Y}'\mathbf{B}' \end{bmatrix} \mathbf{A} \begin{bmatrix} \mathbf{x}_1 & \mathbf{BY} \end{bmatrix} = \mathbf{\Lambda}$ This is true since, the left hand side is

$$\begin{aligned} \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{Y}'\mathbf{B}' \end{bmatrix} \begin{bmatrix} \mathbf{A}\mathbf{x}_1 & \mathbf{ABY} \end{bmatrix} &= \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{Y}'\mathbf{B}' \end{bmatrix} \begin{bmatrix} \lambda_1\mathbf{x}_1 & \mathbf{A}'\mathbf{BY} \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 & \lambda_1\mathbf{x}_1'\mathbf{BY} \\ \lambda_1\mathbf{Y}'\mathbf{B}'\mathbf{x}_1 & \mathbf{D} \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \end{aligned}$$

which is the right hand side. Note that in the first and the second equality above we have used $\mathbf{A}\mathbf{x}_1 = \lambda_1\mathbf{x}_1$ and also the fact that $\mathbf{A} = \mathbf{A}'$, and in the last equality, we have made use of the fact that $\mathbf{x}_1$ is orthogonal to columns of $\mathbf{B}$, and that $\mathbf{Y}'\mathbf{B}'\mathbf{A}\mathbf{B}\mathbf{Y} = \mathbf{D}$.

Now that we know that there is an orthogonal matrix $\mathbf{X}$ and a diagonal matrix $\mathbf{\Lambda}$, with $\mathbf{X}'\mathbf{A}\mathbf{X} = \mathbf{\Lambda}$ it is easy to interpret these objects. First, observe that

$$\begin{aligned} |\mathbf{A} - \lambda\mathbf{I}| = 0 &\iff |\mathbf{X}'||\mathbf{A} - \lambda\mathbf{I}||\mathbf{X}| = 0 \\ &\iff |\mathbf{X}'\mathbf{A}\mathbf{X} - \lambda\mathbf{I}| = 0 \\ &\iff |\mathbf{\Lambda} - \lambda\mathbf{I}| = 0 \end{aligned}$$

which shows that the diagonal elements of $\mathbf{\Lambda}$ are precisely the eigenvalues of $\mathbf{A}$. Lastly observe that the relation $\mathbf{X}'\mathbf{A}\mathbf{X} = \mathbf{\Lambda}$, and the orthogonality of $\mathbf{X}$, allows us to write $\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{\Lambda}$.

This means that if $\mathbf{x}_i$ is the i-th column of $\mathbf{X}$ and λ_i is the ith diagonal element of $\mathbf{\Lambda}$, then, $\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i$, revealing thereby that $\mathbf{x}_i$ is an eigenvector corresponding to the eigenvalue λ_i. $\clubsuit$

One of the usefulness of the theorem stems from the fact that it allows us to compute arbitrary powers of symmetric matrices easily. Since such a matrix $\mathbf{A}$ is expressible as as $\mathbf{X}\mathbf{\Lambda}\mathbf{X}'$, utilizing the orthogonality of $\mathbf{X}$, we can see that $\mathbf{A}^t = \mathbf{X}\mathbf{\Lambda}^t\mathbf{X}'$, and hence, for instance one can see that if all the eigenvalues are strictly less than 1 in absolute value, as t tends to infinity, $\mathbf{A}^t$ tends towards the zero matrix. The theorem also allows us to claim:

Fact 13 *The rank of a symmetric matrix* $\mathbf{A}$ *is equal to the number of its nonzero eigenvalues.*

Proof: Since pre or post multiplication by a nonsingular matrix does not change its rank, $\mathbf{X}'\mathbf{A}\mathbf{X}$ has the same rank as $\mathbf{\Lambda}$ which is the number of nonzero elements along the diagonal of the latter. $\clubsuit$

The next fact is an assertion about idempotent matrices.

Fact 14 *The eigenvalues of any idempotent matrices are 0 or 1.*

Proof: By Fact 10, if λ is an eigenvalue of $\mathbf{A}$, $\lambda^2 = \lambda$. The only solutions to this equation are 0 and 1. $\clubsuit$

Next, we have an assertion about symmetric, idempotent matrices.

Fact 15 *For a symmetric idempotent matrix, its trace is equal to its rank.*

Proof: Let $\mathbf{A}$ be symmetric and idempotent, and admit a decomposition $\mathbf{X}'\mathbf{A}\mathbf{X} = \mathbf{\Lambda}$. Again, because pre or post multiplaication by nonsingular matrices does not change rank, $rank(\mathbf{A}) = rank(\mathbf{\Lambda}) = trace(\mathbf{\Lambda})$ (because of Fact (14)). But $trace(\mathbf{\Lambda}) = trace(\mathbf{X}'\mathbf{A}\mathbf{X}) = trace(\mathbf{A}\mathbf{X}\mathbf{X}') = trace(\mathbf{A})$. $\clubsuit$

We end this section discussing a few results on positive definite and semidefinite matrices.

Fact 16 *A symmetric matrix is positive definite (semidefinite) if and only if all its eigenvalues are strictly positive (nonnegative). Similarly, a symmetric matrix is negative definite (semidefinite) if and only if all its eigenvalues are strictly negative (nonpositive).*

Proof: I will prove the result for positive definite matrices. The arguments for the other three cases (positive semidefinite, negative definite and negative semidefinite) are similar.

Suppose $\mathbf{A}$ is an $n \times n$ symmetric and positive definite. We show that all eigenvalues must be strictly positive. Since $\mathbf{A}$ is symmetric, we have $\mathbf{X'AX} = \mathbf{\Lambda}$, where $\mathbf{X}$ is orthogonal and $\mathbf{\Lambda}$ contains $\mathbf{A}$'s eigenvalues. Suppose the ith diagonal entry of $\mathbf{\Lambda}$, λ_i, is not strictly positive (i.e. it is 0 or negative). Define $\mathbf{y} = \mathbf{X}\mathbf{e}_i$, where $\mathbf{e}_i$ is a unit vector with 1 in its i-th coordinate and 0 elsewhere. Clearly, $\mathbf{y}$ is nonzero. Now, $\mathbf{y'Ay} = \mathbf{e}_i'\mathbf{X'AX}\mathbf{e}_i = \mathbf{e}_i'\mathbf{\Lambda}\mathbf{e}_i = \lambda_i \leq 0$. But this contradicts the positive definiteness of $\mathbf{A}$.

To see the reverse implication, now, suppose all the eigenvalues of $\mathbf{A}$ are strictly positive. Let $\mathbf{y}$ be any nonzero vector. Then, there exists a nonzero vector $\mathbf{z}$ such that $\mathbf{y} = \mathbf{Xz}$ (since, the columns of $\mathbf{X}$, being orthonormal, is a basis for $\mathcal{R}^n$). Now, $\mathbf{y'Ay} = \mathbf{z'X'AXz} = \mathbf{z'\Lambda z} = \sum_i^n \lambda_i z_i^2 > 0$, since each $\lambda_i > 0$ and not all z_i's are 0. This shows $\mathbf{A}$ is positive definite. ♣

Next, we consider the problem of finding 'square roots' of positive semidefinite matrices. The next fact assures the existence of two different types of square roots for such matrices.

Fact 17 *Suppose* $\mathbf{A}$ *is symmetric and positive semidefinite. Then there exists a matrix* $\mathbf{P}$ *such that* $\mathbf{PP'} = \mathbf{A}$. *There also exists a matrix* $\mathbf{Q}$ *such that* $\mathbf{QQ} = \mathbf{A}$.

Proof: Define $\mathbf{\Lambda}^{\frac{1}{2}}$ to be a diagonal matrix containing the square roots of the eigenvalues of $\mathbf{A}$ along its diagonal. It is clear that $\mathbf{\Lambda}^{\frac{1}{2}}$ is symmetric and $\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Lambda}^{\frac{1}{2}} = \mathbf{\Lambda}$. Now define $\mathbf{P} = \mathbf{X\Lambda}^{\frac{1}{2}}$ and $\mathbf{Q} = \mathbf{X\Lambda}^{\frac{1}{2}}\mathbf{X'}$. You can check by direct multiplication that indeed $\mathbf{PP'} = \mathbf{A}$ and $\mathbf{QQ} = \mathbf{A}$. ♣

Section End Questions

1. Give an example of a matrix not all of the eigenvalues of which are real.
2. Create a 3×3 example of a symmetric matrix. Calculate all its eigenvalues and eigenvectors. Now verify by computation if the claim made by the spectral decomposition Theorem holds.
3. Verify Facts 10 and 11.
4. Let us call a real matrix diagonalizable if there exists an orthogonal matrix $\mathbf{U}$ and a diagonal matrix $\mathbf{D}$ such that $\mathbf{A} = \mathbf{U'DU}$. The Spectral Decomposition Theorem says that symmetric matrices are diagonalizable. Actually, a more general class of matrix is diagonalizable. There is a theorem that says that a real matrix $\mathbf{A}$ is diagonalizable if and only if $\mathbf{AA'} = \mathbf{A'A}$ (such matrices are called 'normal' matrices). Using this result, give an example of a matrix that is not diagonalizable. Now give a direct proof as to why your matrix cannot be diagonalized.
5. Create a matrix which is neither positive semidefinite (psd) nor negative semidefinite (nsd). Why is it neither psd or nsd?