Chapter 15
Regression analysis with linear algebra primer

15.1 Overview

This primer is intended to provide a mathematical bridge to a master’s level course that uses linear algebra for students who have taken an undergraduate econometrics course that does not. Why should we make the mathematical shift? The most immediate reason is the huge double benefit of allowing us to generalise the core results to models with many explanatory variables while simultaneously permitting a great simplification of the mathematics. This alone justifies the investment in time – probably not more than ten hours – required to acquire the necessary understanding of basic linear algebra.

In fact, one could very well put the question the other way. Why do introductory econometrics courses not make this investment and use linear algebra from the start? Why do they (almost) invariably use ordinary algebra, leaving students to make the switch when they take a second course?

The answer to this is that the overriding objective of an introductory econometrics course must be to encourage the development of a solid intuitive understanding of the material and it is easier to do this with familiar, everyday algebra than with linear algebra, which for many students initially seems alien and abstract. An introductory course should ensure that at all times students understand the purpose and value of what they are doing. This is far more important than proofs and for this purpose it is usually sufficient to consider models with one, or at most two, explanatory variables. Even in the relatively advanced material, where we are forced to consider asymptotics because we cannot obtain finite-sample results, the lower-level mathematics holds its own. This is especially obvious when we come to consider finite-sample properties of estimators when only asymptotic results are available mathematically. We invariably use a simple model for a simulation, not one that requires a knowledge of linear algebra.

These comments apply even when it comes to proofs. It is usually helpful to see a proof in miniature where one can easily see exactly what is involved. It is then usually sufficient to know that in principle it generalises, without there being any great urgency to see a general proof. Of course, the linear algebra version of the proof will be general and often simpler, but it will be less intuitively accessible and so it is useful to have seen a miniature proof first. Proofs of the unbiasedness of the regression coefficients under appropriate assumptions are obvious examples.

At all costs, one wishes to avoid the study of econometrics becoming an extended exercise in abstract mathematics, most of which practitioners will never use again. They will use regression applications and as long as they understand what is happening in
15. Regression analysis with linear algebra primer

principle, the actual mechanics are of little interest.

This primer is not intended as an exposition of linear algebra as such. It assumes that a basic knowledge of linear algebra, for which there are many excellent introductory textbooks, has already been acquired. For the most part, it is sufficient that you should know the rules for multiplying two matrices together and for deriving the inverse of a square matrix, and that you should understand the consequences of a square matrix having a zero determinant.

15.2 Notation

Matrices and vectors will be written bold, upright, matrices upper case, for example $A$, and vectors lower case, for example $b$. The transpose of a matrix will be denoted by a prime, so that the transpose of $A$ is $A'$, and the inverse of a matrix will be denoted by a superscript $^{-1}$, so that the inverse of $A$ is $A^{-1}$.

15.3 Test exercises

Answers to all of the exercises in this primer will be found at its end. If you are unable to answer the following exercises, you need to spend more time learning basic matrix algebra before reading this primer. The rules in Exercises 3–5 will be used frequently without further explanation.

1. Demonstrate that the inverse of the inverse of a matrix is the original matrix.

2. Demonstrate that if a (square) matrix possesses an inverse, the inverse is unique.

3. Demonstrate that, if $A = BC$, $A' = C'B'$.

4. Demonstrate that, if $A = BC$, $A^{-1} = C^{-1}B^{-1}$, provided that $B^{-1}$ and $C^{-1}$ exist.

5. Demonstrate that $[A']^{-1} = [A^{-1}]'$.

15.4 The multiple regression model

The most obvious benefit from switching to linear algebra is convenience. It permits an elegant simplification and generalisation of much of the mathematical analysis associated with regression analysis. We will consider the general multiple regression model:

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + u_i$$

where the second subscript identifies the variable and the first the observation. In the textbook, as far as the fourth edition, the subscripts were in the opposite order. The reason for the change of notation here, which will be adopted in the next edition of the textbook, is that it is more compatible with a linear algebra treatment.
15.5. The intercept in a regression model

Equation (1) is a row relating to observation \( i \) in a sample of \( n \) observations. The entire layout would be:

\[
\begin{bmatrix}
Y_1 \\
\vdots \\
Y_i \\
\vdots \\
Y_n
\end{bmatrix} = \begin{bmatrix}
\beta_1 X_{11} + \cdots + \beta_j X_{1j} + \cdots + \beta_k X_{1k} \\
\vdots \\
\beta_1 X_{i1} + \cdots + \beta_j X_{ij} + \cdots + \beta_k X_{ik} \\
\vdots \\
\beta_1 X_{n1} + \cdots + \beta_j X_{nj} + \cdots + \beta_k X_{nk}
\end{bmatrix} + \begin{bmatrix}
u_1 \\
\vdots \\
u_i \\
\vdots \\
u_n
\end{bmatrix}.
\]

This, of course, may be written in linear algebra form as:

\[
y = X\beta + u \tag{2}
\]

where:

\[
y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} X_{11} \cdots X_{1j} \cdots X_{1k} \\ \vdots \\ X_{i1} \cdots X_{ij} \cdots X_{ik} \\ \vdots \\ X_{n1} \cdots X_{nj} \cdots X_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_i \\ \vdots \\ \beta_n \end{bmatrix}, \quad \text{and } u = \begin{bmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix}
\]

with the first subscript of \( X_{ij} \) relating to the row and the second to the column, as is conventional with matrix notation. This was the reason for the change in the order of the subscripts in equation (1).

Frequently, it is convenient to think of the matrix \( X \) as consisting of a set of column vectors:

\[
X = [x_1 \cdots x_j \cdots x_k]
\]

where:

\[
x_j = \begin{bmatrix} X_{1j} \\ \vdots \\ X_{nj} \end{bmatrix}.
\]

\( x_j \) is the set of observations relating to explanatory variable \( j \). It is written lower case, bold, not italic because it is a vector.

15.5 The intercept in a regression model

As described above, there is no special intercept term in the model. If, as is usually the case, one is needed, it is accommodated within the matrix framework by including an \( X \) variable, typically placed as the first, with value equal to 1 in all observations:

\[
x_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.
\]
15. Regression analysis with linear algebra primer

The coefficient of this unit vector is the intercept in the regression model. If it is included, and located as the first column, the $X$ matrix becomes:

$$
X = \begin{bmatrix}
1 & X_{12} & \cdots & X_{1j} & \cdots & X_{1k} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
1 & X_{n2} & \cdots & X_{nj} & \cdots & X_{nk} 
\end{bmatrix}
= [1 \ x_2 \ \cdots \ x_j \ \cdots \ x_k].
$$

15.6 The OLS regression coefficients

Using the matrix and vector notation, we may write the fitted equation:

$$
\hat{Y}_i = \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_k X_{ik}
$$

as:

$$
\hat{y} = X\hat{\beta}
$$

with obvious definitions of $\hat{y}$ and $\hat{\beta}$. Then we may define the vector of residuals as:

$$
\hat{u} = y - \hat{y} = y - X\hat{\beta}
$$

and the residual sum of squares as:

$$
RSS = \hat{u}'\hat{u} = (y - X\hat{\beta})'(y - X\hat{\beta})
$$

$$
= y'y - y'X\hat{\beta} - \hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}
$$

(y'$X\hat{\beta} = \hat{\beta}'X'y$ since it is a scalar.) The next step is to obtain the normal equations:

$$
\frac{\partial RSS}{\partial \hat{\beta}_j} = 0
$$

for $j = 1, \ldots, k$ and solve them (if we can) to obtain the least squares coefficients. Using linear algebra, the normal equations can be written:

$$
X'X\hat{\beta} - X'y = 0.
$$

The derivation is straightforward but tedious and has been consigned to Appendix A. $X'X$ is a square matrix with $k$ rows and columns. If assumption A.2 is satisfied (that it is not possible to write one $X$ variable as a linear combination of the others), $X'X$ has an inverse and we obtain the OLS estimator of the coefficients:

$$
\hat{\beta} = (X'X)^{-1}X'y.
$$

Exercises

6. If $Y = \beta_1 + \beta_2 X + u$, obtain the OLS estimators of $\beta_1$ and $\beta_2$ using (3).

7. If $Y = \beta_2 + u$, obtain the OLS estimator of $\beta_2$ using (3).

8. If $Y = \beta_1 + u$, obtain the OLS estimator of $\beta_1$ using (3).
15.7 Unbiasedness of the OLS regression coefficients

Substituting for $y$ from (2) into (3), we have:

$$\hat{\beta} = [X'X]^{-1}X'(X\beta + u)$$
$$= [X'X]^{-1}X'\beta + [X'X]^{-1}X'u$$
$$= \beta + [X'X]^{-1}X'u.$$ 

Hence each element of $\hat{\beta}$ is equal to the corresponding value of $\beta$ plus a linear combination of the values of the disturbance term in the sample. Next:

$$E(\hat{\beta} \mid X) = \beta + E([X'X]^{-1}X'u \mid X).$$

To proceed further, we need to be specific about the data generation process (DGP) for $X$ and the assumptions concerning $u$ and $X$. In Model A, we have no DGP for $X$: the data are simply taken as given. When we describe the properties of the regression estimators, we are either talking about the potential properties, before the sample has been drawn, or about the distributions that we would expect in repeated samples using those given data on $X$. If we make the assumption $E(u \mid X) = 0$, then:

$$E(\hat{\beta} \mid X) = \beta + E([X'X]^{-1}X'u \mid X) = \beta$$

and so $\hat{\beta}$ is an unbiased estimator of $\beta$. It should be stressed that unbiasedness in Model A, along with all other properties of the regression estimators, are conditional on the actual given data for $X$.

In Model B, we allow $X$ to be drawn from a fixed joint distribution of the explanatory variables. The appropriate assumption for the disturbance term is that it is distributed independently of $X$ and hence its conditional distribution is no different from its absolute distribution: $E(u \mid X) = E(u)$ for all $X$. We also assume $E(u) = 0$. The independence of the distributions of $X$ and $u$ allows us to write:

$$E(\hat{\beta} \mid X) = \beta + E([X'X]^{-1}X'u \mid X)$$
$$= \beta + E([X'X]^{-1}X')E(u)$$
$$= \beta.$$

15.8 The variance-covariance matrix of the OLS regression coefficients

We define the variance-covariance matrix of the disturbance term to be the matrix whose element in row $i$ and column $j$ is the population covariance of $u_i$ and $u_j$. By assumption A.4, the covariance of $u_i$ and $u_j$ is constant and equal to $\sigma_u^2$ if $j = i$ and by assumption A.5 it is equal to zero if $j \neq i$. Thus the variance-covariance matrix is:
15. Regression analysis with linear algebra primer

\[
\begin{bmatrix}
\sigma_u^2 & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & \sigma_u^2 & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & \sigma_u^2 & \cdots & 0 & 0 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & \cdots & \sigma_u^2 & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & \sigma_u^2 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & \sigma_u^2 \\
\end{bmatrix}
\]

that is, a matrix whose diagonal elements are all equal to \(\sigma_u^2\) and whose off-diagonal elements are all zero. It may more conveniently be written \(I_n\sigma_u^2\) where \(I_n\) is the identity matrix of order \(n\).

Similarly, we define the variance-covariance matrix of the regression coefficients to be the matrix whose element in row \(i\) and column \(j\) is the population covariance of \(\hat{\beta}_i\) and \(\hat{\beta}_j\):

\[
\text{cov}(\hat{\beta}_i, \hat{\beta}_j) = E[(\hat{\beta}_i - E(\hat{\beta}_i))(\hat{\beta}_j - E(\hat{\beta}_j))] = E[(\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j)].
\]

The diagonal elements are of course the variances of the individual regression coefficients. We denote this matrix \(\text{var}(\hat{\beta})\). If we are using the framework of Model A, everything will be conditional on the actual given data for \(X\), so we should refer to \(\text{var}(\hat{\beta} | X)\) rather than \(\text{var}(\hat{\beta})\). Then:

\[
\text{var}(\hat{\beta} | X) = E((\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))' | X) = E((\hat{\beta} - \beta)(\hat{\beta} - \beta)' | X) = E\left( ([X'X]^{-1}X'u)([X'X]^{-1}X'u)' | X \right) = E\left( [X'X]^{-1}X'uX[X'X]^{-1} | X \right) = [X'X]^{-1}X'E(uu'[X]X[X'X]^{-1}) = [X'X]^{-1}X'I_n\sigma_u^2X[X'X]^{-1} = [X'X]^{-1}\sigma_u^2.
\]

If we are using Model B, we can obtain the unconditional variance of \(b\) using the standard decomposition of a variance in a joint distribution:

\[
\text{var}(\hat{\beta}) = E\left[ \text{var}(\hat{\beta} | X) \right] + \text{var}\left[ E(\hat{\beta} | X) \right].
\]

Now \(E(\hat{\beta} | X) = \beta\) for all \(X\), so \(\text{var}[E(\hat{\beta} | X)] = \text{var}(\beta) = 0\) since \(\beta\) is a constant vector, so:

\[
\text{var}(\hat{\beta}) = E\left( ([X'X]^{-1}\sigma_u^2) \right) = \sigma_u^2 E\left( [X'X]^{-1} \right)
\]

the expectation being taken over the distribution of \(X\).

To estimate \(\text{var}(\hat{\beta})\), we need to estimate \(\sigma_u^2\). An unbiased estimator is provided by \(\hat{u}'\hat{u} / (n - k)\). For a proof, see Appendix B.
15.9 The Gauss–Markov theorem

We will demonstrate that the OLS estimators are the minimum variance unbiased estimators that are linear in $y$. For simplicity, we will do this within the framework of Model A, with the analysis conditional on the given data for $X$. The analysis generalises straightforwardly to Model B, where the explanatory variables are stochastic but drawn from fixed distributions.

Consider the general estimator in this class:

$$\hat{\beta}^* = Ay$$

where $A$ is a $k \times n$ matrix. Let:

$$C = A - [X'X]^{-1}X'.$$

Then:

$$\hat{\beta}^* = ([X'X]^{-1}X' + C)y$$. 

= ([X'X]^{-1}X' + C)(X\beta + u) 

= $\beta + CX\beta + [X'X]^{-1}X'u + Cu.$

Unbiasedness requires:

$$CX = 0_k$$

where $0_k$ is a $k \times k$ matrix consisting entirely of zeros. Then, with $E(\hat{\beta}^*) = \beta$, the variance-covariance matrix of $\hat{\beta}^*$ is given by:

$$E\left[(\hat{\beta}^* - \beta)(\hat{\beta}^* - \beta)'ight] = E\left([(X'X)^{-1}X' + C) uu' (X'X)^{-1}X' + C)'ight]$$ 

= $([X'X]^{-1}X' + C) I_k \sigma^2 u ([X'X]^{-1}X' + C)'$ 

= $([X'X]^{-1}X' + C) ([X'X]^{-1}X' + C)' \sigma^2 u$ 

= $([X'X]^{-1} + C C)' \sigma^2 u$.

Now diagonal element $i$ of $CC'$ is the inner product of row $i$ of $C$ and column $i$ of $C'$. These are the same, so it is given by:

$$\sum_{s=1}^{k} c^2_{ik}$$

which is positive unless $c_{is} = 0$ for all $s$. Hence minimising the variances of the estimators of all of the elements of $\beta$ requires $C = 0$. This implies that OLS provides the minimum variance unbiased estimator.

15.10 Consistency of the OLS regression coefficients

Since:

$$\hat{\beta} = \beta + [X'X]^{-1}X'u$$
15. Regression analysis with linear algebra primer

the probability limit of $\hat{\beta}$ is given by:

$$\text{plim } \hat{\beta} = \beta + \text{plim } [X'X]^{-1}X'u$$

$$= \beta + \text{plim } \left( \left[ \frac{1}{n} X'X \right]^{-1} \frac{1}{n} X'u \right).$$

Now, if we are working with cross-sectional data with the explanatory variables drawn from fixed (joint) distributions, it can be shown that:

$$\text{plim } \left[ \frac{1}{n} X'X \right]^{-1}$$

has a limiting matrix and that:

$$\text{plim } \frac{1}{n} X'u = 0.$$

Hence we can decompose:

$$\text{plim } \left( \left[ \frac{1}{n} X'X \right]^{-1} \frac{1}{n} X'u \right) = \text{plim } \left[ \frac{1}{n} X'X \right]^{-1} \text{plim } \frac{1}{n} X'u = 0$$

and so $\text{plim } \hat{\beta} = \beta$. Note that this is only an outline of the proof. For a proper proof and a generalisation to less restrictive assumptions, see Greene pp.64–65.

15.11 Frisch–Waugh–Lovell theorem

We will precede the discussion of the Frisch–Waugh–Lovell (FWL) theorem by introducing the residual-maker matrix. We have seen that, when we fit:

$$y = X\beta + u$$

using OLS, the residuals are given by:

$$\hat{u} = y - \hat{y} = y - X\hat{\beta}.$$

Substituting for $\hat{\beta}$, we have:

$$\hat{u} = y - X[X'X]^{-1}X'y$$

$$= [I - X[X'X]^{-1}X']y$$

$$= My$$

where:

$$M = I - X[X'X]^{-1}X'.$$

$M$ is known as the ‘residual-maker’ matrix because it converts the values of $y$ into the residuals of $y$ when regressed on $X$. Note that $M$ is symmetric, because $M' = M$, and idempotent, meaning that $MM = M$.

Now suppose that we divide the $k$ variables comprising $X$ into two subsets, the first $s$ and the last $k - s$. (For the present purposes, it makes no difference whether there is or
15.11. Frisch–Waugh–Lovell theorem

is not an intercept in the model, and if there is one, whether the vector of ones responsible for it is in the first or second subset.) We will partition $X$ as:

$$X = [X_1 \ X_2]$$

where $X + 1$ comprises the first $s$ columns and $X_2$ comprises the last $k - s$, and we will partition $\beta$ similarly, so that the theoretical model may be written:

$$y = [X_1 \ X_2] \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + u.$$

The FWL theorem states that the OLS estimates of the coefficients in $\beta_1$ are the same as those that would be obtained by the following procedure: regress $y$ on the variables in $X_2$ and save the residuals as $\hat{u}_y$. Regress each of the variables in $X_1$ on $X_2$ and save the matrix of residuals as $\hat{u}_{X_1}$. If we regress $\hat{u}_y$ on $\hat{u}_{X_1}$, we will obtain the same estimates of the coefficients of $\beta_1$ as we did in the straightforward multiple regression. (Why we might want to do this is another matter. We will come to this later.) Applying the preceding discussion relating to the residual-maker, we have:

$$\hat{u}_y = M_2 y$$

where:

$$M_2 = I - X_2 [X_2'X_2]^{-1}X_2'$$

and:

$$\hat{u}_{X_1} = M_2 X_1.$$

Let the vector of coefficients obtained when we regress $\hat{u}_y$ on $\hat{u}_{X_1}$ be denoted $\hat{\beta}_1^*$. Then:

$$\hat{\beta}_1^* = [\hat{u}_{X_1}' \hat{u}_{X_1}]^{-1} \hat{u}_{X_1}' \hat{u}_y$$

$$= [X_1'M_2'M_2X_1]'X_1'M_2'M_2y$$

$$= [X_1'M_2X_1]'X_1M_2y.$$

(Remember that $M_2$ is symmetric and idempotent.) Now we will derive an expression for $\hat{\beta}_1$ from the orthodox multiple regression of $y$ on $X$. For this purpose, it is easiest to start with the normal equations:

$$X'X\hat{\beta} - X'y = 0.$$

We partition $\hat{\beta}$ as $[\hat{\beta}_1 \ \hat{\beta}_2]$. $X'$ is $[X_1' \ X_2']$, and we have the following:

$$X'X = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}$$

$$X'X\hat{\beta} = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1'X_1\hat{\beta}_1 + X_1'X_2\hat{\beta}_2 \\ X_2'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 \end{bmatrix}$$

$$X'y = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}.$$
15. Regression analysis with linear algebra primer

Hence, splitting the normal equations into their upper and lower components, we have:

\[ X_1'X_1\hat{\beta}_1 + X_1'X_2\hat{\beta}_2 - X_1'y = 0 \]

and:

\[ X_2'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 - X_2'y = 0. \]

From the second we obtain:

\[ X_2'X_2\hat{\beta}_2 = X_2'y - X_2'X_1\hat{\beta}_1 \]

and so:

\[ \hat{\beta}_2 = [X_2'X_2]^{-1}[X_2'y - X_2'X_1\hat{\beta}_1]. \]

Substituting for \( \hat{\beta}_2 \) in the first normal equation:

\[ X_1'X_1\hat{\beta}_1 + X_1'X_2[X_2'X_2]^{-1}[X_2'y - X_2'X_1\hat{\beta}_1] - X_1'y = 0. \]

Hence:

\[ X_1'X_1\hat{\beta}_1 - X_1'X_2[X_2'X_2]^{-1}X_2'X_1\hat{\beta}_1 = X_1'y - X_1'X_2[X_2'X_2]^{-1}X_2'y \]

and so:

\[ X_1'[I - X_2[X_2'X_2]^{-1}X_2']X_1\hat{\beta}_1 = X_1'[I - X_2[X_2'X_2]^{-1}X_2']y. \]

Hence:

\[ X_1'M_2X_1\hat{\beta}_1 = X_1'M_2y \]

and:

\[ \hat{\beta}_1 = [X_1'M_2X_1]^{-1}X_1'M_2y = \hat{\beta}_1^*. \]

Why should we be interested in this result? The original purpose remains instructive. In early days, econometricians working with time series data, especially macroeconomic data, were concerned to avoid the problem of spurious regressions. If two variables both possessed a time trend, it was very likely that ‘significant’ results would be obtained when one was regressed on the other, even if there were no genuine relationship between them. To avoid this, it became the custom to detrend the variables before using them by regressing each on a time trend and then working with the residuals from these regressions. Frisch and Waugh (1933) pointed out that this was an unnecessarily laborious procedure. The same results would be obtained using the original data, if a time trend was added as an explanatory variable.

Generalising, and this was the contribution of Lovell, we can infer that, in a multiple regression model, the estimator of the coefficient of any one variable is not influenced by any of the other variables, irrespective of whether they are or are not correlated with the variable in question. The result is so general and basic that it should be understood by all students of econometrics. Of course, it fits neatly with the fact that the multiple regression coefficients are unbiased, irrespective of any correlations among the variables.

A second reason for being interested in the result is that it allows one to depict graphically the relationship between the observations on the dependent variable and those on any single explanatory variable, controlling for the influence of all the other explanatory variables. This is described in the textbook in Section 3.2.
15.12. Exact multicollinearity

Exercise

9. Using the FKL theorem, demonstrate that, if a multiple regression model contains an intercept, the same slope coefficients could be obtained by subtracting the means of all of the variables from the data for them and then regressing the model omitting an intercept.

15.12 Exact multicollinearity

We will assume, as is to be expected, that \( k \), the number of explanatory variables (including the unit vector, if there is one), is less than \( n \), the number of observations. If the explanatory variables are independent, the \( X \) matrix will have rank \( k \) and likewise \( X'X \) will have rank \( k \) and will possess an inverse. However, if one or more linear relationships exist among the explanatory variables, the model will be subject to exact multicollinearity. The rank of \( X \), and hence of \( X'X \), will then be less than \( k \) and \( X'X \) will not possess an inverse.

Suppose we write \( X \) as a set of column vectors \( x_j \), each corresponding to the observations on one of the explanatory variables:

\[
X = \begin{bmatrix} x_1 & \cdots & x_j & \cdots & x_k \end{bmatrix}
\]

where:

\[
x_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{nj} \end{bmatrix}.
\]

Then:

\[
X' = \begin{bmatrix} x_1' \\ \vdots \\ x_j' \\ \vdots \\ x_k' \end{bmatrix}
\]

and the normal equations:

\[
X'X\hat{\beta} - X'y = 0
\]

may be written:

\[
\begin{bmatrix} x_1'X\hat{\beta} \\ \vdots \\ x_j'X\hat{\beta} \\ \vdots \\ x_k'X\hat{\beta} \end{bmatrix} - \begin{bmatrix} x_1'y \\ \vdots \\ x_j'y \\ \vdots \\ x_k'y \end{bmatrix} = 0.
\]
15. Regression analysis with linear algebra primer

Now suppose that one of the explanatory variables, say the last, can be written as a linear combination of the others:

\[ x_k = \sum_{i=1}^{k-1} \lambda_i x_i. \]

Then the last of the normal equations is that linear combination of the other \( k - 1 \). Hence it is redundant, and we are left with a set of \( k - 1 \) equations for determining the \( k \) unknown regression coefficients. The problem is not that there is no solution. It is the opposite: there are too many possible solutions, in fact an infinite number. One coefficient could be chosen arbitrarily, and then the normal equations would provide a solution for the other \( k - 1 \). Some regression applications deal with this situation by dropping one of the variables from the regression specification, effectively assigning a value of zero to its coefficient.

Exact multicollinearity is unusual because it mostly occurs as a consequence of a logical error in the specification of the regression model. The classic example is the dummy variable trap. This occurs when a set of dummy variables \( D_j, j = 1, \ldots, s \) are defined for a qualitative characteristic that has \( s \) categories. If all \( s \) dummy variables are included in the specification, in observation \( i \) we will have:

\[ \sum_{j=1}^{s} D_{ij} = 1 \]

since one of the dummy variables must be equal to 1 and the rest are all zero. But this is the (unchanging) value of the unit vector. Hence the sum of the dummy variables is equal to the unit vector. As a consequence, if the unit vector and all of the dummy variables are simultaneously included in the specification, there will be exact multicollinearity. The solution is to drop one of the dummy variables, making it the reference category, or, alternatively, to drop the intercept (and hence unit vector), effectively making the dummy variable coefficient for each category the intercept for that category. As explained in the textbook, it is illogical to wish to include a complete set of dummy variables as well as the intercept, for then no interpretation can be given to the dummy variable coefficients.

15.13 Estimation of a linear combination of regression coefficients

Suppose that one wishes to estimate a linear combination of the regression parameters:

\[ \sum_{j=1}^{k} \lambda_j \beta_j. \]

In matrix notation, we may write this as \( \lambda' \beta \) where:

\[ \lambda = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_j \\ \vdots \\ \lambda_k \end{bmatrix}. \]
The corresponding linear combination of the regression coefficients, $\lambda'\hat{\beta}$, provides an unbiased estimator of $\lambda'\beta$. However, we will often be interested also in its standard error, and this is not quite so straightforward. We obtain it via the variance:

$$\text{var}(\lambda'\hat{\beta}) = E\left[(\lambda'\hat{\beta} - E(\lambda'\hat{\beta}))^2\right].$$

Since $$(\lambda'\hat{\beta} - \lambda'\beta)$$

is a scalar, it is equal to its own transpose, and so $(\lambda'\hat{\beta} - \lambda'\beta)^2$ may be written:

$$\text{var}(\lambda'\hat{\beta}) = E\left[\lambda'(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\lambda\right] = \lambda'E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'.\lambda\right] = \lambda'[X'X]^{-1}\lambda\sigma^2_u.$$

The square root of this expression provides the standard error of $\lambda'\hat{\beta}$ after we have replaced $\sigma^2_u$ by its estimator $\hat{u}'\hat{u}/(n - k)$ in the usual way.

### 15.14 Testing linear restrictions

An obvious application of the foregoing is its use in testing a linear restriction. Suppose that one has a hypothetical restriction:

$$\sum_{j=1}^{k} \lambda_j \beta_j = \lambda_0.$$

We can perform a $t$ test of the restriction using the $t$ statistic:

$$t = \frac{\lambda'\hat{\beta} - \lambda_0}{\text{s.e.}(\lambda'\hat{\beta})}$$

where the standard error is obtained via the variance-covariance matrix as just described. Alternatively, we could reparameterise the regression specification so that one of the coefficients is $\lambda'\beta$. In practice, this is often more convenient since it avoids having to work with the variance-covariance matrix. If there are multiple restrictions that should be tested simultaneously, the appropriate procedure is an $F$ test comparing $RSS$ for the unrestricted and fully restricted models.

### 15.15 Weighted least squares and heteroskedasticity

Suppose that the regression model:

$$y = X\beta + u$$

satisfies the usual regression model assumptions and suppose that we premultiply the elements of the model by the $n$ by $n$ matrix $A$ whose diagonal elements are $A_{ii}$,
15. Regression analysis with linear algebra primer

\[ i = 1, \ldots, n, \text{ and whose off-diagonal elements are all zero:} \]

\[
A = \begin{bmatrix}
A_{11} & \cdots & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
0 & \cdots & A_{ii} & \cdots & 0 \\
\vdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & 0 & \cdots & A_{nn}
\end{bmatrix}.
\]

The model becomes:

\[ Ay = AX\beta + Au. \]

If we fit it using least squares, the point estimates of the coefficients are given by:

\[ \hat{\beta}_{WLS} = (X'AX)^{-1}X'Ay \]

(WLS standing for weighted least squares). This is unbiased but heteroskedastic because the disturbance term in observation \( i \) is \( A_{ii}u_i \) and has variance \( A_{ii}^2\sigma_u^2 \).

Now suppose that the disturbance term in the original model was heteroskedastic, with variance \( \sigma_u^2 \) in observation \( i \). If we define the matrix \( A \) so that the diagonal elements are determined by:

\[ A_{ii} = \frac{1}{\sqrt{\sigma_u^2}} \]

the corresponding variance in the weighted regression will be 1 for all observations and the WLS model will be homoskedastic. The WLS estimator is then:

\[ \hat{\beta}_{WLS} = (X'CX)^{-1}X'Cy \]

where:

\[
C = A'A = \begin{bmatrix}
\frac{1}{\sigma_u^2} & \cdots & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
0 & \cdots & \frac{1}{\sigma_u^2} & \cdots & 0 \\
\vdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & 0 & \cdots & \frac{1}{\sigma_u^2}
\end{bmatrix}.
\]

The variance-covariance matrix of the WLS coefficients, conditional on the data for \( X \), is:

\[
\text{var}(\hat{\beta}_{WLS}) = E\left[(\hat{\beta}_{WLS} - E(\hat{\beta}_{WLS}))(\hat{\beta}_{WLS} - E(\hat{\beta}_{WLS}))'\right] = E\left[(\hat{\beta}_{WLS} - \beta)(\hat{\beta}_{WLS} - \beta)'\right] = E\left[(X'AX)^{-1}X'A'Au(\hat{\beta}_{WLS} - \beta)^2\right] = [X'AX]^{-1}X'Auu'A'AX[X'AX]^{-1} = [X'AX]^{-1}X'AAX[X'AX]^{-1} = [X'CX]^{-1}X'CX[X'CX]^{-1}\sigma_u^2 = [X'CX]^{-1}\sigma_u^2
\]
since \( A \) has been defined so that:

\[
AE(uu')A' = I.
\]

Of course, in practice we seldom know \( \sigma^2_u \), but if it is appropriate to hypothesise that the standard deviation is proportional to some measurable variable \( Z_i \), then the WLS regression will be homoskedastic if we define \( A \) to have diagonal element \( i \) equal to the reciprocal of \( Z_i \).

### 15.16 IV estimators and TSLS

Suppose that we wish to fit the model:

\[
y = X\beta + u
\]

where one or more of the explanatory variables is not distributed independently of the disturbance term. For convenience, we will describe such variables as ‘endogenous’, irrespective of the reason for the violation of the independence requirement. Given a sufficient number of suitable instruments, we may consider using the IV estimator:

\[
\hat{\beta}^{IV} = [W'X]^{-1}W'y \quad (4)
\]

where \( W \) is the matrix of instruments. In general \( W \) will be a mixture of (1) those original explanatory variables that are distributed independently of the disturbance term (these are then described as acting as instruments for themselves), and (2) new variables that are correlated with the endogenous variables but distributed independently of the disturbance term. If we substitute for \( y \):

\[
\hat{\beta}^{IV} = [W'X]^{-1}W'(X\beta + u) = \beta + [W'X]^{-1}W'u.
\]

We cannot obtain a closed-form expression for the expectation of the error term, so instead we take plims:

\[
\text{plim} \hat{\beta}^{IV} = \beta + \text{plim} \left( \frac{1}{n}W'X \right)^{-1} \frac{1}{n}W'u.
\]

Now if we are using cross-sectional data, it is usually reasonable to suppose that:

\[
\text{plim} \left( \frac{1}{n}W'X \right)^{-1} \quad \text{and} \quad \text{plim} \left( \frac{1}{n}W'u \right)
\]

both exist, in which case we can decompose the plim of the error term:

\[
\text{plim} \hat{\beta}^{IV} = \beta + \text{plim} \left( \frac{1}{n}W'X \right)^{-1} \text{plim} \left( \frac{1}{n}W'u \right).
\]

Further, if the matrix of instruments has been correctly chosen, it can be shown that:

\[
\text{plim} \left( \frac{1}{n}W'u \right) = 0
\]
15. Regression analysis with linear algebra primer

and hence the IV estimator is consistent.

It is not possible to derive a closed-form expression for the variance of the IV estimator in finite samples. The best we can do is to invoke a central limit theorem that gives the limiting distribution asymptotically and work backwards from that, as an approximation, for finite samples. A central limit theorem can be used to establish that:

$$\sqrt{n}(\hat{\beta}_{\text{IV}} - \beta) \xrightarrow{d} N\left(0, \left\{ \sigma_u^2 \text{ plim} \left[ \frac{1}{n} W'X \right]^{-1} \text{ plim} \left[ \frac{1}{n} W'W \right] \text{ plim} \left[ \frac{1}{n} X'W \right]^{-1} \right\} \right).$$

From this, we may infer, that as an approximation, for sufficiently large samples:

$$\hat{\beta}_{\text{IV}} \sim N\left(\beta, \left\{ \frac{\sigma_u^2}{n} \text{ plim} \left[ \frac{1}{n} W'X \right]^{-1} \text{ plim} \left[ \frac{1}{n} W'W \right] \text{ plim} \left[ \frac{1}{n} X'W \right]^{-1} \right\} \right). \quad (5)$$

We have implicitly assumed so far that \( W \) has the same dimensions as \( X \) and hence that \( W'X \) is a square \( k \) by \( k \) matrix. However, the model may be overidentified, with the number of columns of \( W \) exceeding \( k \). In that case, the appropriate procedure is two-stage least squares. One regresses each of the variables in \( X \) on \( W \) and saves the fitted values. The matrix of fitted values is then used as the instrument matrix in place of \( W \).

**Exercises**

10. Using (4) and (5), demonstrate that, for the simple regression model:

$$Y_i = \beta_1 = \beta_2 X_i + u_i$$

with \( Z \) acting as an instrument for \( X \) (and the unit vector acting as an instrument for itself):

$$\hat{\beta}_{\text{IV}}^1 = \bar{Y} - \hat{\beta}_{\text{IV}}^2 \bar{X}$$

$$\hat{\beta}_{\text{IV}}^2 = \frac{\sum (Z_i - \bar{Z}) (Y_i - \bar{Y})}{\sum (Z_i - \bar{Z}) (X_i - \bar{X})}$$

and, as an approximation:

$$\text{var}(\hat{\beta}_{\text{IV}}^2) = \frac{\sigma_u^2}{\sum (X_i - \bar{X})^2} \times \frac{1}{r_{XZ}^2}$$

where \( Z \) is the instrument for \( X \) and \( r_{XZ} \) is the correlation between \( X \) and \( Z \).

11. Demonstrate that any variable acting as an instrument for itself is unaffected by the first stage of two-stage least squares.

12. Demonstrate that TSLS is equivalent to IV if the equation is exactly identified.
15.17 Generalised least squares

The final topic in this introductory primer is generalised least squares and its application to autocorrelation (autocorrelated disturbance terms). One of the basic regression model assumptions is that the disturbance terms in the observations in a sample are distributed identically and independently of each other. If this is the case, the variance-covariance matrix of the disturbance terms is the identity matrix of order \( n \), multiplied by \( \sigma^2_u \). We have encountered one type of violation, heteroskedasticity, where the values of the disturbance term are independent but not identical. The consequence was that the off-diagonal elements of the variance-covariance matrix remained zero, but the diagonal elements differed. Mathematically, autocorrelation is complementary. It occurs when the values of the disturbance term are not independent and as a consequence some, or all, of the off-diagonal elements are non-zero. It is usual in initial treatments to retain the assumption of identical distributions, so that the diagonal elements of the variance-covariance matrix are the same. Of course, in principle one could have both types of violation at the same time.

In abstract, it is conventional to denote the variance-covariance matrix of the disturbance term \( \Omega \sigma^2_u \), where \( \Omega \) is the Greek upper case omega, writing the model:

\[
y = X\beta \quad \text{with} \quad E(uu') = \Omega \sigma^2_u. \tag{6}
\]

If the values of the disturbance term are iid, \( \Omega = I \). If they are not iid, OLS is in general inefficient and the standard errors are estimated incorrectly. Then, it is desirable to transform the model so that the transformed disturbance terms are iid. One possible way of doing this is to multiply through by some suitably chosen matrix \( P \), fitting:

\[
P y = P X \beta + P u
\]

choosing \( P \) so that \( E(P uu'P') = I \alpha \) where \( \alpha \) is some scalar. The solution for heteroskedasticity was a simple example of this type. We had:

\[
\Omega = \begin{bmatrix}
\sigma^2_u & 0 & \cdots & 0 \\
0 & \sigma^2_u & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma^2_u
\end{bmatrix}
\]

and the appropriate choice of \( P \) was:

\[
P = \begin{bmatrix}
\sqrt{\frac{1}{\sigma^2_{u1}}} & 0 & \cdots & 0 \\
\vdots & \ddots & \cdots & \vdots \\
0 & \cdots & \sqrt{\frac{1}{\sigma^2_{ui}}} & 0 \\
\vdots & \vdots & \cdots & \sqrt{\frac{1}{\sigma^2_{un}}}
\end{bmatrix}
\]

In the case of heteroskedasticity, the choice of \( P \) is obvious, provided, of course, that one knows the values of the diagonal elements of \( \Omega \). The more general theory requires an understanding of eigenvalues and eigenvectors that will be assumed. \( \Omega \) is a
15. Regression analysis with linear algebra primer

symmetric matrix since \( \text{cov}(u_i, u_j) \) is the same as \( \text{cov}(u_j, u_i) \). Hence all its eigenvalues are real. Let \( \Lambda \) be the diagonal matrix with the eigenvalues as the diagonal elements. Then there exists a matrix of eigenvectors, \( C \), such that:

\[
C'\Omega C = \Lambda. \tag{7}
\]

\( C \) has the properties that \( CC' = I \) and \( C' = C^{-1} \). Since \( \Lambda \) is a diagonal matrix, if its eigenvalues are all positive (which means that it is what is known as a ‘positive definite’ matrix), it can be factored as \( \Lambda = \Lambda^{1/2}\Lambda^{1/2} \) where \( \Lambda^{1/2} \) is a diagonal matrix whose diagonal elements are the square roots of the eigenvalues. It follows that the inverse of \( \Lambda \) can be factored as \( \Lambda^{-1} = \Lambda^{-1/2}\Lambda^{-1/2} \). Then, in view of (7):

\[
\Lambda^{-1/2}[C'\Omega C]\Lambda^{-1/2} = \Lambda^{-1/2}\Lambda\Lambda^{-1/2} = \Lambda^{-1/2}\Lambda^{1/2}\Lambda^{1/2}\Lambda^{-1/2} = I. \tag{8}
\]

This, if we define \( P = \Lambda^{-1/2}C' \), (8) becomes:

\[
P\Omega P' = I.
\]

As a consequence, if we premultiply (6) through by \( P \), we have:

\[
Py = PX\beta + Pu
\]

or:

\[
y^* = X'/\beta + u^*
\]

where \( y^* = Py \), \( X^* = PX \), and \( u^* = Pu \), and \( E(u'u'^*) = \sigma^2 \). An OLS regression of \( y^* \) on \( X^* \) will therefore satisfy the usual regression model assumptions and the estimator of \( \beta \) will have the usual properties. Of course, the approach usually requires the estimation of \( \Omega \), \( \Omega \) being positive definite, and there being no problems in extracting the eigenvalues and determining the eigenvectors.

**Exercise**

13. Suppose that the disturbance term in a simple regression model (with an intercept) is subject to AR(1) autocorrelation with \( |\rho| < 1 \), and suppose that the sample consists of just two observations. Determine the variance-covariance matrix of the disturbance term, find its eigenvalues, and determine its eigenvectors. Hence determine \( P \) and state the transformed model. Verify that the disturbance term in the transformed model is iid.

**15.18 Appendix A: Derivation of the normal equations**

We have seen that \( RSS \) is given by:

\[
RSS = y'y - 2y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}. \tag{A.1}
\]

The normal equations are:

\[
\frac{\partial RSS}{\partial \beta_j} = 0 \tag{A.2}
\]

330
Appendix A: Derivation of the normal equations

for $j = 1, \ldots, k$. We will show that they can be written:

$$X'X\hat{\beta} - X'y = 0.$$ 

The proof is mathematically unchallenging but tedious because one has to keep careful track of the dimensions of all of the elements in the equations. As far as I know, it is of no intrinsic interest and once one has seen it there should never be any reason to look at it again.

First note that the term $y'y$ in (A.1) is not a function of any of the $b_j$ and disappears in (A.2). Accordingly we will restrict our attention to the other two terms on the right side of (A.1). Suppose that we write the $X$ matrix as a set of column vectors:

$$X = [x_1 \cdots x_j \cdots x_k] \quad \text{(A.3)}$$

where:

$$x_j = \begin{bmatrix} X_{1j} \\ \vdots \\ X_{nj} \end{bmatrix}.$$ 

Then:

$$y'X\hat{\beta} = [y'x_1 \cdots y'x_j \cdots y'x_k] \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = [y'x_1\hat{\beta}_1 + \cdots + y'x_j\hat{\beta}_j + \cdots + y'x_k\hat{\beta}_k].$$

Hence:

$$\frac{\partial y'X\hat{\beta}}{\partial \hat{\beta}_j} = y'x_j.$$ 

We now consider the $\hat{\beta}'X'X\hat{\beta}$ term. Using (A.3):

$$\hat{\beta}'X'X\hat{\beta} = [x_1\hat{\beta}_1 + \cdots + x_j\hat{\beta}_j + \cdots + x_k\hat{\beta}_k]^\prime [x_1\hat{\beta}_1 + \cdots + x_j\hat{\beta}_j + \cdots + x_k\hat{\beta}_k]$$

$$= \sum_{p=1}^k \sum_{q=1}^k \hat{\beta}_p \hat{\beta}_q x_p'x_q.$$ 

The subset of terms including $\hat{\beta}_j$ is:

$$\sum_{q=1}^k \hat{\beta}_j \hat{\beta}_q x'_jx_q + \sum_{p=1}^k \hat{\beta}_p \hat{\beta}_j x'_p x_j.$$ 

Hence:

$$\frac{\partial \hat{\beta}'X'X\hat{\beta}}{\partial \hat{\beta}_j} = \sum_{q=1}^k \hat{\beta}_q x'_jx_q + \sum_{p=1}^k \hat{\beta}_p x'_p x_j = 2 \sum_{p=1}^k \hat{\beta}_p x'_p x_j.$$
15. Regression analysis with linear algebra primer

Putting these results together:

\[
\frac{\partial RSS}{\partial \hat{\beta}_j} = \frac{\partial [y'y - 2y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}]}{\partial \hat{\beta}_j} = -2y'x_j + 2 \sum_{p=1}^{k} \hat{\beta}_p x'_p x_j.
\]

Hence the normal equation \(\frac{\partial RSS}{\partial \hat{\beta}_j} = 0\) is:

\[
\sum_{p=1}^{k} \hat{\beta}_p x'_p x_j = x'_j y.
\]

(Note that \(x'_p x_j = x'_j x_p\) and \(y'x_j = x'_j y\) since they are scalars.) Hence:

\[
x'_j \left[ \sum_{p=1}^{k} \hat{\beta}_p x_p \right] = x'_j y.
\]

Hence:

\[
x'_j X\hat{\beta} = x'_j y
\]

since:

\[
X\hat{\beta} = [x_1 \cdots x_p \cdots x_k] \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \sum_{p=1}^{k} x_p \hat{\beta}_p.
\]

Hence, stacking the \(k\) normal equations:

\[
\begin{bmatrix} x'_1 X\hat{\beta} \\ \vdots \\ x'_j X\hat{\beta} \\ \vdots \\ x'_k X\hat{\beta} \end{bmatrix} = \begin{bmatrix} x'_1 y \\ \vdots \\ x'_j y \\ \vdots \\ x'_k y \end{bmatrix}.
\]

Hence:

\[
\begin{bmatrix} x'_1 \\ \vdots \\ x'_j \\ \vdots \\ x'_k \end{bmatrix} X\hat{\beta} = \begin{bmatrix} x'_1 \\ \vdots \\ x'_j \\ \vdots \\ x'_k \end{bmatrix} y.
\]

Hence:

\[
X'X\hat{\beta} = X'y.
\]

15.19 Appendix B: Demonstration that \(\hat{\mu}'\hat{\mu}/(n - k)\) is an unbiased estimator of \(\sigma^2_u\)

This classic proof is both elegant, in that it is much shorter than any proof not using matrix algebra, and curious, in that it uses the trace of a matrix, a feature that I have
never seen used for any other purpose. The trace of a matrix, defined for square matrices only, is the sum of its diagonal elements. We will first need to demonstrate that, for any two conformable matrices whose product is square:

\[ tr(AB) = tr(BA). \]

Let \( A \) have \( n \) rows and \( m \) columns, and let \( B \) have \( m \) rows and \( n \) columns. Diagonal element \( i \) of \( AB \) is:

\[ \sum_{p=1}^{m} a_{ip} b_{pi}. \]

Hence:

\[ tr(AB) = \sum_{i=1}^{n} \left( \sum_{p=1}^{m} a_{ip} b_{pi} \right). \]

Similarly, diagonal element \( i \) of \( BA \) is:

\[ \sum_{p=1}^{n} b_{ip} a_{pi}. \]

Hence:

\[ tr(BA) = \sum_{i=1}^{m} \left( \sum_{p=1}^{n} b_{ip} a_{pi} \right). \]

What we call the symbols used to index the summations makes no difference. Re-writing \( p \) as \( i \) and \( i \) as \( p \), and noting that the order of the summation makes no difference, we have \( tr(BA) = tr(AB) \).

We also need to note that:

\[ tr(A + B) = tr(A) + tr(B) \]

where \( A \) and \( B \) are square matrices of the same dimension. This follows immediately from the way that we sum conformable matrices.

By definition:

\[ \hat{u} = y - \hat{y} = y - X\hat{\beta}. \]

Using:

\[ \hat{\beta} = [X'X]^{-1}X'y \]

we have:

\[ \hat{u} = y - X[X'X]^{-1}X'y \]
\[ = X\beta + u - X[X'X]^{-1}X'(X\beta + u) \]
\[ = I_n u - X[X'X]^{-1}X'u \]
\[ = Mu \]

where \( I_n \) is an identity matrix of dimension \( n \) and:

\[ M = I_n - X[X'X]^{-1}X'. \]
15. Regression analysis with linear algebra primer

Hence:

\[ \hat{u}'\hat{u} = u'M'Mu. \]

Now \( M \) is symmetric and idempotent: \( M' = M \) and \( MM = M \). Hence:

\[ \hat{u}'\hat{u} = u'Mu. \]

\( \hat{u}'\hat{u} \) is a scalar, and so the expectation of \( \hat{u}'\hat{u} \) and the expectation of the trace of \( \hat{u}'\hat{u} \) are the same. So:

\[ E(\hat{u}'\hat{u}) = E(tr(\hat{u}'\hat{u})) = E(tr(u'Mu)) = E(tr(Muu')) = tr(E(Muu')). \]

The penultimate line uses \( tr(AB) = tr(BA) \). The last line uses the fact that the expectation of the sum of the diagonal elements of a matrix is equal to the sum of their individual expectations. Assuming that \( X \), and hence \( M \), is nonstochastic:

\[ E(\hat{u}'\hat{u}) = tr(ME(uu')) = tr(MI_n\sigma_u^2) = \sigma_u^2 tr(M) = \sigma_u^2(tr(I_n) - tr(X[X'X]^{-1}X')) = \sigma_u^2(tr(I_n) - tr(X[X'X]^{-1}X')). \]

The last step uses \( tr((A + B) = tr(A) + tr(B) \). The trace of an identity matrix is equal to its dimension. Hence:

\[ E(\hat{u}'\hat{u}) = \sigma_u^2(n-tr(X[X'X]^{-1}X')) = \sigma_u^2(n-tr(X[X'X]^{-1})) = \sigma_u^2(n-tr(I_k)) = \sigma_u^2(n-k). \]

Hence \( \hat{u}'\hat{u}/(n-k) \) is an unbiased estimator of \( \sigma_u^2 \).

15.20 Appendix C: Answers to the exercises

1. Given any square matrix \( C \), another matrix \( D \) is said to be its inverse if and only if \( CD = DC = I \). Thus, if \( B \) is the inverse of \( A \), \( AB = BA = I \). Now focus on the matrix \( B \). Since \( BA = AB = I \), \( A \) is its inverse. Hence the inverse of an inverse is the original matrix.

2. Suppose that two different matrices \( B \) and \( C \) both satisfied the conditions for being the inverse of \( A \). Then \( BA = I \) and \( AC = I \). Consider the matrix \( BAC \). Using \( BA = I \), \( BAC = C \). However, using \( AC = I \), \( BAC = B \). Hence \( B = C \) and it is not possible for \( A \) to have two separate inverses.

3. \( A_{ij} \), and hence \( A'_{ji} \), is the inner product of row \( i \) of \( B \) and column \( j \) of \( C \). If one writes \( D = C'B' \), \( D_{ji} \) is the inner product of row \( j \) of \( C' \) and column \( i \) of \( B' \), that is, column \( j \) of \( C \) and row \( i \) of \( B \). Hence \( D_{ji} = A_{ij} \), so \( D = A' \) and \( C'B' = (BC)' \).

4. Let \( D \) be the inverse of \( A \). Then \( D \) must satisfy \( AD = DA = I \). Now \( A = BC \), so \( D \) must satisfy \( BCD = DBC = I \). \( C^{-1}B^{-1} \) satisfies both of these conditions, since \( BCC^{-1}B^{-1} = BIB^{-1} = I \) and \( C^{-1}B^{-1}BC = C^{-1}IC = I \). Hence \( C^{-1}B^{-1} \) is the inverse of \( BC \) (assuming that \( B^{-1} \) and \( C^{-1} \) exist).
5. Let $B = A^{-1}$. Then $BA = AB = I$. Hence, using the result from Exercise 3, $A^\prime B^\prime = B^\prime A^\prime = I^\prime = I$. Hence $B^\prime$ is the inverse of $A^\prime$. In other words, $[A^{-1}]^\prime = [A^\prime]^{-1}$.

6. The relationship $Y = \beta_1 + \beta_2 X + u$ may be written in linear algebra form as $y = X\beta + u$ where $X = [1 \ x]$ and $1$ is the unit vector and:

$$x = \begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_n \end{bmatrix}.$$ 

Then:

$$X^\prime X = \begin{bmatrix} 1' \\ x' \end{bmatrix} \begin{bmatrix} 1 & x \end{bmatrix} = \begin{bmatrix} 1'1 & 1'x \\ x'1 & x'x \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}.$$ 

The determinant of $X^\prime X$ is:

$$n \sum X_i^2 - (\sum X_i)^2 = n \sum X_i^2 - n^2 X^2.$$ 

Hence:

$$[X^\prime X]^{-1} = \frac{1}{n \sum X_i^2 - n^2 X^2} \begin{bmatrix} \sum X_i^2 & -nX \\ -nX & n \end{bmatrix}.$$ 

We also have:

$$X^\prime y = \begin{bmatrix} 1'y \\ x'y \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_iY_i \end{bmatrix}.$$ 

So:

$$\hat{\beta} = [X^\prime X]X^\prime y = \frac{1}{n \sum X_i^2 - n^2 X^2} \begin{bmatrix} \sum X_i^2 & -nX \\ -nX & n \end{bmatrix} \begin{bmatrix} nY \\ \sum X_iY_i \end{bmatrix} = \frac{1}{n \sum X_i^2 - n^2 X^2} \begin{bmatrix} \sum Y_iX_i^2 - n\sum X_iY_i \\ -n\sum X_iY_i + n\sum X_iY_i \end{bmatrix} = \frac{1}{\sum (X_i - \bar{X})^2} \begin{bmatrix} \sum Y_iX_i^2 - \bar{X}\sum X_iY_i \\ \sum (X_i - \bar{X})(Y_i - \bar{Y}) \end{bmatrix}.$$ 

Thus:

$$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$ 

and:

$$\hat{\beta}_1 = \frac{\bar{Y}\sum X_i^2 - \bar{X}\sum X_iY_i}{\sum (X_i - \bar{X})^2}.$$
15. Regression analysis with linear algebra primer

\( \hat{\beta}_1 \) may be written in its more usual form as follows:

\[
b_1 = \frac{\bar{Y} \left( \sum X_i^2 - n\bar{X}^2 \right) + \bar{Y} n\bar{X}^2 - \bar{X} \sum X_i Y_i}{\sum (X_i - \bar{X})^2}
\]

\[
= \frac{\bar{Y} \left( \sum (X_i - \bar{X})^2 \right) - \bar{X} \left( \sum X_i Y_i - n\bar{Y}^2 \right)}{\sum (X_i - \bar{X})^2}
\]

\[
= \bar{Y} - \frac{\bar{X} \left( \sum (X_i - \bar{X}) (Y_i - \bar{Y}) \right)}{\sum (X_i - \bar{X})^2}
\]

\[
= \bar{Y} - \hat{\beta}_2 \bar{X}.
\]

7. If \( Y = \beta_2 X + u, \ y = X\beta + u \) where:

\[
X = x = \begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_n \end{bmatrix}.
\]

Then:

\[
X'X = x'x = \sum X_i^2.
\]

The inverse of \( X'X \) is \( 1/\sum X_i^2 \). In this model, \( X'y = x'y = \sum X_i Y_i \). So:

\[
\hat{\beta} = [X'X]^{-1}X'y = \frac{\sum X_i Y_i}{\sum X_i^2}.
\]

8. If \( Y = \beta_1 + u, \ y = X\beta + u \) where \( X = 1 \), the unit vector. Then \( X'X = 1'1 = n \) and its inverse is \( 1/n \).

\[
X'y = 1'y = \sum Y_i = n\bar{Y}.
\]

So:

\[
\hat{\beta} = [X'X]^{-1}X'y = \frac{1}{n}n\bar{Y} = \bar{Y}.
\]

9. We will start with \( Y \). If we regress it on the intercept, we are regressing it on \( 1 \), the unit vector, and, as we saw in Exercise 8, the coefficient is \( \bar{Y} \). Hence the residual in observation \( i \) is \( Y_i - \bar{Y} \). The same is true for each of the \( X \) variables when regressed on the intercept. So when we come to regress the residuals of \( Y \) on the residuals of the \( X \) variables, we are in fact using the demeaned data for \( Y \) and the demeaned data for the \( X \) variables.

10. The general form of the IV estimator is \( \hat{\beta}^{IV} = [W'X]^{-1}W'y \). In the case of the simple regression model, with \( Z \) acting as an instrument for \( X \) and the unit vector acting as an instrument for itself, \( W = [1 \ z] \) and \( X = [1 \ x] \). Thus:

\[
W'X = \begin{bmatrix} 1' \\ z' \end{bmatrix} \begin{bmatrix} 1 & x \end{bmatrix} = \begin{bmatrix} 1'1 & 1'x \\ z'1 & z'x \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum Z_i & \sum Z_i X_i \end{bmatrix}.
\]
The determinant of $W'X$ is:

$$n \sum Z_iX_i - \left( \sum Z_i \right) \left( \sum X_i \right) = n \sum Z_iX_i - n^2\bar{Z}\bar{X}.$$ 

Hence:

$$[W'X]^{-1} = \frac{1}{n \sum Z_iX_i - n^2\bar{Z}\bar{X}} \begin{bmatrix}
\sum Z_iX_i & -n\bar{X} \\
-n\bar{Z} & \ 
\end{bmatrix}.$$ 

We also have:

$$W'y = \left[ 1' \bar{y} \right] = \left[ \sum Y_i \sum Z_iY_i \right].$$

So:

$$\hat{\beta}^{IV} = [W'X]^{-1}W'y$$

$$= \frac{1}{n \sum Z_iX_i - n^2\bar{Z}\bar{X}} \left[ \sum Z_iX_i - n\bar{X} \right] \left[ \sum Y_i \sum Z_iY_i \right]$$

Thus:

$$\hat{\beta}_2^{IV} = \frac{\sum \left( Z_i - \bar{Z} \right) \left( Y_i - \bar{Y} \right)}{\sum \left( Z_i - \bar{Z} \right) \left( X_i - \bar{X} \right)}$$

and:

$$\hat{\beta}_1^{IV} = \frac{\sum Z_iX_i - \bar{X}\sum Z_iY_i}{\sum \left( Z_i - \bar{Z} \right) \left( X_i - \bar{X} \right)}.$$ 

$\hat{\beta}_1^{IV}$ may be written in its more usual form as follows:

$$\hat{\beta}_1^{IV} = \frac{\bar{Y} \left( \sum Z_iX_i - n\bar{Z}\bar{X} \right) + \bar{Y}n\bar{Z}\bar{X} - \bar{X}\sum Z_iY_i}{\sum \left( Z_i - \bar{Z} \right) \left( X_i - \bar{X} \right)}$$

$$= \frac{\bar{Y} \left( \sum \left( Z_i - \bar{Z} \right) \left( X_i - \bar{X} \right) \right) - \bar{X} \left( \sum Z_iY_i - n\bar{X}\bar{Y} \right)}{\sum \left( Z_i - \bar{Z} \right) \left( X_i - \bar{X} \right)}$$

$$= \bar{Y} - \frac{\bar{X} \left( \sum \left( Z_i - \bar{Z} \right) \left( Y_i - \bar{Y} \right) \right)}{\sum \left( Z_i - \bar{Z} \right) \left( X_i - \bar{X} \right)}$$

$$= \bar{Y} - \hat{\beta}_2^{IV}\bar{X}.$$ 

11. By definition, if one of the variables in $X$ is acting as an instrument for itself, it is included in the $W$ matrix. If it is regressed on $W$, a perfect fit is obtained by
15. Regression analysis with linear algebra primer

assigning its column in \( W \) a coefficient of 1 and assigning zero values to all the other coefficients. Hence its fitted values are the same as its original values and it is not affected by the first stage of Two-Stage Least Squares.

12. If the variables in \( X \) are regressed on \( W \) and the matrix of fitted values of \( X \) saved:

\[
\hat{X} = W[W'W]^{-1}W'X.
\]

If \( \hat{X} \) is used as the matrix of instruments:

\[
\hat{\beta}^{\text{TLS}} = [\hat{X}'X]^{-1}\hat{X}'y
\]

\[
= [X'W[W'W]^{-1}W'X']^{-1}X'W[W'W]^{-1}W'y
\]

\[
= [W'X]^{-1}W'W[X'W]^{-1}X'W[W'W]^{-1}W'y
\]

\[
= [W'X]^{-1}W'X
\]

\[
= \hat{\beta}^{\text{IV}}.
\]

Note that, in going from the second line to the third, we have used 
\[
[ABC]^{-1} = C^{-1}B^{-1}A^{-1},
\]
and we have exploited the fact that \( W'X \) is square and possesses an inverse.

13. The variance-covariance matrix of \( u \) is:

\[
\begin{bmatrix}
1 & \rho \\
\rho & 1
\end{bmatrix}
\]

and hence the characteristic equation for the eigenvalues is:

\[
(1 - \lambda)^2 - \rho^2 = 0.
\]

The eigenvalues are therefore \( 1 - \rho \) and \( 1 + \rho \). Since we are told \( |\rho| < 1 \), the matrix is positive definite.

Let:

\[
c = \begin{bmatrix}
c_1 \\
c_2
\end{bmatrix}.
\]

If \( \lambda = 1 - \rho \), the matrix \( A - \lambda I \) is given by:

\[
A - \lambda I = \begin{bmatrix}
\rho & \rho \\
\rho & \rho
\end{bmatrix}
\]

and hence the equation:

\[
[A - \lambda I]c = 0
\]

yields:

\[
\rho c_1 + \rho c_2 = 0.
\]

Hence, also imposing the normalisation:

\[
c'c = c_1^2 + c_2^2 = 1
\]
we have \( c_1 = 1 / \sqrt{2} \) and \( c_2 = -1 / \sqrt{2} \), or vice versa. If \( \lambda = 1 + \rho \):

\[
A - \lambda I = \begin{bmatrix} -\rho & \rho \\ \rho & -\rho \end{bmatrix}
\]

and hence \([A - \lambda I]c = 0\) yields:

\[-\rho c_1 + \rho c_2 = 0.\]

Hence, also imposing the normalisation:

\[c'c = c_1^2 + c_2^2 = 1\]

we have \( c_1 = c_2 = 1 / \sqrt{2} \). Thus:

\[
C = \begin{bmatrix} 1 / \sqrt{2} & 1 / \sqrt{2} \\ -1 / \sqrt{2} & 1 / \sqrt{2} \end{bmatrix}
\]

and:

\[
P = \Lambda^{-1/2}C' = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 / \sqrt{1+\rho} & 0 \\ 0 & 1 / \sqrt{1+\rho} \end{bmatrix} \begin{bmatrix} 1 / \sqrt{1-\rho} & -1 / \sqrt{1-\rho} \\ 1 / \sqrt{1-\rho} & 1 / \sqrt{1+\rho} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 / \sqrt{1-\rho} & -1 / \sqrt{1-\rho} \\ 1 / \sqrt{1-\rho} & 1 / \sqrt{1-\rho} \end{bmatrix}.
\]

It may then be verified that \( P\Omega P' = I \):

\[
\frac{1}{\sqrt{2}} \begin{bmatrix} 1 / \sqrt{1-\rho} & -1 / \sqrt{1-\rho} \\ 0 & 1 / \sqrt{1+\rho} \end{bmatrix} \begin{bmatrix} 1 / \sqrt{1-\rho} & 1 / \sqrt{1-\rho} \\ 1 / \sqrt{1+\rho} & 1 / \sqrt{1+\rho} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 / \sqrt{1-\rho} & 1 / \sqrt{1+\rho} \\ 1 / \sqrt{1+\rho} & 1 / \sqrt{1-\rho} \end{bmatrix} \begin{bmatrix} 1 / \sqrt{1-\rho} & -1 / \sqrt{1-\rho} \\ 1 / \sqrt{1+\rho} & 1 / \sqrt{1+\rho} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.
\]

The transformed model has:

\[
y^* = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 / \sqrt{1-\rho} (y_1 - y_2) \\ 1 / \sqrt{1+\rho} (y_1 + y_2) \end{bmatrix}
\]

and parallel transformations for the \( X \) variables and \( u \). Given that:

\[
u^* = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 / \sqrt{1-\rho} (u_1 - u_2) \\ 1 / \sqrt{1+\rho} (u_1 + u_2) \end{bmatrix}
\]
Regression analysis with linear algebra primer

none of its elements is the white noise \( \varepsilon \) in the AR(1) process, but nevertheless its elements are iid.

\[
\begin{align*}
\text{var}(u_1^*) &= \frac{1}{2} \frac{1}{1 - \rho} \left( \text{var}(u_1) + \text{var}(u_2) - 2\text{cov}(u_1, u_2) \right) \\
&= \frac{1}{2} \frac{1}{1 - \rho} \left( \sigma_u^2 + \sigma_u^2 - 2\rho\sigma_u^2 \right) = \sigma_u^2 \\
\text{var}(u_2^*) &= \frac{1}{2} \frac{1}{1 + \rho} \left( \text{var}(u_1) + \text{var}(u_2) + 2\text{cov}(u_1, u_2) \right) \\
&= \frac{1}{2} \frac{1}{1 + \rho} \left( \sigma_u^2 + \sigma_u^2 + 2\rho\sigma_u^2 \right) = \sigma_u^2 \\
\text{cov}(u_1^*, u_2^*) &= \frac{1}{2} \frac{1}{\sqrt{1 - \rho^2}} \text{cov} \left( (u_1 - u_2), (u_1 + u_2) \right) \\
&= \frac{1}{2} \frac{1}{\sqrt{1 - \rho^2}} \left( \text{var}(u_1) + \text{cov}(u_1, u_2) - \text{cov}(u_2, u_1) - \text{var}(u_2) \right) \\
&= 0.
\end{align*}
\]

Hence \( E(u^* u'^*) = I \sigma_u^2 \). Of course, this was the objective of the \( P \) transformation.