A Primer on Linear Algebra and a Review of Econometrics[1]

Linear Algebra

A *matrix* is simply a collection of individual elements. For example, the matrix A may be:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \tag{1}$$

Here, $a_{1,1}$ is the element from the first row and first column of A., $a_{2,1}$ is the element from the second row and first column of A...

A matrix is said to be MxN if it consists of M rows and N columns. A, for example, is $2x2$.

A *vector* is a matrix consisting of either one row or one column and a *scalar* is a $1x1$ matrix.

Addition

Adding matrices is very simple. We simply sum the matching elements for each matrix. For example:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 6 & 8 \\ 10 & 12 \end{bmatrix} \tag{2}$$

In order for matrices to be added, they must be of the same dimension (same number of rows and columns). If not, they are not conformable and their sum is not defined.

Multiplication

To multiply matrices, we do not simply take the product of the corresponding elements:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} * \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} \neq \begin{bmatrix} 5 & 12 \\ 21 & 32 \end{bmatrix} \tag{3}$$

Considering $A * B = C$, the first row of A multiplies the first column of B to obtain $c_{1,1}$.

[1] These are undergraduate lecture notes. They do not represent academic work. Expect typos, sloppy formatting, and occasional (possibly stupefying) errors.

$$c_{1,1} = a_{1,1} * b_{1,1} + a_{1,2} * b_{2,1} + a_{1,m} * b_{m,1} \tag{4}$$

The second row of A then multiplies the first column of B to obtain $c_{2,1}$.

$$c_{2,1} = a_{2,1} * b_{1,1} + a_{2,2} * b_{2,1} + a_{2,m} * b_{m,1} \tag{5}$$

The first row of A multiplies the second column of B to obtain $c_{1,2}$.

$$c_{1,2} = a_{1,1} * b_{1,2} + a_{1,2} * b_{2,2} + a_{1,m} * b_{m,2} \tag{6}$$

It thus follows that:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} * \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix} \tag{7}$$

For matrix multiplication to work, the number of rows of A must equal the number of columns of B. Otherwise their product is not defined. If A is MxN, and B is NxP, then their product is MxP.

Note that, in general, $A * B \neq B * A$.

Multiplying a matrix by a scalar is different. The following example shows how this is done:

$$\kappa \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} \kappa 1 & \kappa 2 \\ \kappa 3 & \kappa 4 \end{bmatrix} \tag{8}$$

Inversion

The *identity matrix* is the matrix equivalent of the number one. $I(3)$, the $3x3$ identity matrix is:

$$I(3) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{9}$$

For $I(n)$, the identity matrix consist of ones on all n diagonals and zero for all off-diagonal elements. The inverse of A, denoted as A^{-1}, has the property that:

$$A * A^{-1} = I = A^{-1} * A \tag{10}$$

It is also the case that:

$$A * I = A = I * A \tag{11}$$

For large matrices, there is no simple formula for computing an inverse. For a two by two matrix, however, a simple formula does exist:

$$\begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} = \frac{1}{a_{1,1}a_{2,2} - a_{1,2}a_{2,1}} \begin{bmatrix} a_{2,2} & -a_{1,2} \\ -a_{2,1} & a_{1,1} \end{bmatrix} \tag{12}$$

The inverse of a matrix is only defined for square matrices (those with the same number of rows and columns). A square matrix is not always invertible, an inverse exists if and only if the matrix's *determinate* is non-zero. For a $2x2$ matrix, the determinate equals $a_{1,1}a_{2,2} - a_{1,2}a_{2,1}$.

A $2x2$ matrix is non-invertible if $a_{1,1}a_{2,2} = a_{1,2}a_{2,1}$. This occurs if one row (or column) is a multiple of the other row (or column). For example, the following matrix is not invertible:

$$\begin{bmatrix} 1 & 5 \\ -2 & -10 \end{bmatrix}$$

Transpose

The transpose of a matrix is replaces its columns with its rows and vice-versa. For example:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \tag{13}$$

$$A' = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \tag{14}$$

Eigenvalues and Eigenvectors

The *eigenvalues* of square matrices will be useful throughout the term. The eigenvalues (denoted as the vector λ) is obtained through the *characteristic equation*:

$$Det\,(A - \lambda I) = 0 \tag{15}$$

For each eigenvector, λ_i, it is possible to obtain an eigenvector, x using the following formula:

$$(A - \lambda_i I)x = 0 \tag{16}$$

Do not be alarmed if you do not see the usefulness of eigenvalues. It will become clearer when we apply this background to macroeconomic material later in the term.

An Example

Consider the following matrix:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

To obtain this matrix's eigenvalues, we set the determinate of the following matrix equal to zero:

$$\begin{bmatrix} 1-\lambda & 2 \\ 3 & 4-\lambda \end{bmatrix} \tag{17}$$

which using the formula for the determinate of a $2x2$ matrix sets:

$$(1-\lambda)(4-\lambda) - 6 = 0; \tag{18}$$

The solution to (18) is a quadratic it may be rewritten as:

$$\lambda^2 - 5\lambda - 2 = 0 \tag{19}$$

Using the quadratic formula:

$$\lambda_i = \frac{5 \pm -\sqrt{25+8}}{2} = -0.37, 5.37 \tag{20}$$

To find eigenvectors, consider $\lambda_i = 5.37$:

$$\begin{bmatrix} 1-5.37 & 2 \\ 3 & 4-5.37 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0 \tag{21}$$

which yields:

$$\begin{bmatrix} -4.37x_1 + 2x_2 \\ 3x_1 - 1.37x_2 \end{bmatrix} = 0 \tag{22}$$

Both equations reduce to $x_2 = 2.19x_1$. Any non-zero vector that satisfies this relationship is thus an eigenvector. The solution may thus be written as:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2.19 \end{bmatrix} \tag{23}$$

Eigenvectors are not unique. Multiplying an eigenvector by any non-zero scalar (zero vectors are excluded from being eigenvectors) also returns a valid eigenvector.

Application: Representing a system of equations:

Serious macroeconomic models almost always consist of a set of equations. Consider the following example:

$$\pi_t + \alpha y_t + \beta i_t = u_t \tag{24}$$

$$\pi_t + \gamma y_t + \omega i_t = g_t \tag{25}$$

$$\pi_t + \chi y_t + \kappa i_t = e_t \tag{26}$$

Where π_t is inflation y_t is output, i_t is the interest rate, and the variables on the right hand side are exogenous shocks. Solving this model consists of representing the three endogenous variables as functions of exogenous shocks and parameters. Because there are three equations and three endogenous variables, the system is well-specified. It is possible through tedious substitution, to do this using (24)-(26). Our life is much easier, however if we represent the model using matrices:

$$Ax_t = z_t \tag{27}$$

$$\begin{bmatrix} 1 & \alpha & \beta \\ 1 & \gamma & \omega \\ 1 & \chi & \tau \end{bmatrix} \begin{bmatrix} \pi_t \\ y_t \\ ei_t \end{bmatrix} = \begin{bmatrix} u_t \\ g_t \\ e_t \end{bmatrix} \tag{28}$$

Notice that by multiplying the matrices on the left hand side we return (24)-(26). Using linear algebra we can work with a single equation instead of three. Solving the model is easy:

$$x_t = A^{-1} z_t \tag{29}$$

A Review of Linear Regression with Matrices

Denote y as a vector consisting of the dependent variable. For now, it does not matter if this vector consists of a time series or a cross sectional variable. Denote x as a matrix of independent variables. If we wish to include an intercept (which is typical), then the first column of x will consist of a column of ones.

We define our regression as:

$$\hat{y}_t = x_t^{'} b \tag{30}$$

Here, I have chosen to use the t subscript to indicate a time series, But I could have included an i subscript instead, it doesn't matter for now. $\hat{y}_t$ is the fitted, not actual value, of y_t. b is our vector of regression coefficients.

The regression will not fit the data exactly. We thus define e_t as the vector of residuals:

$$e_t = y_t - x_t^{'} b \tag{31}$$

The next step is to choose how we select b. We will consider the ordinary least squares (OLS) estimator that minimizes the sum of squared errors. Keep in mind, that this choice is somewhat arbitrary. We could instead choose the estimator that minimizes the absolute value of the error terms, but this is mathematically complex. We could also choose the estimator that minimizes the sum of the error terms raised to the power 4.

Note that the sum of squares errors may be written as $e_t^{'}e_t$. We thus set up the following minimization problem:

$$Min_b \;\; e_t^{'}e_t = (y - xb)^{'}(y - xb) \tag{32}$$

Solving for the OLS estimator is then a matter of differentiating with respect to b and setting this equal to zero. The following process involves some subtle matrix manipulation that will probably confuse you if this is the first time that you have worked with matrices. But I still think there is value in seeing the simplicity of this derivation. Expanding (32) yields:

$$Min_b \;\; e_t^{'}e_t = y^{'}y - b^{'}xy - y^{'}xb + b^{'}x^{'}xb \tag{33}$$

Re-arranging yields:

$$Min_b \;\; e_t^{'}e_t = y^{'}y - 2y^{'}xb + b^{'}x^{'}xb \tag{34}$$

Differentiating with respect to b then yields the first-order condition:[2]

$$-2x^{'}y + 2x^{'}xb = 0 \tag{35}$$

Which then yields:

$$b = (x^{'}x)^{-1}x^{'}y \tag{36}$$

Equation (36) is the OLS estimator. Note that this derivation does not depend on the number of independent variables or on whether or not x includes a constant.It can be verified that this is just the matrix representation of the same OLS estimator that your derived in ECO 255.

Equation (36) gives us an estimate of the true regression parameters. We are also typically interested in the accuracy of this estimate. This is especially important for hypothesis testing. The OLS estimator's covariance matrix is:

$$Var[b|X] = \sigma^2(x^{'}x)^{-1} \tag{37}$$

where σ^2 is an estimate of the true error term (which we cannot observe). It can be shown that an unbiased estimator of σ^2 is:

[2] Just trust me that second order conditions hold.

$$s^2 = \frac{e^{'}e}{T-k} \tag{38}$$

Here T is the sample size and k is the number of independent variables. $T-k$ is known as the degrees of freedom. We obtain our sample variance by taking:

$$Var[b|x] = s^2(x^{'}x)^{-1} \tag{39}$$

It is then straightforward to do hypothesis testing. Suppose, for example, that we wish to test whether or not the third element of b, denoted b_3 is statistically different from zero. We may then write our null hypothesis as $H_0 : b_3 = 0$. We first take the square root of the third element of b to obtain this variable's standard error (recall that the standard deviation is the square root of the variance). We then divide b_3 by this standard error to obtain a t-statistic. If this statistic is greater than 1.96, we can then, for example, claim that the variable is statistically significant at the 95% confidence level. We thus reject the null hypothesis. If our t-statistic is less than the critical value then we fail to reject the null hypothesis. We never accept a hypothesis.

Estimating OLS is easy. And as long as the Gauss-Markov conditions hold, then OLS will be BLUE (Best Linear Unbiased Estimator). In this case, OLS is probably all we need. Recall these conditions:

1. No Omitted Variable Bias. In other words, OLS is only unbiased if we include all of the relevant variables on the right hand side. Unfortunately, all econometric models are misspecified in this dimension, we always omit something that matters. The goal is thus to use economic intuition and theory to convincingly argue that we have minimized omitted variable bias. We can't just mindlessly run regressions.

2. Exogeneity. Formally, $E[\sigma|x] = 0$. This requires that the error terms be uncorrelated with any of the independent variables. A common way in which exogeneity is violated is simultaneity. Suppose, for example, that output is our dependent variable and interest rates are our independent variable. Macroeconomic theory suggests that these variables are determined at the same time. Suppose, for example, that random recessions cause Central Banks to lower interest rates. This shows up in our specification as both a negative error term and a lower interest rate. In words, interest rates and the error term are correlated. This endogeneity biases the results.

3. No perfect colinearity. This requires that no independent variable be a perfect function of other independent variables. Suppose, for example, that x has two variables and that the second is simply twice the first. Our matrix may thus look like:

$$x = \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \end{bmatrix} \tag{40}$$

Now note that:

$$x^{'}x = \begin{bmatrix} 14 & 28 \\ 28 & 56 \end{bmatrix} \tag{41}$$

When we then try to obtain $(x^{'}x)^{-1}$, crucial for obtaining the OLS estimator, we get the following:

Perfect multicolinearity causes this critical matrix to be non-invertible. Intuitively, because the two independent variables are perfectly correlated, OLS cannot tell how they individually affect the dependent variable.

Now suppose that we have near perfect colinearity. This may yield:

$$x^{'}x = \begin{bmatrix} 14 & 28 \\ 27.99 & 56 \end{bmatrix} \tag{42}$$

We can now invert $x^{'}x$ but examining (12) (where we see the determinate is close to zero) shows that each element of the inverted matrix becomes very large. Using (39), our standard errors also become very large and it is nearly impossible to find statistical significance.

There are two solutions for dealing with too much multicolinearity. The first is to get more data. Using (39), this will drive down the resulting standard errors. But telling a researcher to get more data is a little like telling a student to just get smarter or telling Mark Sanchez to stop being such a horrible quarterback. They would already be doing it if it were feasible. The second is to drop one (or more if needed) variables and recognize that we are estimating a joint effect and interpret our results accordingly.

4. Homoscedasticity. Formally $E[e_t^2|x] = 0$. In words, the variance of the error term may not depend on any right hand side variables. Suppose for example, that wages are our dependent variable while occupations are our right hand side variables. If some occupations not only affect the mean value of wages, but also increase volatility (for example professional athletes are sometimes very poor and sometimes fabulously wealthy), then we may have heteroscedasticity.

Heteroscedasticity tends to be a major issue when working with cross sectional data. It does not bias the coefficient estimates but does bias the standard errors upwards. It is easy to deal with by obtaining White 's standard errors. In Stata, instead of running

$$reg\ y\ x; \tag{43}$$

we run

$$reg\ y\ x\ robust; \tag{44}$$

5. Nonautocorrelation. Formally $E[e_t e_s|X] = 0$ for any $t \neq s$. In words, the error term may not be correlated with its lags. Autocorrelation is usually a problem with time series data. This course will develop techniques to fix it.

Spatial autocorrelation may, however, also be a problem with cross sectional data. It may occur, for example, if certain geographical areas are likely to have high (low) error terms for multiple observations.

Some Examples

#1. Suppose that we wish to regress y only on a constant. x is thus simply a vector of ones. $(x'x)^{-1}$ therefore equals the sample size T. $x'y$ is thus simply $\sum_{t=1}^{T} y_t$. The OLS estimator thus equals:

$$b = \frac{\sum_{t=1}^{T} y_t}{T} \tag{45}$$

Note that (45) is simply the sample average.

#2. Now suppose our right hand side consists only one variable and does not include a constant. Now, $(x^{'}x)^{-1} = \frac{1}{\sum_{t=1}^{T} x_t^2}$ and $x^{'}y = \sum_{t=1}^{T} x_t y_t$. Our OLS estimator is thus:

$$b = \frac{\sum_{t=1}^{T} x_t y_t}{\sum_{t=1}^{T} x_t^2} \tag{46}$$

Other Issues

1. Endogeneity (see #1 of the Gauss-Markov conditions) is a widespread problem that threatens to bias many econometric results. The solution to this problem is to find an instrument. A good instrument is correlated with the independent variable that exhibits endogeneity but is uncorrelated with the error term. Denote a set of instruments as z. An unbiased and consistent estimator is then:

$$b_{iv} = [x^{'}z(z^{'}z)^{-1}z^{'}x]^{-1}x^{'}z(z^{'}z)^{-1}z^{'}y \tag{47}$$

Although messy, it is straightforward to calculate this estimator as well as a related vector of standard errors. The challenge is thus to find an appropriate instrument for the problem at hand.

In much cross sectional work, there is no obvious instrument. A major part of the project is thus to find, justify, and defend the choice of an instrument. The researcher must be prepared to argue that the instrument is genuinely uncorrelated with the error term and the acceptance, of lack thereof, of this argument may determine the project's success or failure. The popular book *Freakonomics* is essentially a set of examples where Steve Levitt, the author, finds clever instruments that allow him to answer certain questions.

In time series analysis, however, natural instruments are much easier to find. Suppose that x_t is correlated with the error term. In most cases, its lag, x_{t-1} will be correlated with x_t but uncorrelated with e_t. It is thus typically used as an instrument.

2. We often consider measures of fit. The most common is R^2. This tells us the fraction of the dependent variable's variation that is explained by the independent variables as opposed to the error terms. Because OLS maximizes OLS for a set of independent variables, and because it can always select a regression coefficient of 0, adding an additional independent variable can only increase R^2. As a result, R^2 is an interesting metric, but we should take care to make sure

that we are not trying to maximize it. Over-fitting a regression model is a potentially serious error.

To see the dangers of over-fitting, suppose that we keep adding explanatory variables, formally k increases, so that our degrees of freedom become very small. From (38) and (39), we see that as $T - k$ becomes small, our standard errors increase as well. As a result, we have less confidence in our estimates and we would expect our estimate to do poorly when taken to new data. Thus well we do a good job of fitting the sample, the out of sample fit is poor.

To determine the right value of k, there are numerous *informational criteria.* these all work in a similar way. They reward the specification for good fit. At the same time, they penalize the specification for including an additional variable. If the additional fit is larger than the penalty, then we should add another variable. the Unfortunately, there is no clear theoretical basis for determining this penalty. As a result, there are several popular informational criteria including adjusted R^2, the Akaike Information Criteria (AIC), and the Schwartz Information Criteria (SIC).

Later in the term, it will be important to determine which, for example, AR process best fits a time series., Adding lags will always improve fit. We will use information criteria to determine the optimal number of lags.