OLS in Matrix Form

Nathaniel Beck
Department of Political Science
University of California, San Diego
La Jolla, CA 92093
beck@ucsd.edu
http://weber.ucsd.edu/∼nbeck

April, 2001

Some useful matrices

If $\mathbf{X}$ is a matrix, its *transpose*, $\mathbf{X}'$ is the matrix with rows and columns flipped so the ijth element of $\mathbf{X}$ becomes the jith element of $\mathbf{X}'$.

Matrix forms to recognize:

For vector x, $x'x =$ sum of squares of the elements of x (scalar)

For vector x, $xx' = N \times N$ matrix with ijth element $x_i x_j$

A square matrix is *symmetric* if it can be flipped around its main diagonal, that is, $x_{ij} = x_{ji}$. In other words, if $\mathbf{X}$ is symmetric, $\mathbf{X} = \mathbf{X}'$. xx' is symmetric.

For a rectangular $m \times N$ matrix $\mathbf{X}$, $\mathbf{X}'\mathbf{X}$ is the $N \times N$ square matrix where a typical element is the sum of the cross products of the elements of row i and column j; the diagonal is the sum of the squares of row i.

OLS

Let $\mathbf{X}$ be an $N \times k$ matrix where we have observations on K variables for N units. (Since the model will usually contain a constant term, one of the columns has all ones. This column is no different than any other, and so henceforth we can ignore constant terms.) Let $\mathbf{y}$ be an n-vector of observations on the dependent variable. IF $\boldsymbol{\epsilon}$ is the vector of errors and β is the K-vector of unknown parameters:

We can write the general linear model as

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}. \tag{1}$$

The vector of residuals is given by

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\beta} \tag{2}$$

where the hat over β indicates the OLS estimate of β.

We can find this estimate by minimizing the sum of

squared residuals. Note this sum is $\mathbf{e}'\mathbf{e}$. Make sure you can see that this is very different than $\mathbf{e}\mathbf{e}'$.

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) \tag{3}$$

which is quite easy to minimize using standard calculus (on matrices quadratic forms and then using chain rule).

This yields the famous normal equations

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \tag{4}$$

or, if $\mathbf{X}'\mathbf{X}$ is non-singular,

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{5}$$

Under what conditions will $\mathbf{X}'\mathbf{X}$ be non-singular (of full rank)?

$\mathbf{X}'\mathbf{X}$ is $K \times K$.

One necessary condition, based on a trivial theorem on rank, is that $N \geq K$. This assumptions is usually met trivially, N is usually big, K is usually small.

Next must have all of the columns of $\mathbf{X}$ be linearly independent (this is why we did all this work), that is no variable is a linear combination of the other variables.

This is the assumption of no (perfect) multicolinearity.

Note that only linear combinations are ruled out, NOT non-linear combinations.

Gauss-Markov assumptions

The critical assumption is that we get the mean function right, that is $E(\mathbf{y}) = \mathbf{X}\beta$.

The second critical assumption is either that $\mathbf{X}$ is non-stochastic, or, if it is, that it is independent of $\mathbf{e}$.

We can very compactly write the Gauss-Markov (OLS) assumptions on the errors as

$$\mathbf{\Omega} = \sigma^2 I \tag{6}$$

where $\mathbf{\Omega}$ is the variance covariance matrix of the error process,

$$\mathbf{\Omega} = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'). \tag{7}$$

Make sure you can unpack this into

- Homoskedasticity
- Uncorrelated errors

VCV Matrix of the OLS estimates

We can derive the variance covariance matrix of the OLS estimator, $\hat{\beta}$.

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{8}$$
$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \boldsymbol{\epsilon}) \tag{9}$$
$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \tag{10}$$
$$= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}. \tag{11}$$

This shows immediately that OLS is unbiased so long as either X is non-stochastic so that

$$E(\hat{\beta}) = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\epsilon}) = \beta \tag{12}$$

or still unbiased if X is stochastic but independent of ϵ, so that $E(\mathbf{X}\epsilon) = 0$.

The variance covariance matrix of the OLS estimator

is then

$$E((\hat{\beta}-\beta)(\hat{\beta}-\beta)') = E\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}]'\right) \tag{13}$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \tag{14}$$

and then given our assumption about the variance covariance of the errors, Equation 6

$$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{15}$$

Robust (Huber or White) standard errors

Note how the second to last formulation makes sense of both White's heteroskedasticity consistent standard errors and my panel consistent standard errors.

Heteroskedasticity will lead to incorrect standard errors insofar as

$$\mathbf{X}'E\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right)\mathbf{X} \neq \sigma^2(\mathbf{X}'\mathbf{X}) \tag{16}$$

We don't know the ϵ but we do know the residuals, $\mathbf{e}$. Obviously the each individual residual is not a good estimator of the corresponding ϵ, but White showed that $\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}$ is a good estimator of the corresponding expectation term.

Thus White suggested a test for seeing how far this estimator diverges from what you would get if you just used the OLS standard errors. This test is to regress the squared residuals on the terms in $\mathbf{X}'\mathbf{X}$, that is the squares and cross-products of the independent variables. If the R^2 exceeds a critical

value (NR^2 is χ^2_k), then heteroskedasticity causes problems. At that point use the White estimate. (By and large always using the White estimate can do little harm and some good.)

Partitioned matrix and partial regression - the FWL theorem

In all the below, any matrix or submatrix that is inverted is square, but other matrices may be rectangular so long as everything is conformable and only square matrices ended up being inverted.

Direct multiplication tell us that

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22}^{-1} \end{bmatrix} \tag{17}$$

Matrices like the above are called block diagonal. As we shall see, this tells us a lot about when we can ignore that we might have added other variables to a regression.

The situation is more complicated for a general

matrix.

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1}(I + \mathbf{A}_{12}\mathbf{F}\mathbf{A}_{21}A_{11}^{-1}) & -A_{11}^{-1}A_{12}\mathbf{F} \\ -\mathbf{F}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{F} \end{bmatrix} \tag{18}$$

where

$$\mathbf{F} = (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \tag{19}$$

There are a lot of ways to show this, but can be done (tediously) by direct multiplication. NOTHING DEEP HERE.

Why do we care? The above formulae allow us to understand what it means to add variables to a regression, and when it matters if we either have too many or too few (omitted variable bias) variables.

First note that if we have two sets of independent variables, say $\mathbf{X}_1$ and $\mathbf{X}_2$ that are *orthogonal* to each other, then the sums of cross products of the variables in $\mathbf{X}_1$ with $\mathbf{X}_2$ are zero (by definition). Thus the $\mathbf{X}'\mathbf{X}$ matrix formed out of $\mathbf{X}_1$ and $\mathbf{X}_2$ is block diagonal and so the theorem on the inverse of block

diagonal matrices tells us that the OLS estimates of the coefficients of the first set of variables estimated separately is the same as what we would get if we estimated using both sets of variables.

What does it mean for the two sets of variables to be orthogonal. Essentially, it means they are independent, that is, one has nothing to do with the other. So if we have regressions involving political variables, and we think that hair color is unrelated to any of these, then we can not worry about what would happen if we included hair color in the regression. But if we leave out race or party id, it will make a difference.

The more interesting question is what happens if the two sets of variables are not orthogonal; in particular, what happens if we estimate a regression using a set of variables $\mathbf{X}_1$ but omit relevant $\mathbf{X}_2$. That is, suppose the "true" model is

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon \tag{20}$$

but we "mistakenly" omit the $\mathbf{X}_2$ variables from the regression.

The true normal equation is:

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1\mathbf{y} \\ \mathbf{X}_2\mathbf{y} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \qquad (21)$$

Now we can use the results on partitioned inverse to see that

$$\hat{\beta}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\hat{\beta}_2 \qquad (22)$$

Note that the first term in this (up to the minus sign) is just the OLS estimates of the $\hat{\beta}_1$ in the regression of $\mathbf{y}$ on the $\mathbf{X}_1$ variables alone.

Thus it is only irrelevant to ignore "omitted" variables if the second term, after the minus sign, is zero.

What is that term.

The first part of that term, up the $\hat{\beta}_2$ is just the regression of the variables in $\mathbf{X}_2$ (done separately and then put together into a matrix) on all the variables in $\mathbf{X}_1$.

This will only be zero if the variables in $\mathbf{X}_1$ are linearly unrelated to the variables in $\mathbf{X}_2$ (political variables and hair coloring).

The second term will also be zero if $\hat{\beta}_2 = 0$, that is, the $\mathbf{X}_2$ variables have no impact on $\mathbf{y}$.

Thus you can ignore all potential omitted variables that are *either* unrelated to the variables you do include *or* unrelated to the dependent variables.

But any other variables that do not meet this condition will change your estimates of $\hat{\beta}_1$ if you do include them.

To study this further, we need some more matrices!

The residual maker and the hat matrix

There are some useful matrices that pop up a lot.

Note that

$$
\begin{aligned}
e &= \mathbf{y} - \mathbf{X}\hat{\beta} && (23)\\
&= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} && (24)\\
&= (I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} && (25)\\
&= My && (26)
\end{aligned}
$$

where $M =$ and M *Makes* residuals out of $\mathbf{y}$. Note that M is $N \times N$, that is, big!

A square matrix A is *idempotent* if $A^2 = AA = A$ (in scalars, only 0 and 1 would be idempotent). M is

idempotent.

$$MM = (I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \tag{27}$$

$$= I^2 - 2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{28}$$

$$= I - 2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{29}$$

$$= M \tag{30}$$

This will prove useful

A related matrix is the *hat* matrix which makes $\hat{\mathbf{y}}$, the predicted $\mathbf{y}$ out of $\mathbf{y}$. Just note that

$$\hat{\mathbf{y}} = \mathbf{y} - \mathbf{e} = [I - M]y = Hy \tag{31}$$

where

$$H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{32}$$

Greene calls this matrix P, but he is alone. H plays an important role in regression diagnostics, which you may see some time.

Back to comparing big and small regressions

If we "uninvert" the normal equation (Equation 21) we get

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1\mathbf{y} \\ \mathbf{X}_2\mathbf{y} \end{bmatrix} \quad (33)$$

and we can simplify the equation for $\hat{\beta}_1$ when all variables are included (Equation 22 to

$$\hat{\beta}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\hat{\beta}_2) \quad (34)$$

Direct multiplication for the second element in Equation 33 gives

$$\mathbf{X}_2'\mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2'\mathbf{X}_2\hat{\beta}_2 = \mathbf{X}_2'\mathbf{y} \quad (35)$$

and then substituting for $\hat{\beta}_1$ using Equation 34 gives

$$\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - \mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\hat{\beta}_2 + \mathbf{X}_2'\mathbf{X}_2\hat{\beta}_2 = \mathbf{X}_2\mathbf{y} \quad (36)$$

which simplifies to

$$\mathbf{X}_2'[I - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']\mathbf{X}_2\hat{\beta}_2 = \mathbf{X}_2'[I - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']\mathbf{y} \quad (37)$$

and then we get $\hat{\beta}_2$ by premultiplying both sides by the inverse of the term that premultiplies $\hat{\beta}_2$.

Note that the term in brackets is the M matrix which makes residuals for regressions on the $\mathbf{X}_1$ variables; $M\mathbf{y}$ is the vector of residuals from regressing $\mathbf{y}$ on the $\mathbf{X}_1$ variables and $M\mathbf{X}_2$ is the matrix made up of the column by column residuals of regressing each variable (column) in $\mathbf{X}_2$ on all the variables in $\mathbf{X}_1$.

Because M is both idempotent and symmetric, we can then write

$$\hat{\beta}_2 = (\mathbf{X}_2^*\mathbf{X}_2)^{-1}\mathbf{X}_2^*\mathbf{y}^* \quad (38)$$

where $\mathbf{X}_2^* = M_1\mathbf{X}_2$ and $\mathbf{y}^* = M_1\mathbf{y}$.

Note Equation 38 shows that $\hat{\beta}_2$ is just obtained from regressing $\mathbf{y}^*$ on $\mathbf{X}_2^*$ (get good at spotting regressions, that is, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ forms).

But what are the starred variables. They are just the residuals of the variables after regressing on the $\mathbf{X}_1$ variables.

So the difference between regressing only on $\mathbf{X}_2$ and both $\mathbf{X}_2$ and $\mathbf{X}_1$ variables is the latter first regresses both the dependent variable and all the $\mathbf{X}_2$ variables on the $\mathbf{X}_1$ variables and then regresses the residuals on each other, while the smaller regression just regresses $\mathbf{y}$ on the $\mathbf{X}_2$ variables.

This is what is means to hold the $\mathbf{X}_1$ variables "constant" in a multiple regression, and explains why we have so many controversies about what variables to include in a multiple regression.