

## Regression Basics in Matrix Terms

### 1 The Normal Equations of least squares

Let  $y$  denote the dependent variable, a  $n \times 1$  vector, and let  $X$  denote the  $n \times k$  matrix of regressors. Write  $b$  for the  $k$ -vector of regression coefficients, and write  $e$  for the  $n$ -vector of residuals,  $e_i = y_i - X_i b$ .

The vector  $b$  is the least squares solution if and only if it is chosen such that the sum of squared residuals,

$$\text{SSR} = \sum_{i=1}^n e_i^2,$$

is at a minimum.

Attaining the minimum SSR can be approached as a calculus problem. In matrix notation, we can write the SSR as

$$\begin{aligned} e'e &= (y - Xb)'(y - Xb) \\ &= (y' - (Xb)')(y - Xb) \\ &= y'y - y'Xb - (Xb)'y + (Xb)'Xb \\ &= y'y - 2b'X'y + b'X'Xb \end{aligned}$$

The manipulations above exploit the properties of the matrix transpose, namely  $(A + B)' = A' + B'$  and  $(AB)' = B'A'$ . Now take the derivative of the last expression above with respect to  $b$  and set it to zero. This gives

$$\begin{aligned} -2X'y + 2X'Xb &= 0 \\ \Rightarrow -X'y + X'Xb &= 0 \\ \Rightarrow X'Xb &= X'y \end{aligned}$$

If the matrix  $X'X$  is non-singular (i.e. it possesses an inverse) then we can multiply by  $(X'X)^{-1}$  to get

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

This is a classic equation in statistics: it gives the least squares regression coefficients as a function of the data matrices,  $X$  and  $y$ .

**Orthogonality of residuals and regressors.** Here's another perspective. As we said above, the regression residuals are  $e = y - Xb$ . Suppose we require that the residuals be *orthogonal* to the regressors,  $X$ . We then have

$$\begin{aligned} X'e &= 0 \\ \Rightarrow X'(y - Xb) &= 0 \\ \Rightarrow X'Xb &= X'y \end{aligned}$$

which replicates the solution above. So another way of thinking about the least squares coefficient vector,  $b$ , is that it satisfies the condition that the residuals are orthogonal to the regressors. Why is that "a good thing"? Well, intuitively, if this orthogonality condition were not satisfied, that would mean that the predictive power of  $X$  with regard to  $y$  is not exhausted. The residuals represent the "unexplained" variation in  $y$ ; if they are not orthogonal to  $X$  it follows that more explanation can be squeezed out of  $X$  by a different choice of coefficients.

#### Proof without calculus

Here is a cute derivation of the Normal Equations of least squares which does not rely on calculus.<sup>1</sup>

<sup>1</sup>With acknowledgement to Robert E. White, Professor of Mathematics at North Carolina State University.

The vector  $b$  is the least squares solution if and only if  $b$  is such that the sum of squared residuals,

$$SSR = e'e = (y - Xb)'(y - Xb),$$

is a minimum of all  $(y - Xc)'(y - Xc)$ . That is, if  $b$  is the least squares solution, then for all  $c$

$$(y - Xb)'(y - Xb) \leq (y - Xc)'(y - Xc) \quad (2)$$

Let us write  $c = b + (c - b)$ , so that  $Xc = Xb + X(c - b)$ . Again using the properties of the matrix transpose, we get

$$\begin{aligned} (y - Xc)'(y - Xc) &= [(y - Xb) - X(c - b)]' \cdot [(y - Xb) - X(c - b)] \\ &= [(y - Xb)' - (X(c - b))'] \cdot [(y - Xb) - X(c - b)] \\ &= (y - Xb)'(y - Xb) \\ &\quad - (X(c - b))'(y - Xb) \\ &\quad - (y - Xb)'X(c - b) \\ &\quad + (X(c - b))'X(c - b) \end{aligned} \quad (3)$$

$$\geq (y - Xb)'(y - Xb) - 2(c - b)'X'(y - Xb) \quad (4)$$

Why must the inequality (4) hold? Consider the terms on the right-hand side of (3), all of which are  $1 \times 1$  matrices or scalars. The first term recurs in (4), and the second and third are transposes of each other, both equal to

$$-(c - b)'X'(y - Xb),$$

since the transpose of a  $1 \times 1$  matrix is that same matrix. So the RHS of (3) differs from the RHS of (4) only in the term  $(X(c - b))'X(c - b)$ . But this is the scalar product of a vector and its own transpose, which is a sum of squares and so necessarily non-negative.

Now note that (4), which is assuredly true, is equivalent to (2), the condition that defines the least-squares  $b$ , if the last term on the right of (4), namely

$$2(c - b)'X'(y - Xb),$$

is zero. But this is bound to be the case (regardless of  $c$ ) if

$$\begin{aligned} X'(y - Xb) &= 0 \\ \Rightarrow X'y - X'Xb &= 0 \\ \Rightarrow X'Xb &= X'y \end{aligned}$$

So once again we arrive at the classic solution, equation (1).

## 2 The expected value of the least squares estimator

Up to this point we have been concerned with the arithmetic of least squares. From an econometric viewpoint, however, we're interested in the properties of least squares as an *estimator* of some data generating process (DGP) that we presume to be operating in the economy. Let us write the DGP as

$$y = X\beta + u$$

where  $u$  is the error or disturbance term, satisfying  $E(u) = 0$ . (The error term is sometimes represented by a Greek letter, often  $\epsilon$  or  $\varepsilon$ , but I'll use  $u$  here to keep the notation simpler.)

From this standpoint we consider the least squares coefficient vector  $b$  as an estimator,  $\hat{\beta}$ , of the unknown parameter vector  $\beta$ . By the same token our regression residuals,  $e$ , can be considered as estimates of the unknown errors  $u$ , or  $e = \hat{u}$ . In this section we consider the expected value of  $\hat{\beta}$ , with a question of particular interest being: under what conditions does  $E(\hat{\beta})$  equal  $\beta$ ? In the following section we examine the variance of  $\hat{\beta}$ .

Since, as we saw above, the least squares  $\hat{\beta}$  equals  $(X'X)^{-1}X'y$ , we can write

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

The next step is to substitute the presumed DGP for  $y$ , giving

$$\begin{aligned} E(\hat{\beta}) &= E[(X'X)^{-1}X'(X\beta + u)] \\ &= E[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'u] \\ &= E[\beta + (X'X)^{-1}X'u] \\ &= \beta + E[(X'X)^{-1}X'u] \end{aligned}$$

(Since  $\beta$  itself is not a random variable, it can be moved through the expectation operator.) We then see that  $E(\hat{\beta}) = \beta$ —in other words, the least squares estimator is unbiased—on condition that

$$E[(X'X)^{-1}X'u] = 0$$

What does it take to satisfy this condition? If we take expectations conditional on the observed values of the regressors,  $X$ , the requirement is just that

$$E(u|X) = 0 \tag{5}$$

i.e., the error term is independent of the regressors.

So: the least squares estimator is unbiased provided that (a) our favored specification of the DGP is correct and (b) the independent variables in  $X$  offer no predictive power over the errors  $u$ . We might say that the errors must be *uncorrelated* with the regressors, but the requirement in (5) is actually stronger than that.

### 3 The variance of the least squares estimator

Recall the basic definition of variance:

$$\text{Var}(x) = E[x - E(x)]^2 = E[(x - E(x))(x - E(x))]$$

The variance of a random variable  $x$  is the expectation of the squared deviation from its expected value. The above holds good for a scalar random variable. For a random *vector*, such as the least squares  $\hat{\beta}$ , the concept is similar except that squaring has to be replaced with the matrix counterpart, multiplication of a vector into its own transpose.

This requires a little thought. The vector  $\hat{\beta}$  is a column vector of length  $k$ , and so is the difference  $\hat{\beta} - E(\hat{\beta})$ . Given such a  $k$ -vector,  $v$ , if we form  $v'v$  (the inner product) we get a  $1 \times 1$  or scalar result, the sum of squares of the elements of  $v$ . If we form  $vv'$  (the outer product) we get a  $k \times k$  matrix. Which do we want here? The latter: by the variance of a random vector we mean the (co-)variance matrix, which holds the variances of the elements of the vector on its principal diagonal, and the covariances in the off-diagonal positions.

Therefore, from first principles,

$$\text{Var}(\hat{\beta}) = E \left[ \left( \hat{\beta} - E(\hat{\beta}) \right) \left( \hat{\beta} - E(\hat{\beta}) \right)' \right]$$

If we suppose that the conditions for the estimator to be unbiased are met, then  $E(\hat{\beta}) = \beta$  and, substituting in the least squares formula for  $\hat{\beta}$ , we get

$$\text{Var}(\hat{\beta}) = E \left[ \left( (X'X)^{-1}X'y - \beta \right) \left( (X'X)^{-1}X'y - \beta \right)' \right]$$

Then, once again, substitute the DGP,  $X\beta + u$ , for  $y$ , to get

$$\begin{aligned} \text{Var}(\hat{\beta}) &= E \left[ \left( (X'X)^{-1}X'(X\beta + u) - \beta \right) \left( (X'X)^{-1}X'(X\beta + u) - \beta \right)' \right] \\ &= E \left[ \left( \beta + (X'X)^{-1}X'u - \beta \right) \left( \beta + (X'X)^{-1}X'u - \beta \right)' \right] \\ &= E \left[ \left( (X'X)^{-1}X'u \right) \left( (X'X)^{-1}X'u \right)' \right] \\ &= E \left[ (X'X)^{-1}X'uu'X(X'X)^{-1} \right] \end{aligned} \tag{6}$$

(Note that the symmetric matrix  $(X'X)^{-1}$  is equal to its transpose.) As with figuring the expected value of  $\hat{\beta}$ , we will take expectations conditional on  $X$ . And in assuming that  $\hat{\beta}$  is unbiased, we have already assumed that  $E(X'u) = 0$ . So we can write

$$\text{Var}(\hat{\beta}) = (X'X)^{-1}X'E(uu')X(X'X)^{-1} \quad (7)$$

The awkward bit is the matrix in the middle,  $E(uu')$ . This is the covariance matrix of  $u$ ; note that it is  $n \times n$  and so potentially quite large. However, under certain conditions it simplifies greatly—specifically, if the error term is *independently and identically distributed* (i.i.d.). In that case  $E(uu')$  boils down to an  $n \times n$  version of the following,

$$\begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix},$$

that is, a matrix with a constant value on the diagonal and zeros elsewhere. The constant  $\sigma^2$  reflects the “identical” part of the i.i.d. condition (the error variance is the same at each observation) and the zeros off the diagonal reflect the “independent” part (there’s no correlation between the error at any given observation and the error at any other observation). Such an error term is sometimes referred to as “white noise”.

Under the assumption that  $u$  is i.i.d., then, we can write  $E(uu') = \sigma^2 I$ , where  $I$  is the  $n \times n$  identity matrix. And the scalar  $\sigma^2$  can be moved out, giving

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \sigma^2(X'X)^{-1}X'IX(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned} \quad (8)$$

The somewhat nasty equation (7) therefore gives the variance of  $\hat{\beta}$  in general case, while equation (8) gives the nice and simple version that arises when the error term is assumed to be i.i.d.

How do we actually get a value for (that is, an estimate of) the variance of  $\hat{\beta}$ ? The true error variance,  $\sigma^2$ , is unknown, but we can substitute an estimate,  $s^2$ , based on the regression residuals:

$$s^2 = \frac{\sum_{i=1}^n e_i^2}{n - k}$$

and our estimated variance is then  $s^2(X'X)^{-1}$ —provided, of course, that we are willing to assume that the error term is i.i.d.